Markov Chain Monte Carlo Inference of Parametric Dictionaries for Sparse Bayesian Approximations

Theodora Chaspari, *Student Member, IEEE*, Andreas Tsiartas, *Member, IEEE*, Panagiotis Tsilifis, and Shrikanth S. Narayanan, *Fellow, IEEE*

***Abstract*—Parametric dictionaries can increase the ability of sparse representations to meaningfully capture and interpret the underlying signal information, such as encountered in biomedical problems. Given a mapping function from the atom parameter space to the actual atoms, we propose a sparse Bayesian framework for learning the atom parameters, because of its ability to provide full posterior estimates, take uncertainty into account and generalize on unseen data. Inference is performed with Markov Chain Monte Carlo, that uses block sampling to generate the variables of the Bayesian problem. Since the parameterization of dictionary atoms results in posteriors that cannot be analytically computed, we use a Metropolis–Hastings–within-Gibbs framework, according to which variables with closed-form posteriors are generated with the Gibbs sampler, while the remaining ones with the Metropolis Hastings from appropriate candidate-generating densities. We further show that the corresponding Markov Chain is uniformly ergodic ensuring its convergence to a stationary distribution independently of the initial state. Results on synthetic data and real biomedical signals indicate that our approach offers advantages in terms of signal reconstruction compared to previously proposed Steepest Descent and Equiangular Tight Frame methods. This paper demonstrates the ability of Bayesian learning to generate parametric dictionaries that can reliably represent the exemplar data and provides the foundation towards inferring the entire variable set of the sparse approximation problem for signal denoising, adaptation, and other applications.**

***Index Terms*—Dictionary learning, parametric dictionaries, Bayesian inference, Markov chain Monte Carlo, sparse representation, uniform ergodicity.**

I. Introduction

SIGNAL representations are fundamental for denoising, interpolation, estimation, classification and recognition. Recently there has been an increased focus on sparse representations [1] which model a signal with a small number of components from a large overcomplete set of exemplar atoms, called dictionary. A dictionary $\mathbf{D} \in \Re^{P\times K}$ contains K atoms $\mathbf{d_1}, \ldots, \mathbf{d_K} \in \Re^P$ that constitute the building blocks of a P-dimensional signal $\mathbf{x} \in \Re^P$. Specifically a signal can be expressed as an exact or approximate linear combination of a small number of atoms from the dictionary as $\mathbf{x} \approx \mathbf{Dc}$, where $\mathbf{c} \in \Re^K$ contains the coefficients of the corresponding atoms. In the typical case where $K > P$, an infinite set of solutions arise. This can be addressed by imposing a sparsity constraint on $\mathbf{x}$, according to which $\mathbf{x}$ should be represented by the smallest number of dictionary atoms and the sparse representation problem can be expressed as an $l0$-norm minimization.

The problem of minimizing $||\mathbf{c}||_0$ subject to the constraint $\mathbf{x} = \mathbf{Dc}$ has been proven to be NP-hard and several directions have been proposed to solve it. One approach includes greedy strategies that abandon exhaustive search in favor of locally optimal updates resulting in sub-optimal solutions. Examples include matching pursuit [2] and orthogonal matching pursuit [3], [4] algorithms. An alternative is the relaxation of the discontinuous $l0$-norm leading to the more computationally expensive basis pursuit [5] and focal underdetermined system solver [6], [7] reaching global solutions. Bayesian methods with appropriate statistical assumptions have been further used to identify the desired sparse solution [8]–[10].

An essential step towards compact and reliable representations is the dictionary selection. Traditionally, analytic pre-designed dictionaries comprising Gabor [11], wavelet [12], curvelet [13], or other atoms have been used, because of their localization, directionality and multi-resolution properties. Dictionary learning (DL) focuses on learning atoms from the available training data. It includes several well-known algorithms, such as the K-SVD [14] and the MOD [15], as well as probabilistic approaches [16]. Although non-parametric DL is effective for signal reconstruction [14], restoration [17], and classification [18], it depicts a highly non-convex nature, mostly yields non-structured dictionaries [19] and typically requires a large amount of training data [20].

These disadvantages can be mitigated by imposing a predetermined structure through the use of carefully selected knowledge-driven parametric functions mapping a parameter space to structured dictionary atoms. This results in parametric dictionaries bridging the gap between pre-defined analytic dictionaries and purely numerical DL [21]. Dictionary atoms are expressed through an application-specific function ϕ of a parameter set, say $\mathbf{d_k} = \phi(\boldsymbol{\theta_k})$, where $\boldsymbol{\theta_k} \in \Re^Q$, $Q < P$, are the atom parameters optimized with respect to desirable properties [22]–[24]. Parametric DL is more likely to converge faster and have more efficient implementations compared to the non-

Manuscript received July 11, 2015; revised November 22, 2015 and February 10, 2016; accepted February 12, 2016. Date of publication March 07, 2016; date of current version April 21, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wenwu Wang. This work was funded by National Science Foundation and National Institute of Health. *(Corresponding author: Theodora Chaspari.)*

T. Chaspari and S. S. Narayanan are with the Signal Analysis and Interpretation Laboratory (SAIL), Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: chaspari@usc.edu; shri@sipi.usc.edu).

A. Tsiartas is with SRI International, Menlo Park, CA 94025-3493 USA (e-mail: andreas.tsiartas@sri.com).

P. Tsilifis is with the Department of Mathematics at the Dana and David Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA 90089 USA (e-mail: tsilifis@usc.edu).

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

Digital Object Identifier 10.1109/TSP.2016.2539143

parametric problem [19]. It further provides higher signal interpretability yielding important meta-information [25]–[28].

This paper proposes a Bayesian framework for learning the parameters of a dictionary given a predetermined parametric function. The formulation of the sparse representation problem from the Bayesian perspective assumes a probabilistic distribution for all variables in an effort to provide a posterior belief for their values [10], [29]. This allows the estimation of full posteriors rather than single estimates which can result in better handling uncertainty and benefits noise estimation.

Related Work

DL methods usually alternate between sparse decoding and dictionary update [19]. In the context of non-parametric DL, Ophir *et al.* have proposed a sequential learning algorithm by identifying the orthogonal directions to a data subset [30]. Engan *et al.* used matrix inversion to compute the dictionary matrix [15], while Aharon *et al.* introduced K-SVD, which is a constrained optimization approach performing atom-by-atom update [14]. Generalized principal component analysis models the data as a union of low-dimensional subspaces with orthogonal bases [31], while structured dictionaries have been proposed in an effort to enforce additional translation invariant, hierarchical and multiscale properties [32]–[34].

Parametric dictionaries depict higher interpretability, lower density of local minima and compact representation [19]. Dictionaries can be learnt as a result of translation of elementary signal segments over space and time [35], [36]. They can also contain atoms of predetermined structure, such as wavelets [37] or Gabor functions [22], whose parameters are adapted with steepest-descent (SD) [22] and other least-squares-based methods [25], [37]. Yanghoobi *et al.* proposed a method to find dictionaries close to the one with minimum coherence, called the equiangular tight frame (ETF) [24]. Finally, Thanou *et al.* [21] proposed a parametric DL for signals residing on weighted graphs.

Although the aforementioned deterministic DL methods perform well in various signal processing applications [14], [17], [18], they provide single-point estimates and cannot handle noise uncertainty. Bayesian approaches offer a way to address those disadvantages. They have been proposed for compressed sensing [8], [9], [38], [39] as well as for non-parametric DL.

The problem of ensuring sparseness in Bayesian approximations has been addressed in a variety of ways. Early approaches have used continuous sparse-promoting distributions to the atom coefficients $\mathbf{c_n}$, such as the Laplacian [40], Cauchy [16], [41] and Student-t [42], [10]. Another way is the use of indicator variables that permit to independently sample each atom from the Bernoulli distribution [43]–[45]. Despite their computational efficiency, these can yield an arbitrarily large amount of non-zero coefficients, which might not be practical for many applications. More inline with our problem, previous studies have explored the use of appropriate probabilistic assumptions for keeping a constant number of dictionary atoms in the representation with proposal distributions iteratively conditioning on the atoms selected at previous steps [39], [46], [38]. Gaussian assumptions are typically imposed on the dictionary atoms [43], [44] and the signal noise [38], [43], [44], [47], the latter being consistent with biomedical applications [48], [49].

Bayesian inference casts DL into an optimization problem for maximizing the posterior distribution. The overcomplete dictionaries containing many atoms in combination with the large amount of training data yield a high-dimensional framework, for which closed form solutions are usually difficult to derive and approximate inference methods are followed. Common approaches include the evidence maximization [8], [9], relevance vector machine [10], Markov Chain Monte Carlo (MCMC) [38], [44], and variational approximation [43].

Contributions

We propose a Bayesian framework for learning the parameters of dictionary atoms, given a parametric function that maps the dictionary parameters to the actual atoms. Our approach imposes probabilistic distributions to the variables of the sparse representation problem that are estimated through MCMC methods because of their simplicity and ability to fully perform Bayesian inference [50]. Compared to previous Bayesian DL [43], [44], our approach introduces parametric dictionaries where non-closed-form solutions are handled with a combination of Gibbs and Metropolis-Hastings (MH) sampling (MH-within-Gibbs). Our approach differs from previous parametric DL [22], [24] because of its stochastic framework that yields estimation of the full problem variables. This results in parametric dictionaries that take into account the structure of the training data and are less prone to overfitting. The parametric nature of our problem further requires the use of appropriate sparse-imposing priors that keep the selected number of dictionary atoms within a pre-determined range. We perform atom sampling with and without replacement formalized through the Multinomial and the Wallenius' hypergeometric distribution, respectively.

One key challenge with MCMC is to determine its asymptotic behavior, i.e., whether it provides accurate posterior approximations. The goal is to create an aperiodic and irreducible Markov Chain (MC) with stationary distribution same as the posterior distribution of interest [51]. Irreducibility ensures that any state of the space is accessible, while aperiodicity makes sure that the chain does not return to the same state at regular times. Uniformly ergodic MCs are a special case in which the MC converges to the invariant distribution independently of the initial state. They guarantee geometrically fast convergence and are key sufficient conditions in order to establish central limit theorems for empirical averages and provide consistent estimators of MCMC standard errors [51], [52]. Because of these, we discuss the geometric ergodicity of MCMC in our proposed Bayesian inference framework that ensures convergence. We further perform qualitative and quantitative diagnostics to evaluate the reliability of the generated samples and resulting distributions.

We demonstrate the ability of our algorithm for parallel processing with experiments on synthetic data and real biomedical signals. DL is performed for each sample separately and the resulting dictionaries from each exemplar data are further combined into a unified model. Our results indicate that the proposed approach yields benefits in terms of superior signal reconstruction compared to previous SD [22] and ETF [24] methods. When we have precise a priori knowledge of the optimal parametric function ϕ representing the data, our parametric framework also yields better performance than the classical non-parametric approach of K-SVD [14].

In the following, we provide the problem formulation (Section II) and describe the proposed MCMC approach for learning the parameters for the considered problem (Section III). We describe the parallel implementation framework of our algorithm (Section IV) and discuss the use of various parametric functions and their effect on the sample-generating procedure (Section V). We further analyze the geometric ergodicity properties of the proposed MH-within-Gibbs algorithm yielding uniform convergence (Section VI). In Section VII, we provide experimental results and the results of MCMC diagnostics. Finally, we discuss our results and offer conclusions in Sections VIII and IX.

Notation

We denote matrices with bold uppercase letters $\mathbf{X}$ and vectors with bold lowercase letters $\mathbf{x}$. Lowercase letters with appropriate numerical indices will either refer to the columns of a matrix, i.e., $\mathbf{X} = [\mathbf{x_1} \dots \mathbf{x_N}]$ or the elements of a vector, i.e., $\mathbf{x} = [x_1 \dots x_N]^T$. We denote $(\mathbf{X})_{ij}$ the entry of matrix $\mathbf{X}$ corresponding to the ith row and jth column. The pth order norm of a vector is symbolized as $||\mathbf{x}||_p$, while the $N \times 1$ identity vector and $N \times N$ identity matrix are noted as $\mathbf{1_N}$ and $\mathbf{I_N}$, respectively. The vectorization of matrix $\mathbf{X}$, obtained by stacking its columns on top of one another, is defined as $\text{vec}(\mathbf{X}) = [\mathbf{x_1^T}, \dots, \mathbf{x_N^T}]^T$. Finally the gradient and Hessian of a scalar valued function $f(\mathbf{x})$, $\mathbf{x} \in \Re^N$, are denoted as $\nabla f(\mathbf{x}) \in \Re^{N \times 1}$ and $H_f = \nabla^2 f(\mathbf{x}) \in \Re^{N \times N}$, where $(\nabla f(\mathbf{x}))_i = \frac{\theta f}{\theta x_i}$ and $(H_f)_{ij} = \frac{\vartheta^2 f}{\vartheta x_i \vartheta x_j}$, respectively.

II. Problem Formulation

Let $\mathbf{X} = [\mathbf{x_1} \dots \mathbf{x_N}] \in \Re^{P \times N}$ be a data matrix of N examples $\mathbf{x_n} \in \Re^P$. We formulate the parametric DL problem by assuming an overcomplete dictionary $\mathbf{D_n} \in \Re^{P \times K}$ containing K prototypical atoms $\mathbf{d_{nk}} \in \Re^P$ for each exemplar data $\mathbf{x_n}$ separately. This approach, also found in similar studies [43], enjoys computational benefits compared to batch methods, since it can yield faster reliable estimates of the considered variables (Section III-B) and can be easily parallelized to run on multiple computational threads (Section III-E).

In the case of parametric dictionaries, the atoms can be expressed with an appropriate domain-specific function $\phi : \Re^Q \to \Re^P$ in terms of a parameter vector $\boldsymbol{\theta_{nk}} \in \Re^Q, Q < P$, as $\mathbf{d_{nk}} = \phi(\boldsymbol{\theta_{nk}})$, therefore $\mathbf{D_n} = [\phi(\boldsymbol{\theta_{n1}}) \dots \phi(\boldsymbol{\theta_{nK}})]$. Typically dictionary atoms have a unitary $l2$-norm, i.e., $\|\phi(\boldsymbol{\theta_{nk}})\|_2 = 1$. Each signal $\mathbf{x_n}$ can be expressed as a linear combination of a small number of atoms, $L \ll K$, with additive noise $\boldsymbol{\epsilon_n} \in \Re^P$

$$\mathbf{x_n} = \mathbf{D_n c_n} + \boldsymbol{\epsilon_n} \tag{1}$$

where $\boldsymbol{\epsilon_n}$ is the error and $\mathbf{c_n} \in \Re^K, ||\mathbf{c_n}||_0 = L$, are the coefficients with non-zero values only for the used atoms. According to the Bayesian framework, each variable in (1) is assumed to follow an underlying probabilistic distribution.

Especially in parametric DL, where we jointly sample the parameters of the selected dictionary atoms (Section III-B), a small number of selected atoms can keep the implementation computationally tractable. For this reason, we are interested in imposing explicit sparseness constraints, similarly to previous studies [38], [39], [46] (Section I). We will describe two different ways to approach this using the Multinomial and the Wallenius' hypergeometric distribution, that allow sampling with and without replacement.

A. Atom Sampling With Replacement

A straightforward method to sample the dictionary atoms is to relax the $l0$-sparsity norm constraint into $||\mathbf{c_n}||_0 \leq L$ allowing independent sampling of the dictionary atoms L times with replacement through the Multinomial distribution. Since the population size is much larger than the sample size ($L \ll K$), duplicate atoms are rare [53]. If we assume a discrete multinomial distribution for selecting one dictionary atom out of the possible K, (1) can be re-written as

$$\mathbf{x_n} = \mathbf{D_n} \sum_{l=1}^{L} s_{nl} \mathbf{z_{nl}} + \boldsymbol{\epsilon_n} \tag{2}$$

where $\mathbf{z_{nl}} \in \bigcup \mathbf{u_i}, \mathbf{u_i} = [0, 0, \dots, 1, \dots, 0]^T$ with 1 in the ith entry ($i \leq K$) and 0 in the rest K-1 entries. The vector $||\mathbf{z_{nl}}||_0 = 1$ is binary activating one dictionary atom at a time and $\mathbf{s_n} = [s_{n1} \dots s_{nL}]^T \in \Re^L$ only contains the coefficients of the selected atoms. If atom $\mathbf{d_k}$ is the lth representation term, then $z_{nl_k} = 1, z_{nl_{k'}} = 0, \forall k' \neq k$, and s_{nl} consists the kth entry of vector $\mathbf{c_n}$ in (1), i.e., $s_{nl} = c_{nk}$. The probability of selecting atom k for data $\mathbf{x_n}$ is π_{nk} such that $\mathbf{z_{nl}} \sim \text{Multinomial}(1, \boldsymbol{\pi_n})$ with $\boldsymbol{\pi_n} = [\pi_{n1}, \dots, \pi_{nK}]^T \in \Re^K$. If the same atom is selected more than once, the corresponding coefficient is only once estimated.

B. Atom Sampling Without Replacement

Sampling without replacement avoids duplicate atoms and keeps the $l0$-sparsity constraint intact. The problem of selecting L atoms out of the possible K can be formalized similarly to the classical experiment of taking colored balls at random from an urn without replacement [54], [55]. If the balls have a different weight, the result follows the Wallenius' noncentral hypergeometric distribution [56].

In the considered problem, we can assume K dictionary atoms each of a different type (i.e., each ball in the urn has a different color) and selection probability $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$. If $\mathbf{z_n} = [z_{n1}, \dots, z_{nK}] \in \Re^K, z_{nk} \in \{0, 1\}$ and $\|\mathbf{z_n}\|_0 = L$, indicates the selected atoms for $\mathbf{x_n}$, then (1) becomes

$$\mathbf{x_n} = \mathbf{D_n}(\mathbf{s_n} \circ \mathbf{z_n}) + \boldsymbol{\epsilon_n} \tag{3}$$

where "$\circ$" represents the Hadamard or entrywise product and $\mathbf{s_n} \in \Re^K$ and $\mathbf{z_n}$ follows the Wallenius distribution.

C. Additional Probabilistic Assumptions

We assume independent atom coefficients $s_{n1}, \dots, s_{nL}$ each following a normal distribution with mean $\mu_{s_{nl}}$ and precision γ_s, i.e., $\mathbf{s_n} \sim \text{Normal}(\boldsymbol{\mu_{s_n}}, \gamma_s^{-1} \mathbf{I_L})$. We have considered a different mean $\mu_{s_{nl}}$ for each exemplar data n and selected dictionary atom l to account for the various signal energy levels and possible atom configurations. We further hypothesize dictionary parameters of normal distribution $\boldsymbol{\theta_{nk}} \sim \text{Normal}(\mathbf{g_{nk}}, \mathbf{G_n^{-1}})$ with mean $\mathbf{g_k}$ and precision $\mathbf{G_n}$. Finally, we assume zero mean Gaussian noise with variance $\gamma_{\epsilon_n}^{-1}$, i.e., $\boldsymbol{\epsilon_n} \sim \text{Normal}(\mathbf{0}, \gamma_{\epsilon_n}^{-1} \mathbf{I_Q})$. Similar distributions have been also hypothesized in prior work [16], [41], [44] (Section I).

The Bayesian framework further treats the parameters of the above variables as random components in order to better

TABLE I
PRIOR DISTRIBUTIONS OF BAYESIAN DICTIONARY LEARNING VARIABLES

Variable	Type	Expression	(Hyper) Parameters
$\mathbf{z_{nl}}^{\dagger} \in \bigcup \mathbf{u_i}$	Multinomial $(1, \boldsymbol{\pi_n})$	$\prod_{k=1}^{K} \pi_{nk}^{z_{nl_k}}$	$\boldsymbol{\pi_n}$: outcome probability
$\mathbf{z_n}^{\ddagger} \in \Re^K$	Wallenius $(\mathbf{1_K}, L, \boldsymbol{\pi_n})$	$\int_0^1 \prod_{k=1}^{K} \left(1 - t^{\pi_{nk}/d}\right)^{z_{nk}} dt,\ d = \sum_{k=1}^{K} \pi_{nk}(1 - z_{nk})$	
$\mathbf{s_n} \in \Re^L$	Normal $(\boldsymbol{\mu_{s_n}}, \gamma_s^{-1}\mathbf{I_L})$	$(2\pi)^{-L/2}\gamma_s^{1/2} \exp\left[-\frac{\gamma_s}{2}(\mathbf{s_n} - \boldsymbol{\mu_{s_n}})^T(\mathbf{s_n} - \boldsymbol{\mu_{s_n}})\right]$	$\boldsymbol{\mu_{s_n}}$: mean γ_s: precision
$\boldsymbol{\theta_{nk}} \in \Re^Q$	Normal $(\mathbf{g_{nk}}, \mathbf{G_n}^{-1})$	$(2\pi)^{-Q/2}\lvert\mathbf{G_n}\rvert^{1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta_{nk}} - \boldsymbol{g_{nk}})^T\mathbf{G_n}(\boldsymbol{\theta_{nk}} - \boldsymbol{g_{nk}})\right]$	$\boldsymbol{g_{nk}}$: mean $\mathbf{G_n}$: precision
$\boldsymbol{\epsilon_n} \in \Re^P$	Normal $(0, \gamma_{\epsilon_n}^{-1}\mathbf{I_P})$	$(2\pi)^{-P/2}\gamma_{\epsilon_n}^{1/2} \exp\left[-\frac{\gamma_{\epsilon_n}}{2}\boldsymbol{\epsilon_n}^T\boldsymbol{\epsilon_n}\right]$	γ_{ϵ_n}: precision
$\boldsymbol{\pi_n} \in \Re^K$	Dirichlet $(\boldsymbol{\alpha})$	$\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\sum_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{nk}^{\alpha_k - 1}$	$\boldsymbol{\alpha}$: concentration parameters
$\mathbf{g_{nk}} \in \Re^Q$	Normal $(\mathbf{g_0}, \mathbf{G_0}^{-1})$	$(2\pi)^{-Q/2}\lvert\mathbf{G_0}\rvert^{1/2} \exp\left[-\frac{1}{2}(\boldsymbol{g_{nk}} - \boldsymbol{g_0})^T\mathbf{G_0}(\boldsymbol{g_{nk}} - \boldsymbol{g_0})\right]$	$\boldsymbol{g_0}$: mean $\mathbf{G_0}$: precision
$\mathbf{G_n} \in \Re^{Q\times Q}$	Wishart $(\nu_0, \mathbf{R_0})$	$\left(\lvert\mathbf{G_n}\rvert^{\frac{\nu_0 - Q - 1}{2}} \exp\left[-\frac{tr(\mathbf{R_0}^{-1}\mathbf{G_n})}{2}\right]\right)\left(2^{\frac{Q\nu_0}{2}}\lvert\mathbf{R_0}\rvert^{\frac{\nu_0}{2}}\Gamma_Q(\frac{\nu_0}{2})\right)^{-1}$	$\nu_0 > Q - 1$: dof $\mathbf{R_0}$: scale matrix
$\gamma_{\epsilon_n} \in \Re$	Gamma (e, f)	$\frac{f^e}{\Gamma(e)}\gamma_{\epsilon_n}^{e-1} \exp\left[-f\gamma_{\epsilon_n}\right]$	e: shape f: rate

Γ, Γ_Q: (multivariate) gamma functions, dof: degrees of freedom, $\mathbf{1_K} = [1, \ldots, 1]^T \in \Re^K$
†: atom sampling with replacement ($\|\mathbf{z_{nl}}\|_0 = 1$), $\mathbf{u_i} = [0, 0, \ldots, 1, \ldots, 0]^T$ with 1 in the i^{th} entry, 0 otherwise
‡: atom sampling without replacement ($\|\mathbf{z_n}\|_0 = L$)

capture uncertainty. We introduce conjugate prior distributions that simplify computations. Specifically, we assume that $\boldsymbol{\pi_n}$ follows a Dirichlet prior, i.e., $\pi_{nk} \sim \text{Dirichlet}\,(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = [a_1 \ldots a_K]^T$, and the precision of the Gaussian noise follows a Gamma prior, i.e., $\gamma_{\epsilon_n} \sim \text{Gamma}\,(e, f)$. The mean vector $\mathbf{g_{nk}}$ and precision matrix $\mathbf{G_n}$ of the dictionary atom parameters are modeled with Gaussian and Wishart distributions, respectively, i.e., $\mathbf{g_{nk}} \sim \text{Normal}\,(\mathbf{g_0}, \mathbf{G_0^{-1}})$ and $\mathbf{G_n} \sim \text{Wishart}\,(\nu_0, \mathbf{R_0})$.

D. Objective

The goal is to find $\mathbf{y_n^*}$ such that

$$\mathbf{y_n^*} = \arg\max_{\mathbf{y_n}} p(\mathbf{y_n}|\mathbf{x_n}, \mathcal{H}_n) \tag{4}$$

$$\mathbf{y_n} = [\mathbf{z_n}^T, s_{n1}, \ldots, s_{nL}, \boldsymbol{\theta_{n1}}^T, \ldots, \boldsymbol{\theta_{n}}_K{}^T, \mathbf{g_{n1}}^T, \ldots, \mathbf{g_{nK}}^T, \text{vec}(\mathbf{G_n})^T, \boldsymbol{\epsilon_n}^T, \boldsymbol{\pi_n}^T, \gamma_{\epsilon_n}]^T \tag{5}$$

$$\mathcal{H}_n = \{\boldsymbol{\alpha}, \mathbf{g_0}, \mathbf{G_0}, \nu_0, \mathbf{R_0}, e, f, \boldsymbol{\mu_{s_n}}, \gamma_s\} \tag{6}$$

The probabilistic assumptions for (4)–(6) are summarized in Table I.

III. INFERENCE WITH MCMC SAMPLING

The inference problem aims at finding solutions $\mathbf{y_n^*}, n = 1 \ldots, N$, that maximize (4). Since (4) is not analytically tractable, we use MCMC for approximate inference because of its simple and fast implementation. We describe the inference procedure in Section III-A and provide the corresponding derivations in Sections III-B, III-C, and III-D.

A. MCMC Sampling

The large number of variables in our problem renders the simultaneous sampling from the full posterior quite prohibitive. For this reason, we divide the variable space $\mathbf{y}$ into blocks, a technique which is usually referred as "block-at-a-time" MCMC [57], [58]. Suppose that the variable space is split into B blocks specified according to the problem characteristics, i.e., $\mathbf{y} = [\mathbf{y_1}^T, \ldots, \mathbf{y_B}^T]^T$. Without loss of generality, we hypothesize these blocks are sampled sequentially from $\mathbf{y_1}$ to $\mathbf{y_B}$. This is referred as "deterministic-scan" MCMC sampling [59]–[61] and will be the focus of our paper.

When the posterior probability of a block b yields a known probabilistic distribution, we use the Gibbs sampler, otherwise we sample based on the MH [62], [63]. MH generates a sample with a candidate-generating (or proposal) density q. The MC transitions to the generated sample with a predefined probability of move. A critical component is the selection of proposal density, based on which MH samplers are divided into two categories. The first is the random-walk [62] with samples generated around the current value of the corresponding variables. Despite its simplicity, this method often depicts slow convergence depending on the variance of q [58], [64]. The second type, called independent MH [63], samples independently of the previous state. Its proposal density is close to the target distribution in a certain sense benefiting convergence. Previous work has considered Student-t distributions tailored to the target density [65]–[67], whose long tails ensure that no areas of the state space are left unexplored. The "MH-within-Gibbs" sampler uses MH to generate samples for blocks whose posterior does not yield a known distribution, and Gibbs to generate samples for the remaining blocks.

We will further discuss the use of "MH-within-Gibbs sampler" for our problem and the derivation of posteriors for each variable (Table II). We will assume that $\mathbf{y}_{-\mathbf{y_b}}$ contains all variables except the ones included in the bth block (i.e., $\mathbf{y_b}$), which are currently being generated.

B. Sampling Dictionary Parameters

Let $\mathcal{I}_{D_n} = \{k'_1, \ldots, k'_L\}$ be the indices of dictionary atoms that represent signal $\mathbf{x_n}$ and $\boldsymbol{\theta_{n}}_{k'_1} \ldots \boldsymbol{\theta_{n}}_{k'_L}$ the

TABLE II
DESCRIPTION OF METROPOLIS-HASTINGS-WITHIN-GIBBS SAMPLING DISTRIBUTIONS

Variable	Sampling Distribution/Proposal	Sampling Method
$\mathbf{z_{nl}}^{\dagger} \in \Re^K$	Multinomial $(1, \mathbf{p_{nl}})$ $p_{nl_k} = \pi_{nk} \exp\left[-\frac{1}{2}\gamma_{\epsilon_n} s_{nl}\left(s_{nl} - 2\boldsymbol{\epsilon}_{\mathbf{nl}}^T \phi(\boldsymbol{\theta_{nk}})\right)\right]$	Gibbs
$\mathbf{z_n}^{\ddagger} \in \Re^K$	Wallenius $(\mathbf{1_k}, L, \pi_{\mathbf{n}})$	Metropolis-Hastings
$s_{nl} \in \Re$	Normal $\left(\frac{\gamma_s \mu_{s_{nl}} + \gamma_{\epsilon_n} (\mathbf{D_n z_{nl}})^T \boldsymbol{\epsilon}_{\mathbf{nl}}}{\gamma_s + \gamma_{\epsilon_n}}, (\gamma_s + \gamma_{\epsilon_n})^{-1}\right)$	Gibbs
${}^{*}\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} = \text{vec}\left(\left[\boldsymbol{\theta}_{\boldsymbol{nk}'_1} \dots \boldsymbol{\theta}_{\boldsymbol{nk}'_L}\right]\right) \in \Re^{QL}$	Student-t $\left(\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}, \hat{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}, \nu_1\right)$	Metropolis-Hastings
${}^{\dagger}\boldsymbol{\theta}_{\boldsymbol{nk}} \in \Re^Q,\ k \notin \mathcal{I}_{D_n}$	Normal $(\mathbf{g_{nk}}, \mathbf{G_n})$	Gibbs
$\boldsymbol{\epsilon}_{\boldsymbol{n}} \in \Re^P$	Normal $(0, \gamma_{\epsilon_n}^{-1} \mathbf{I_P})$	Gibbs
$\boldsymbol{\pi}_{\boldsymbol{n}} \in \Re^K$	Dirichlet $\left(\left[\alpha_1 + \sum_{l=1}^L z_{nl_1}, \dots, \alpha_K + \sum_{l=1}^L z_{nl_K}\right]^T\right)$	Gibbs
$\mathbf{g_{nk}} \in \Re^Q$	Normal $\left((\mathbf{G_n} + \mathbf{G_0})^{-1}(\mathbf{G_n}^T \boldsymbol{\theta_{nk}} + \mathbf{G_0}^T \mathbf{g_0}), (\mathbf{G_n} + \mathbf{G_0})^{-1}\right)$	Gibbs
$\mathbf{G_n} \in \Re^{Q \times Q}$	Wishart $\left(\nu_0 + K, \left[\mathbf{R_0}^{-1} + \sum_{k=1}^K (\boldsymbol{\theta_{nk}} - \boldsymbol{g_{nk}})(\boldsymbol{\theta_{nk}} - \boldsymbol{g_{nk}})^T\right]^{-1}\right)$	Gibbs
$\gamma_{\epsilon_n} \in \Re$	Gamma $\left(e + \frac{1}{2}, f + \frac{1}{2}\left\|\mathbf{x_n} - \sum_{l=1}^L s_{nl}\phi(\boldsymbol{\theta}_{\boldsymbol{nk}'_l})\right\|^2\right)$	Gibbs

${}^{*}\mathbf{x_n} = \sum_{l=1}^L s_{nl}\phi(\boldsymbol{\theta}_{\boldsymbol{nk}'_l}) + \boldsymbol{\epsilon_n}$, ${}^{\dagger}\mathcal{I}_{D_n} = \{k'_1, \dots, k'_L\}$, $\mathbf{1_K} = [1, \dots, 1]^T \in \Re^K$
†, ‡: atom sampling with/without replacement

corresponding atom parameters. The joint posterior of $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} = \text{vec}\left(\left[\boldsymbol{\theta}_{\boldsymbol{n}k'_1} \dots \boldsymbol{\theta}_{\boldsymbol{n}k'_{\boldsymbol{L}}}\right]\right) \in \Re^{QL}$ can be written as

$$
\begin{aligned}
&p\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} | \mathbf{y}_{-\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}, \mathbf{x_n}, \mathcal{H}_n\right) \\
&\quad \propto \prod_{k \in \mathcal{I}_{D_n}} p\left(\boldsymbol{\theta}_{\boldsymbol{n}k} | \mathbf{g_{nk}}, \mathbf{G_n}\right) \cdot p\left(\mathbf{x_n} - \sum_{l=1}^L s_{nl}\phi(\boldsymbol{\theta}_{\boldsymbol{n}k'_l}) \middle| \gamma_{\epsilon_n}\right) \\
&\quad \propto \exp\left[-\frac{\gamma_{\epsilon_n}}{2}\left\|\mathbf{x_n} - \sum_{l=1}^L s_{nl}\phi(\boldsymbol{\theta}_{\boldsymbol{n}k'_l})\right\|_2^2\right] \\
&\qquad \times \exp\left[-\frac{1}{2}\sum_{k \in \mathcal{I}_{D_n}} (\boldsymbol{\theta}_{\boldsymbol{n}k} - \mathbf{g_{nk}})^T \mathbf{G_n} (\boldsymbol{\theta}_{\boldsymbol{n}k} - \mathbf{g_{nk}})\right] \\
&\quad \propto \exp\left[-\frac{\gamma_{\epsilon_n}}{2}\left\|\mathbf{x_n} - \mathbf{S_n}\tilde{\boldsymbol{\phi}}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right\|_2^2\right] \\
&\qquad \times \exp\left[-\frac{1}{2}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \tilde{\mathbf{g}}_{\mathbf{n}})^T \tilde{\mathbf{G}}_{\mathbf{n}} (\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \tilde{\mathbf{g}}_{\mathbf{n}})\right] \qquad (7)
\end{aligned}
$$

where $\tilde{\boldsymbol{\phi}}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) = \text{vec}\left(\left[\phi(\boldsymbol{\theta}_{\boldsymbol{n}k'_1}) \dots \phi(\boldsymbol{\theta}_{\boldsymbol{n}k'_{\boldsymbol{L}}})\right]\right) \in \Re^{PL}$, $\tilde{\mathbf{g}}_n = \text{vec}\left(\left[\mathbf{g}_{\mathbf{nk}'_1} \dots \mathbf{g}_{\mathbf{nk}'_{\mathbf{L}}}\right]\right) \in \Re^{QL}$ and

$$
\tilde{\mathbf{G}}_{\mathbf{n}} = \begin{pmatrix} \mathbf{G_n} & & \\ & \ddots & \\ & & \mathbf{G_n} \end{pmatrix} \in \Re^{QL \times QL}
$$

$$
\mathbf{S_n} = \left(\begin{array}{ccc|c|ccc} s_{n1} & & & & s_{nL} & & \\ & \ddots & & \ddots & & \ddots & \\ & & s_{n1} & & & & s_{nL} \end{array}\right) \in \Re^{P \times PL}
$$

Because of ϕ, we cannot find a known probabilistic distribution for (7), therefore we use MH for sampling $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$. The proposal density is a multivariate Student-t distribution with location $\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}$ tailored to the target density with ν_1 degrees of freedom and identity scale matrix $\hat{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}$, defined as

$$
\begin{aligned}
q(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} | \mathbf{x_n}, \mathcal{H}_n, \mathbf{y}_{-\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}) &= \frac{\Gamma\left(\frac{\nu_1 + Q}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)} \left[\pi(\nu_1 + Q)\right]^{-\frac{Q}{2}} \left|\hat{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\right|^{-\frac{1}{2}} \\
&\quad \times \left[1 + \frac{1}{\nu_1 + Q}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\right)^T \hat{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}^{-1}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\right)\right]^{-\frac{\nu_1 + Q}{2}} \qquad (8)
\end{aligned}
$$

$$
\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}} = \arg\max_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}} \ln\left(p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} | \mathbf{y}_{-\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}, \mathbf{x_n}, \mathcal{H}_n)\right) \qquad (9)
$$

$$
\hat{\boldsymbol{V}}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}} = \tau^2 \mathbf{I_{QL}} \qquad (10)
$$

The choice of Student-t is crucial for the uniform ergodicity of MCMC (Section VI). In the experiments (Section VII), we find a numerical solution to (9) using the Nelder Mead's simplex [68] because of its simplicity compared to gradient-based approaches. From (8)–(9), we are able to jointly generate the parameters of the selected atoms, rather than generating each one separately. A large number of selected atoms increases the computational cost towards maximizing (9).

The posterior for the remaining atoms $\boldsymbol{\theta}_{\boldsymbol{n}k} \notin \mathcal{I}_{D_n}$, generated with Gibbs, is

$$
p\left(\boldsymbol{\theta}_{\boldsymbol{n}k} | \mathbf{y}_{-\boldsymbol{\theta}_{\boldsymbol{n}k}}, \mathbf{x_n}, \mathcal{H}_n\right) \propto p\left(\boldsymbol{\theta}_{\boldsymbol{n}k} | \mathbf{g_{nk}}, \mathbf{G_n}\right) \qquad (11)
$$

C. Sampling Atom Indices and Coefficients

The probabilistic selection of dictionary atoms gives the opportunity to test unseen combinations. This can alleviate disadvantages of greedy algorithms, since the randomization might overcome locally optimal solutions, as also observed in [39]. Because the atom indices and coefficients are interdependent, we will describe the sampling procedure for both.

In the case of sampling with replacement, the atom indices and coefficients are generated with the Gibbs sampler, as their posterior can be analytically derived based on the assumptions for their priors (Section II). In the following, we will denote $\boldsymbol{\epsilon_{nl}} = \mathbf{x_n} - \sum_{l' \neq l} s_{nl'} \mathbf{D_n z_{nl'}}$ the error for the exemplar data $\mathbf{x_n}$ when we exclude the lth atom from the representation. Then the total representation error can be expressed as

$$\boldsymbol{\epsilon_n} = \boldsymbol{\epsilon_{nl}} - s_{nl}\mathbf{D_n z_{nl}} = \boldsymbol{\epsilon_{nl}} - s_{nl}\sum_{k=1}^{K} z_{nl_k}\phi(\boldsymbol{\theta_{nk}}) \tag{12}$$

The posterior distribution of $\mathbf{z_{nl}}$ can be written as

$$\begin{aligned} p(\mathbf{z_{nl}}|\mathbf{y}_{-\{\mathbf{z_{nl}}\}}, \mathbf{x_n}, \mathcal{H}_n) &\propto p(\mathbf{z_{nl}}|\boldsymbol{\pi_n}, L)\cdot p(\boldsymbol{\epsilon_n}|\gamma_{\epsilon_n}) \\ &\propto \prod_{k=1}^{K} \pi_{nk}^{z_{nl_k}} \cdot \exp\left(-\frac{\gamma_{\epsilon_n}}{2}\left\|\boldsymbol{\epsilon_{nl}} - s_{nl}\sum_{k=1}^{K} z_{nl_k}\phi(\boldsymbol{\theta_{nk}})\right\|_2^2\right) \\ &\propto \prod_{k=1}^{K}\left[\pi_{nk}\exp\left(\gamma_{\epsilon_n} s_{nl}\boldsymbol{\epsilon}_{\mathbf{nl}}^T\phi(\boldsymbol{\theta_{nk}})\right)\right]^{z_{nl_k}} \end{aligned} \tag{13}$$

For deriving the above expression we took into account that $z_{nl_k} \in \{0,1\}$ is a binary variable, implying that $z_{nl_k}^2 = z_{nl_k}$ and $a z_{nl_k} = a^{z_{nl_k}}, \forall a \in \Re$. Also $\mathbf{z_{nl}}$ has unit $l0$-norm, i.e., $\|\mathbf{z_{nl}}\|_0 = 1$, resulting in $z_{nl_k} z_{nl_{k'}} = 0, \forall k' \neq k$. As indicated in (13), this update procedure considers the similarity of dictionary atoms to the signal residual and the prior knowledge for selecting atom k. Finally, the posterior of s_{nl} is

$$\begin{aligned} p(s_{nl}|\mathbf{y}_{-s_{nl}}, \mathbf{x_n}, \mathcal{H}_n) &\propto p(s_{nl}|\gamma_s)\cdot p(\boldsymbol{\epsilon_n}|\gamma_{\epsilon_n}) \\ &\propto \exp\left[-\frac{\gamma_s}{2}\left(s_{nl}-\mu_{s_{nl}}\right)^2 - \frac{\gamma_{\epsilon_n}}{2}\left\|\boldsymbol{\epsilon_{nl}} - s_{nl}\mathbf{D_n z_{nl}}\right\|_2^2\right] \end{aligned} \tag{14}$$

By completing the square of the above quadratic formula with respect to s_{nl} and assuming dictionary atoms of unit norm, s_{nl} can be generated from a normal distribution (Table II).

To the best of our knowledge, we found no conjugate prior for the Wallenius' hypergeometric distribution [53], therefore we used the independent MH with proposal distribution tailored to the Wallenius prior for sampling without replacement.

D. Sampling the Parameters of the Priors

The conjugate assumptions for the distribution of the parameters of the aforementioned variables allow us to use the Gibbs sampler to generate the corresponding samples. For the sake of completeness, we briefly sketch the derivation of the posteriors with sampling distributions shown in Table II.

$$p(\boldsymbol{\pi_n}|\mathbf{y}_{-\boldsymbol{\pi_n}}, -)) \propto p(\boldsymbol{\pi_n}|\boldsymbol{\alpha})\prod_{l=1}^{L} p(\mathbf{z_{nl}}|\boldsymbol{\pi_n}, \boldsymbol{\alpha}) \tag{15}$$

$$p(\mathbf{g_{nk}}|\mathbf{y}_{-\mathbf{g_{nk}}}, -) \propto p(\boldsymbol{\theta_{nk}}|\mathbf{g_{nk}}, \mathbf{G_n})\cdot p(\mathbf{g_{nk}}|\mathbf{g_0}, \mathbf{G_0}) \tag{16}$$

$$p(\mathbf{G_n}|\mathbf{y}_{-\mathbf{G_n}}, -) \propto \prod_{k=1}^{K} p(\boldsymbol{\theta_{nk}}|\mathbf{g_{nk}}, \mathbf{G_n})\cdot p(\mathbf{G_n}|\nu_0, \mathbf{R_0}) \tag{17}$$

$$p(\gamma_{\epsilon_n}|\mathbf{y}_{-\gamma_{\epsilon_n}}, -) \propto p(\gamma_{\epsilon_n}|e, f)\cdot p\left(\mathbf{x_n} - \mathbf{D_n}\sum_{l=1}^{L} s_{nl}\mathbf{z_{nl}}\middle|\gamma_{\epsilon_n}\right) \tag{18}$$

where the dash "$-$" denotes $\{\mathbf{x_n}, \mathcal{H}_n\}$.

E. Implementation of Bayesian DL

The problem variables (5) are inferred for each exemplar data separately. This method can yield signal-specific estimations of the noise variance, useful for denoising applications. It also allows sharing computational cost in MCMC inference, which typically requires a large amount of iterations to converge. Since more training samples are likely to require more MCMC iterations, if we had trained one dictionary on the entire dataset, the sequential nature of MCMC would have significantly slowed down convergence. It is worth mentioning that similar studies have acknowledged the difficulty of performing Bayesian inference in large datasets [43]. The MCMC inference procedure is outlined in Algorithm 1.

Algorithm 1: MCMC inference of parametric dictionary learning variables

Require: Data $\mathbf{x_n}$, hyperparameters $\mathcal{H}_n$
1: **for** $n = 1, \ldots, N$ **do**
2: **for** $m = 1, \ldots, M$ **do**
3: Sample atom indices $\mathbf{z_{n1}^{(m)}}, \ldots \mathbf{z_{nL}^{(m)}}$ or $\mathbf{z_n^{(m)}}$
4: Find $\mathcal{I}_{D_n}^{(m)} = \{k_1', \ldots, k_L'\}$ s.t. $z_{nl_l'}^{(m)} = 1$
5: Sample atom coefficients $s_{n1}^{(m)}, \ldots, s_{nL}^{(m)}$
6: Sample noise vector $\boldsymbol{\epsilon}_{\mathbf{n}}^{(m)}$
7: Sample dictionary parameters $\boldsymbol{\theta}_{\mathbf{nk}}^{(m)}, k \notin \mathcal{I}_{D_n}^{(m)}$
8: Sample dictionary priors $\mathbf{g_{n1}^{(m)}}, \ldots, \mathbf{g_{nK}^{(m)}}, \mathbf{G_n^{(m)}}$
9: Sample atom selection probability priors $\boldsymbol{\pi}_{\mathbf{n}}^{(m)}$
10: Sample noise variance $\gamma_{\epsilon_n}^{(m)}$
11: Sample $\tilde{\boldsymbol{\theta}}_{\mathbf{n}}^{(m)}$ for generating $\boldsymbol{\theta}_{\mathbf{nk}}^{(m)}, k \in \mathcal{I}_{D_n}^{(m)}$
12: Compute $\mathbf{D}_{\mathbf{n}}^{(m)} = \left[\phi\left(\boldsymbol{\theta}_{\mathbf{n1}}^{(m)}\right)\ldots\phi\left(\boldsymbol{\theta}_{\mathbf{nK}}^{(m)}\right)\right]$
13: **end for**
14: **end for**

IV. Combination of Generated Dictionaries

The variables of the considered Bayesian problem are inferred for each exemplar data separately yielding sample-specific information useful for denoising and other applications, and allowing parallel implementations (Sections II, III). The latter is beneficial because of the large amount of data in many applications, that render batch DL methods computationally expensive or even prohibitive. However, in order to obtain generalizable dictionaries, that are able to reliably represent unseen data, we need to combine the corresponding results into a unified model (Algorithm 2). While the Bayesian framework aims to maximize the posterior probability of the model parameters, DL is usually evaluated based on the reconstruction error. For this reason, the combination of the dictionaries that result from the Bayesian inference is performed based on a root mean square (RMS) error criterion.

Let $\mathbf{D}_{\mathbf{n}}^{(m)}$ be the dictionaries generated with the MH-within-Gibbs sampler (Section III-E). In order to evaluate their performance using signal reconstruction criteria, we need to use a sparse decomposition algorithm, that can represent the original signal based on the inferred dictionaries. We use OMP [3], [4] because of its simplicity and effectiveness. Let $\mathrm{Err}_n^{(m)}$ be the relative RMS error that yields from decomposing $\mathbf{x_n}$ based on dictionary $\mathbf{D}_{\mathbf{n}}^{(m)}$. Also let $\mathbf{D}_{\mathbf{n}}^{(m_n^*)} \in \Re^{P\times K}$ be the dictionaries

that yield the lowest relative RMS error for each signal n and $\boldsymbol{\Theta}^{(m_n^*)} = \left[\boldsymbol{\theta}_{\mathbf{nk_1'}}^{(m_n^*)}, \ldots, \boldsymbol{\theta}_{\mathbf{nk_L'}}^{(m_n^*)}\right] \in \Re^{Q\times L}$ the parameters of the atoms that were selected by OMP based on $\mathbf{D}_{\mathbf{n}}^{(m_n^*)}$. $\boldsymbol{\Theta}^{(m_n^*)}$ are concatenated into a unified dictionary

$$\boldsymbol{\Theta_U} = \left[\boldsymbol{\Theta}^{(m_1^*)} \ldots \boldsymbol{\Theta}^{(m_N^*)}\right] \in \Re^{Q\times NL}$$
$$m_n^* = \arg\min_{m\in[1,M]} \mathrm{Err}_n^{(m)}, n = 1, \ldots, N$$

Ideally, we could have used $\boldsymbol{\Theta_U}$ as the parameters of our final dictionary. However, in practice the large amount of data renders $\boldsymbol{\Theta_U}$ computationally expensive. For this reason, the parameter vectors of $\boldsymbol{\Theta_U}$ are further quantized with K-means clustering using N_{bin} centers. This results in parameter matrix $\boldsymbol{\Theta_Q}$ with corresponding final dictionary $\mathbf{D_Q}$. The aforementioned procedure is performed on the training data, while the test data are not seen at all during this step (Section VII-B5).

This approach can be applied to combine dictionaries learnt from any other method, therefore it also provides a unified platform that allows comparison of the considered parametric DL approaches in our paper (Section VII).

Algorithm 2: Combination of generated dictionaries

Require: Data $\mathbf{x_n}$, generated dictionaries $\mathbf{D}_{\mathbf{n}}^{(m)}$, number of K-means clusters N_b

1: **for** $n = 1, \ldots, N$ **do**
2: **for** $m = 1, \ldots, M$ **do**
3: Reconstruct $\mathbf{x_n}$ based on $\mathbf{D}_{\mathbf{n}}^{(m)}$ with OMP
4: Compute relative RMS error $Err_n^{(m)}$
5: **end for**
6: Find m_n^* such that $m_n^* = \arg\min_{m=[1,M]} Err_n^{(m)}$
7: Retrieve parameters of dictionary atoms selected by OMP based on $\mathbf{D}_{\mathbf{n}}^{(m_n^*)}, \boldsymbol{\Theta}^{(m_n^*)} = \left[\boldsymbol{\theta}_{\mathbf{nk_1'}}^{(m_n^*)}, \ldots, \boldsymbol{\theta}_{\mathbf{nk_L'}}^{(m_n^*)}\right]$
8: **end for**
9: Concatenate $\boldsymbol{\Theta_U} = \left[\boldsymbol{\Theta}^{(m_1^*)} \ldots \boldsymbol{\Theta}^{(m_N^*)}\right]$
10: Quantize dictionary parameters $\boldsymbol{\Theta_Q} = K\text{-}means(\boldsymbol{\Theta_U})$
11: Compute the final dictionary $\mathbf{D_Q} = \phi(\boldsymbol{\Theta_Q})$ where $\boldsymbol{\phi}$ operates on each column of matrix $\boldsymbol{\Theta_Q}$

V. Choosing the Parametric Dictionary Function

The choice of the parametric dictionary function is not trivial and is usually guided by the application of interest. If designed appropriately, parametric dictionaries can yield interpretable information about meaningful signal characteristics for a variety of applications. Previous studies have used Gabor [69] and Gammatone [24] atoms to represent speech signals because of their good localization properties and similarities to the human auditory system. Other efforts have proposed Gaussian-like functions to efficiently capture spherical stereo images [70], diffusion-based dictionaries to model MRI [25], and other wavelet-like atoms for digitizing fingerprint images [71]. Gabor dictionaries have been used for the electroencephalogram (EEG) [72], spline wavelets for the electrocardiogram (ECG) [73], and sigmoid-exponential functions for the electrodermal activity (EDA) [28].

Our Bayesian parametric DL approach generates the parameters of the dictionary atoms based on the MH sampler (Section III-B). The mean of the corresponding distribution depends on the considered parametric function and is selected so that it maximizes the corresponding posterior distribution (7). We will further show that the concavity of (7) generally depends on function ϕ. It is unusual that a function ϕ is concave with respect to all the parameters of interest, but in practice even the estimation of local optima is enough to generate useful samples (Section VII-C).

The posterior (7) of the dictionary atom parameters is:

$$\begin{aligned} U(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) &= \ln\left(p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}|\mathbf{y}_{\_\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}, \mathbf{X}, \mathcal{H})\right) = \\ &\propto -\frac{\gamma_{\epsilon_n}}{2}\left\|\mathbf{x_n} - \mathbf{S_n}\tilde{\boldsymbol{\phi}}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right\|_2^2 \\ &\quad - \frac{1}{2}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \tilde{\mathbf{g}}_{\mathbf{n}})^T\tilde{\mathbf{G}}_{\mathbf{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \tilde{\mathbf{g}}_{\mathbf{n}}) \end{aligned} \quad (19)$$

If we set $\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) = \mathbf{x_n} - \mathbf{S_n}\tilde{\boldsymbol{\phi}}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})$, then (19) becomes

$$U(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) = -\frac{1}{2}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \tilde{\mathbf{g}}_{\mathbf{n}})^T\tilde{\mathbf{G}}_{\mathbf{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \tilde{\mathbf{g}}_{\mathbf{n}}) - \frac{\gamma_{\epsilon_n}}{2}\left\|\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right\|^2 \quad (20)$$

The gradient vector and Hessian matrix of (20) are

$$\begin{aligned} \nabla_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}} U(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) &= -\tilde{\mathbf{G}}_{\mathbf{n}}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} - \tilde{\mathbf{g}}_{\mathbf{n}}\right) \\ &\quad - \gamma_{\epsilon_n}\left(\nabla_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right)^T\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}) \end{aligned} \quad (21)$$

$$\mathbf{H}_U = \nabla^2_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}} U(\tilde{\theta}_{\mathbf{n}}) = -\tilde{\mathbf{G}}_{\mathbf{n}} - \frac{\gamma_{\epsilon_n}}{2}\nabla^2_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\left\|\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right\|_2^2 \quad (22)$$

where

$$\begin{aligned} \left(\nabla^2_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\left\|\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right\|_2^2\right)_{ij} &= \frac{\vartheta^2}{\vartheta\tilde{\theta}_{n_i}\vartheta\tilde{\theta}_{n_j}}\left\|\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right\|_2^2 \\ &= \frac{\vartheta}{\vartheta\tilde{\theta}_{n_i}}\left(2\sum_{p=1}^{PL}\psi_{n_p}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\frac{\vartheta\psi_{n_p}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})}{\vartheta\tilde{\theta}_{n_j}}\right) \\ &= 2\sum_{p=1}^{PL}\frac{\vartheta\psi_{n_p}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})}{\vartheta\tilde{\theta}_{n_i}}\frac{\vartheta\psi_{n_p}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})}{\vartheta\tilde{\theta}_{n_j}} \\ &\quad + 2\sum_{p=1}^{PL}\psi_{n_p}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\frac{\vartheta^2\psi_{n_p}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})}{\vartheta\tilde{\theta}_{n_i}\vartheta\tilde{\theta}_{n_j}} \end{aligned} \quad (23)$$

If $\mathbf{H}_{\psi_{n_p}}$ is the Hessian of $\psi_{n_p}, p = 1, \ldots, PL$, then from (23)

$$\begin{aligned} \nabla^2_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\left\|\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right\|_2^2 &= 2\left(\nabla_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right)^T\left(\nabla_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right) \\ &\quad + 2\textstyle\sum_{p=1}^{P}\psi_{n_p}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\mathbf{H}_{\psi_{n_p}} \end{aligned} \quad (24)$$

The Hessian of ψ_{n_p} can be computed as

$$\left(\mathbf{H}_{\psi_{n_p}}\right)_{ij} = -\sum_{l=1}^{L} s_{nl}\frac{\vartheta^2\phi_p(\boldsymbol{\theta}_{nk_l'})}{\vartheta\theta_{nk'_{li}}\vartheta\theta_{nk'_{ll}}} \quad (25)$$

$$\mathbf{H}_{\psi_{n_p}} = -\sum_{l=1}^{L} s_{nl}\mathbf{H}_{\phi_p}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{n}k_{\boldsymbol{l}}'}} \quad (26)$$

From (22), (24) and (26), we get

$$\begin{aligned} \mathbf{H}_U &= -\tilde{\mathbf{G}}_{\mathbf{n}} - \gamma_{\epsilon_n}\left(\nabla_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right)\left(\nabla_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}}\boldsymbol{\psi}_{\boldsymbol{n}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\right)^T \\ &\quad + \gamma_{\epsilon_n}\sum_{p=1}^{P}\psi_{n_p}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})\sum_{l=1}^{L}s_{nl}\mathbf{H}_{\phi_p}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{n}k_{\boldsymbol{l}}'}} \end{aligned} \quad (27)$$

The first two terms of (27) are postive-definite matrices, whereas the positive-definiteness of the third term depends on ϕ. In our experiments (Section VII) and in the literature mentioned in this section, function ϕ is selected with a knowledge-driven approach based on the application of interest. Although this does not guarantee that the generated atom parameters are the global maxima, in practice the corresponding sampling method yields satisfactory results (Section VII-C).

VI. MCMC Convergence for Bayesian DL

Ergodicity properties of large-dimensional MC have been the subject of several studies [60], [74]. We discuss the uniform ergodicity property of the considered MC which implies convergence independently of the initial state [51], [52]. We focus on the sampling with replacement case, although our discussion can be extended for atom sampling without replacement. We show that the MC used for inferring the variables of the Bayesian DL problem is uniformly ergodic by proving that it meets the conditions of Theorem 6.1. This theorem establishes uniform ergodicity for the MH-within-Gibbs sampler and its proof can be found in [61].

Let P be a Markov transition kernel (Section III-A) and $\mathbf{Y}^{(m)}, \mathbf{Y}^{(m+1)}$ be two consecutive MCMC states generating observations $\mathbf{y}^{(m)}, \mathbf{y}^{(m+1)}$. We will assume that $\mathbf{y}_{\mathbf{b}}^{(m)}$ and $\mathbf{y}_{\mathbf{b}}^{(m+1)}$ are the values of block b at the current and previous MCMC state, noted as $m+1$ and m, respectively. Also, $\mathbf{y}_{-\mathbf{y_b}}^{(m+1)}$ contains all variables that have already been sampled at state $m+1$ and $\mathbf{y}_{-\mathbf{y_b}}^{(m)}$ the ones from the previous state m.

Definition 6.1 (Conditional Weight Function): If block b generated with the MH sampler, its conditional weight function is defined as

$$w_{\mathbf{y_b}}(\mathbf{y_b}^{(m+1)}|\mathbf{y_b}^{(m)}, \mathbf{y}_{-\mathbf{y_b}}{}^{(m)}, \mathbf{y}_{-\mathbf{y_b}}{}^{(m+1)}) = \frac{p\left(\mathbf{y_b}^{(m+1)} \left| \mathbf{y}_{-\mathbf{y_b}}{}^{(m)}, \mathbf{y}_{-\mathbf{y_b}}{}^{(m+1)} \right.\right)}{q\left(\mathbf{y_b}^{(m+1)} \left| \mathbf{y_b}^{(m)}, \mathbf{y}_{-\mathbf{y_b}}{}^{(m)}, \mathbf{y}_{-\mathbf{y_b}}{}^{(m+1)} \right.\right)} \quad (28)$$

Theorem 6.1 (MH-Within-Gibbs Uniform Ergodicity): Given $p(\mathbf{y}|\mathbf{X}, \mathcal{H})$ on state space $\mathbf{y} = [\mathbf{y_1}^T \dots \mathbf{y_B}^T]^T$, let $P = P_1 \dots P_B$ be the Markov kernel of the Gibbs sampler and $Q_b, b \in \mathcal{I}_{B'}$, where $\mathcal{I}_{B'} = \{b_{i_1}, \dots, b_{i_{B'}}\}, B' < B$, be the Markov kernel of the MH sampler with conditional weight $w_{\mathbf{y_b}'}$ as in (28). If each conditional weight $w_{\mathbf{y_b}'}, b' \in \mathcal{I}_{B'}$, of the MH sampler is bounded, $\sup w_{\mathbf{y_b}'} < \infty$, and the Gibbs sampler with Markov kernel $P_1 \dots P_B$ is uniformly ergodic, then the MH-within-Gibbs sampler, resulting from substituting kernels $P_{b'}$ with $Q_{b'}, b' \in \mathcal{I}_{B'}$, is also uniformly ergodic.

Based on Theorem 6.1, we show the MCMC uniform ergodicity for our case. Theorem 6.2 describes the minorization condition of the Gibbs sampler. This consists a multivariate extension from Jones *et al.* [61] (Proposition 2). Lemma 6.1 assists with its proof and ensures the minorization of the first $B-1$ blocks. Lemma 6.2 ensures that the conditional weight for $\boldsymbol{\theta_k}$ is bounded away from infinity. Finally, Theorem 6.3 combines all the aforementioned to prove the uniform ergodicity of the considered MC. The proofs of the theorems 6.2, 6.3 and lemmas 6.1, 6.2 can be found in Appendix A

Lemma 6.1 (Minorization Condition of $B-1$ Blocks): Let $P((\mathbf{Y_1}, \mathbf{Y_2}, \dots, \mathbf{Y_B}), A_1 \times A_2 \times \dots \times A_B)$ be the Markov kernel of a Gibbs sampler, where $A_1, \dots, A_B$ are elements of the Borel σ-algebra on the variable space $\mathbf{Y_1}, \dots, \mathbf{Y_B}$, respectively. We further assume that all updates of the Gibbs sampler, except the one corresponding to $\mathbf{Y_B}$, are minorisable, in the sense that for $b = 1, \dots, B-1$, there is $\epsilon_b > 0$ and a probability measure ν_b, such that $P_{\mathbf{Y_b}}(\mathbf{y}_{-\mathbf{y_b}}, A_{-b}) \geq \epsilon_b \nu_b(A_{-b})$, where $A_{-b} = A_1 \times \dots \times A_{b-1} \times A_{b+1}, \dots, A_B$. The Gibbs sampler with Markov kernel $P_1 \dots P_{B-1}$ is minorisable, i.e., there exist $\epsilon_0 > 0$ and probability measure ν_0 such that

$$P((\mathbf{Y_1}, \mathbf{Y_2}, \dots, \mathbf{Y_{B-1}}), A_{-B}) \geq \epsilon_0 \nu_0(A_{-B})$$

Theorem 6.2 (Partial Minorization Condition for Gibbs): Let the same assumptions from Lemma 6.1 hold, then the Gibbs sampler with Markov kernel $P_1 \dots P_B$ is minorisable.

Lemma 6.2 (Bounded Conditional Weight of Dictionary Parameters): Let the same assumptions from Lemma 6.1 hold. The conditional weight of the Bth block that includes the "super-vector" $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ of dictionary parameters and is generated with MH according to (8)–(10), is bounded, i.e., $\sup w_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}} < \infty$.

Theorem 6.3 (MC Uniform Ergodicity for Bayesian DL Inference): Let $\mathbf{y_n}$ be the variables of the parametric DL problem (4) generated with the MH-within-Gibbs sampler (Algorithm 1), according to which all variables except $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ are sampled with Gibbs (first $B-1$ blocks), while $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ is sampled with MH (Bth block). Then the corresponding MC $\{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \dots\}$ is uniformly ergodic.

VII. Experiments

We compare the Bayesian DL model against the previously proposed parametric SD [22] and ETF [24], which are conceptually closer to our approach. We further perform experiments with the non-parametric K-SVD [14] yielding dictionaries of arbitrary structure, therefore not directly comparable to our approach. We use synthetic and real biomedical data from EDA signals. Because of their characteristic structure, these types of signals favor sparse representation approaches with parametric dictionaries providing interpretable information [28].

EDA is decomposed into a slow moving tonic part depicting the general trend and a phasic part which contains fast fluctuations superimposed onto the tonic signal, also called skin conductance responses (SCR). The tonic part is mathematically expressed as a straight line, while SCRs are represented by sigmoid-exponential functions with a steep rise and slow recovery. Taking this into account, dictionaries contain tonic and phasic atoms as shown in Table III. Since SCR shapes typically contain higher variability than the signal level, which remains fairly constant throughout an analysis window, for the sake of simplicity we perform DL on the phasic atoms for learning the parameters $\boldsymbol{\theta} = [t_0, T_{\text{rise}}, T_{\text{decay}}]^T$. The initial dictionaries are created from the combination of all parameters reported in Table III, resulting in 63 tonic and 144 phasic atoms. The analysis window is 5 sec, i.e., 160 samples with typical sampling frequency of 32 Hz [28].

A critical issue of MCMC is whether a certain number of iterations is enough to stop sampling. In high-dimensional problems, all inferred variables need to converge to the target distribution. Since examining each variable separately is not always feasible, we use a combination of monitoring and diagnostic strategies to quantitatively assess MCMC convergence.

TABLE III
DESCRIPTION OF EDA-SPECIFIC DICTIONARY ATOMS AND INITIAL PARAMETERS

Tonic Atoms	
$\phi_1(t) = \Delta_0 + \Delta \cdot t$	$\Delta_0 \in \{-20, -10, 1\}$ $\Delta \in \{-0.01, -0.009, \ldots, 0, 0.01, 0.02, \ldots, 0.1\}$
Phasic Atoms	
$\phi_2(t) = \frac{e^{\frac{st-t_0}{T_{decay}}}}{\left[1+\left(\frac{st-t_0}{T_{rise}}\right)^{-2}\right]^2} u(t-t_0)$	$T_{rise} \in \{8, 14, 18\}$ $T_{decay} \in \{10, 15, 20\}$ $t_0 \in \{0, 10, 20, \ldots, 150\}$
$u(t) = 1$, $t \geq 0$ and $u(t) = 0$ otherwise	

In the following, we describe the data (Section VII-A), the experimental setup (Section VII-B), and the results evaluating the learned dictionaries (Section VII-C) and providing diagnostics for MCMC convergence (Section VII-D).

A. Data Description

1) Synthetic Data: We randomly generate 1000 synthetic signals that simulate the EDA structure. Each signal is expressed as the sum of a constant c and R number of SCRs

$$x(t) = c + \sum_{r=1}^{R} \phi_2^{(r)}(t) \tag{29}$$

where $\phi_2^{(r)}$ is given in Table III with parameters $T_{\text{rise}}^{(r)}, T_{\text{decay}}^{(r)}$ and $t_0^{(r)}$. In contrast to the rest of the paper, in which superscript "$(\cdot)$" denotes the MCMC state, here it symbolizes the SCR index. The parameters of the synthetic data are randomly generated within a pre-specified range $t_0^{(r)} \in [1, 150]$ samples, $T_{\text{rise}}, ^{(r)}, T_{\text{decay}}^{(r)} \in [1, 20]$, and $R \in [1, 5]$. Since the number of SCRs for each signal is known a priori, DL was performed with $K = R+1$ number of atoms, from which one captures the tonic part and the rest, the phasic.

2) Real Data: We further evaluate the considered DL methods on human EDA data from the database of emotion analysis using physiological signals (DEAP) [75]. DEAP contains 40 one-minute recordings from 32 subjects watching long excerpts of music videos designed to study the relation of multimedia content with mood and temperament [75]. Because of the expected SCR rate in the considered 5sec analysis window [76], training was performed with $K = 3$ atoms, although similar results yield for different values.

B. Experimental Setup

1) Bayesian DL: The proposed MCMC inference (Algorithm 1) is performed with 1000 and 500 iterations for the synthetic and real data, respectively. The atom indices $\mathbf{z_{nl}}$, coefficients $\mathbf{s_n}$ and selection probabilities $\boldsymbol{\pi_n}$ are initialized based on the decomposition of each exemplar signal with OMP. The mean $\boldsymbol{\mu_{s_n}}$ of the coefficients' prior was initialized with the average of the coefficients of the selected atoms from OMP. The mean and precision of dictionary atoms' prior $\mathbf{g_0}$ and $\mathbf{G_0}$ were initialized with the mean and inverse covariance matrix of the initial parameters (Table III). The scale matrix $\mathbf{R_0}$ of the Wishart distribution for sampling $\mathbf{G_n}$ was also initialized with the covariance of dictionary parameters. The remaining hyperparameters were empirically set to $\alpha_k = 2, e = 1, f = 2, \nu_0 = 3, \gamma_s = 10$, and $\gamma_e = 8$.

Dictionary combination (Algorithm 2) was performed on the generated vector parameters $[T_{\text{rise}}, T_{\text{decay}}]^T$ with $N_b = 100, 225, 400$. We did not include the time offset t_0 in this procedure, since it does not contribute to the shape of the dictionary atoms; the final dictionaries should contain the learnt versions of phasic atoms shifted across the entire analysis frame. Therefore, the quantized matrix of dictionary parameters $\boldsymbol{\Theta_Q} \in \Re^{2\times N_b}$ is replicated for all values of t_0 (Table III), yielding final dictionaries with $16N_b$ phasic atoms.

2) Steepest Descent DL: Dictionaries are trained in parallel for each exemplar data using the SD [22], in which OMP alternates with a least-squares fit step for estimating the parameters of the selected atoms. SD was run over 250 iterations. We perform the same procedure for combining the learnt parameters, yielding dictionaries of same size with MCMC.

3) Equiangular Tight Frame DL: This method alternates between finding the dictionary with minimum Frobenius distance from the ETF and updating the corresponding parameters [24]. These steps involve an ETF relaxation constraint value and a gradient descent step, crucial for the overall performance, referred in [24] as α_k and ϵ. In our experiments, dictionaries were trained with different combinations of $\alpha_k \in \{0.1, 0.5, 0.9\}$ and $\epsilon \in \{0.001, 0.002, 0.003, 0.004\}$. The dictionary $\{\alpha_k^*, \epsilon^*\}$ that showed the lowest relative RMS error on the training data was used to evaluate the corresponding test data. In order to ensure the same dictionary size as the other approaches, parameters are uniformly initialized within the intervals $T_{\text{rise}} \in [8, 18], T_{\text{decay}} \in [10, 20]$ with $\sqrt{N_b} = 10, 15, 20$ different values.

4) K-SVD: K-SVD is a classical non-parametric DL method generalizing K-means and iteratively estimating each dictionary atom in order to minimize the reconstruction error [14]. The original dictionaries used in this method were the same as the initial ones of the aforementioned approaches.

5) Evaluation: Evaluation of the final dictionaries is performed based on the relative RMS error computed between the original signal and its corresponding approximation through OMP based on the final dictionaries. RMS error is a common metric in related studies [14], [15], [24]. In order to ensure that our results reflect the ability of the algorithms to represent unseen data, all experiments are performed within a 10-fold cross-validation for the synthetic data and a leave-one-subject-out cross-validation setup for the real EDA signals. We note that OMP has two functionalities in our framework. It serves as a decomposition technique, based on which the learned dictionaries are evaluated against the test data, and is also used at the dictionary combination step (Section IV) in order to "prune" the atoms of the generated dictionaries. OMP does not operate on the test data during this dictionary combination framework, but rather at the evaluation step, during which the final dictionary is already created independently of the test set.

Besides signal reconstruction, dictionary coherence is an additional evaluation criterion. It is defined as the absolute value of the largest inner-product between any atoms of the dictionary and is an important property related to signal reconstruction quality [24]. We computed the resulting coherence of the dictionaries after training.

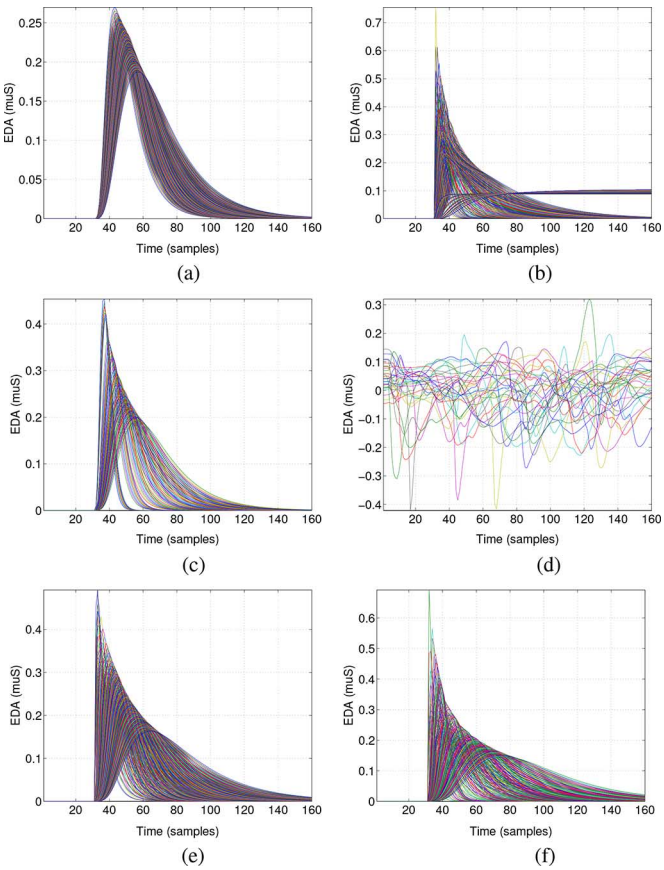


Fig. 1. Example of initial dictionary and dictionaries learnt Steepest Descent (SD), Equiangular Tight Frame (ETF), K-SVD and Markov Chain Monte Carlo Bayesian inference (MCMC) using atom sampling with and without replacement (w-,w-o rplcm). Dictionary combination is performed with $N_b = 400$. An instance of phasic atoms shifted with $t_0 = 30$ is shown. (a) Initial, (b) SD, (c) ETF, (d) K-SVD, (e) MCMC (w- rplcm), (f) MCMC (w-o rplcm).

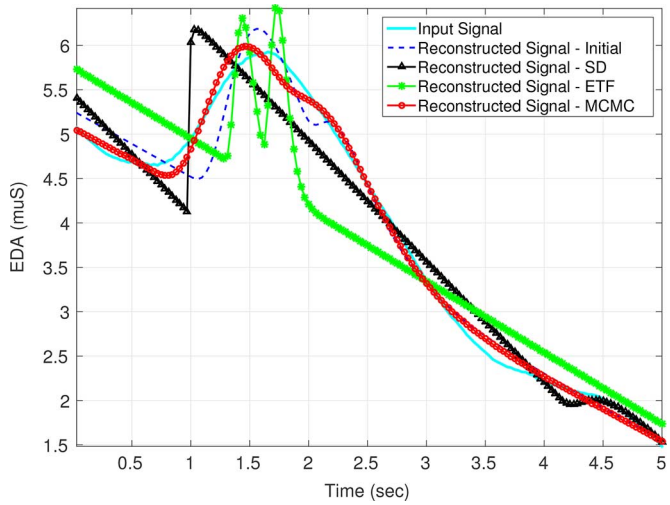


Fig. 2. Example of input EDA signal (solid cyan line) and reconstructed signals using the original dictionary (blue dashed line) and the dictionaries learnt with Steepest Descent (SD), Equiangular Tight Frame (ETF), and Markov Chain Monte Carlo Bayesian inference (MCMC) (black-triangle, green-asterisk, and red-circled lines, respectively). Reconstruction is performed using orthogonal matching pursuit with 4 iterations.

C. Results

Visual inspection of the final dictionaries indicates that SD does not always preserve the initial dictionary structure (Figs. 1(a)–(b)), while ETF yields less variable dictionaries (Fig. 1(c)). MCMC results in a variety of atoms preserving the original shape (Fig. 1(e)–(f)). In contrast to parametric DL, non-parametric K-SVD yields very unstructured non-interpretable atoms (Fig. 1(d)). Further comparison between an exemplar input and the reconstructed signals suggests that our approach depicts superior signal representation (Fig. 2).

Dictionaries learnt from our proposed Bayesian DL yield lower reconstruction errors compared to the initial ones and the ones learnt through SD and ETF (Fig. 3). Atom sampling with and without replacement are not significantly different, since the two distributions are very similar for $K \gg L$. Dictionaries trained using SD perform quite poorly on unseen synthetic data (Figs. 3(a)–(c)), which might occur because their simple structure causes significant overfitting to least-squares-based methods. Despite the fact that ETF DL is not prone to overfitting, since it does not take into account exemplar data during training, it lacks adaptation to more complex real data (Figs. 3(d)–(f)). K-SVD appears more accurate for real signals than the parametric approaches (Figs. 3(d)–(f)), indicating its ability to learn more complex patterns, not necessarily of the same structure as the initial dictionaries. Similar results occur for different dictionary sizes, omitted here for the sake of simplicity.

The large number of atoms generally yields dictionaries of high coherence. All considered methods appear quite equivalent, with ETF achieving the lowest coherence, since this is included as an optimization metric (Table IV).

D. MCMC Diagnostics

Besides theoretical analysis (Section VI), another way to examine MCMC convergence is to see how well the MC is mixing, which is usually achieved through visual inspection. Traceplots of several problem variables from an exemplar signal indicate that the considered chains move around the parameter space and are not only limited in certain areas suggesting good mixing (Figs. 4(a)–(d)). Density plots of the generated samples further validate that the estimated posteriors are close to the target distributions (Figs. 4(e)–(h)).

Convergence diagnostics can further quantitatively assess if there is a bias from generated samples. We use the Geweke diagnostic [77], because it only requires one running chain and attempts to address issues of both bias and variance [78]. It takes two non-overlapping parts of the MC (usually the first 0.1 and last 0.5) and compares their mean using a difference of means test. If the samples are drawn from the stationary distribution of the chain, the two means are equal. The test statistic is a standard Z-score with values under convergence within two standard deviations from zero, i.e., $|z| < 2$ for the standard normal distribution. We perform the Geweke test for both datasets and atom sampling methods (Sections II-A, II-B) with the first 100 samples used as a burn-in period.

The high-dimensionality of our problem prohibits us to report diagnostics for each variable separately, therefore we will summarize the results for groups of variables. For each group of

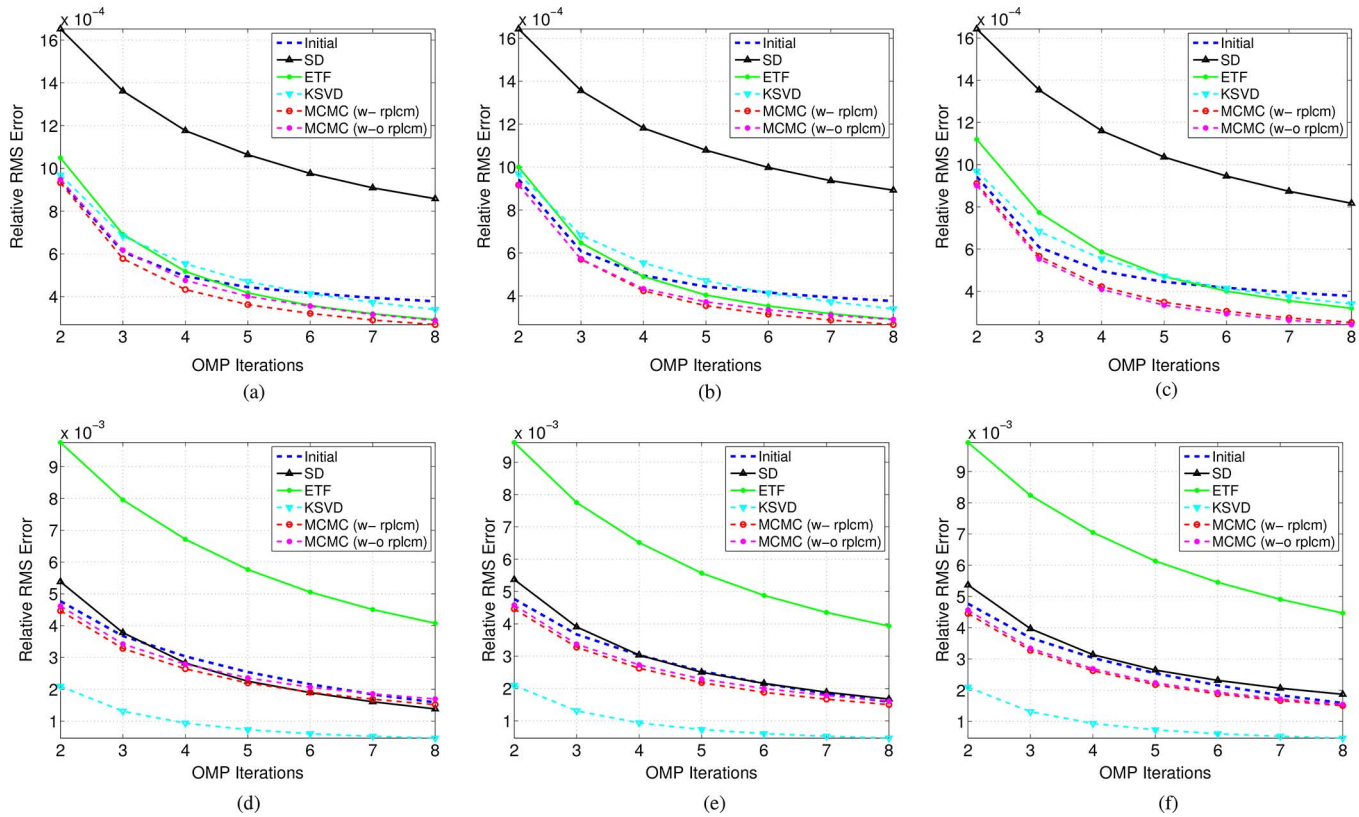


Fig. 3. Relative root mean square (RMS) error between original and reconstructed signal with respect to the number of (orthogonal) matching pursuit (OMP) iterations. Dictionaries are learnt with Steepest Descent (SD), Equiangular Tight Frame (ETF), K-SVD, and Markov Chain Monte Carlo Bayesian inference (MCMC). During MCMC atom sampling is performed with (w-) and without (w-o) replacement (rplcm). (a) Synthetic, 1600 phasic + 63 tonic atoms, (b) Synthetic, 3600 phasic + 63 tonic atoms, (c) Synthetic, 6400 phasic + 63 tonic atoms, (d) Real, 1600 phasic + 63 tonic atoms, (e) Real, 3600 phasic + 63 tonic atoms, (f) Real, 6400 phasic + 63 tonic atoms.

TABLE IV
FINAL DICTIONARY COHERENCE

Method	Synthetic Data	Real Data
Initial	1	1
SD	0.9990	0.9992
ETF	0.9988	0.9990
K-SVD	0.9989	0.9991
MCMC w- rplcm	0.9989	0.9991
MCMC w-o rplcm	0.9989	0.9991

variables, we report the proportion of chains for which $|z| < 2$ (Table V). Most of the variables in our framework succeed on the Geweke diagnostic. The dictionary parameters, generated with MH, usually have a lower success percentages. Although there is a reduced number of cases where the Geweke diagnostic fails, given the large number of variables and the signal reconstruction results (Section VII-C), the performed number of MCMC iterations appears to result in meaningful dictionaries learned through this process.

MH acceptance rate refers to the fraction of candidate draws that are accepted and is important for convergence. Very high acceptance rates suggest that the chain is not mixing well, while very low rates might be inefficient. The acceptance rates for the variables of our problem (Table VI) are close to previously proposed optimal values in the literature [79] suggesting a good mixing of the considered MC.

VIII. DISCUSSION

An important benefit of Bayesian methods yields in providing estimates of the entire variable set of a problem. In the considered setup this can help inferring the dictionary size, the optimal number of dictionary atoms for a given signal, and the corresponding noise levels. Estimation of the dictionary size is not as crucial in our greedy sparse representation approach, that is not as prone to overcomplete dictionaries as basis pursuit methods [5]. Inferring the optimal number of dictionary atoms is important, as it can yield high compression rates and help interpret the underlying signal information. An extension of our method could have imposed discrete probabilistic distributions onto the number of atoms appropriately inferred through MH.

Another advantage of Bayesian methods is that they are less prone to overfitting, which usually occurs with deterministic algorithms that might describe the random noise instead of the actual data [81]. This is reflected in our results (Section VII-C), since deterministic SD, that does not include prior knowledge of the signal structure, performs more poorly on unseen data. On the other side, ETF avoids overfitting, but does not learn the morphology of training data. Bayesian inference appears to compromise between the two.

Inherent differences exist between parametric and non-parametric DL methods. The first impose a predetermined structure on the input space (through function ϕ) and learn the parameters that represent this structure from the training data. Their major

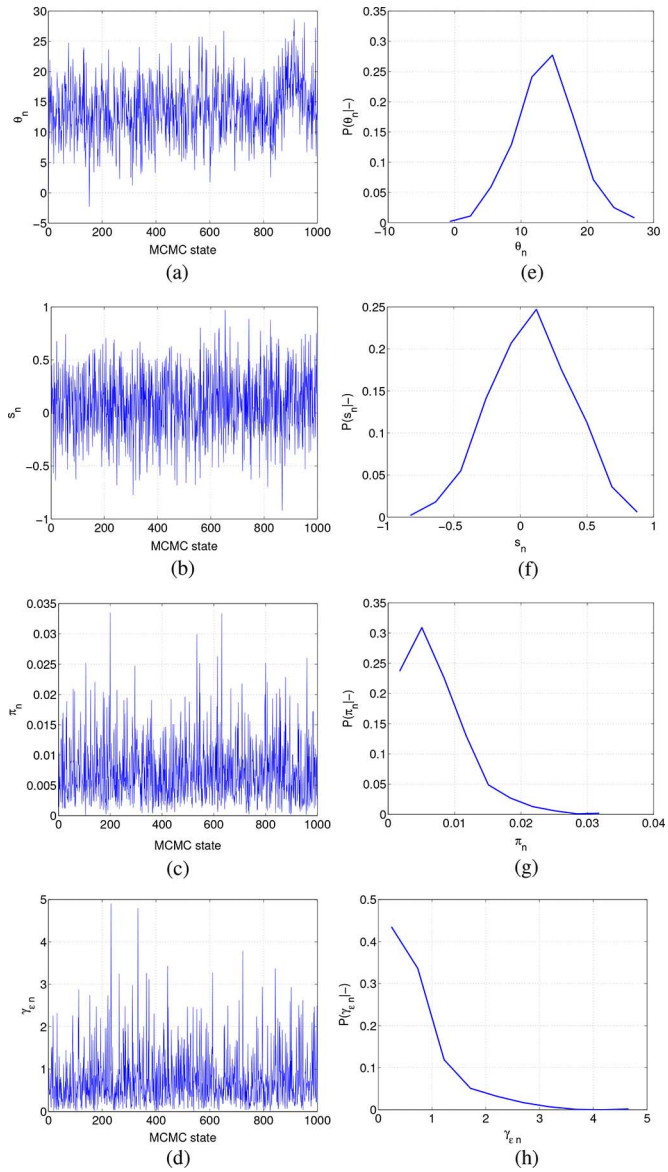


Fig. 4. Example MCMC trace plots and generated posteriors for the first element of vectors containing the (a),(e) dictionary atoms $\boldsymbol{\theta}_n$, (b),(f) atom coefficients $\mathbf{s_n}$, (c),(g) atom priors $\boldsymbol{\pi_n}$, and (d),(h) for the noise precision γ_{ϵ_n}. (a) MCMC trace of θ_{n1}, (b) MCMC trace of s_{n1}, (c) MCMC trace of π_{n1}, (d) MCMC trace of γ_{ϵ_n}, (e) Posterior of θ_{n1}, (f) Posterior of s_{n1}, (g) Posterior of π_{n1}, (h) Posterior of γ_{ϵ_n}.

TABLE V
GEWEKE MCMC DIAGNOSTIC - $P(|z| < 2)$

		Synthetic Data		Real Data	
		w- rplcm	w-o rplcm	w- rplcm	w-o rplcm
$\mathbf{c_n}$		0.93	0.98	0.92	0.95
$\mathbf{z_n}$		0.93	0.93	0.91	0.94
γ_{ϵ_n}		0.91	0.98	0.99	0.92
$\pi_{\mathbf{n}}$		0.93	0.96	0.96	0.92
$\boldsymbol{\theta_n}$	t_0	0.68	0.70	0.72	0.75
	T_{rise}	0.91	0.92	0.87	0.85
	T_{decay}	0.92	0.92	0.86	0.85
$\mathbf{g_n}$	t_0	0.75	0.80	0.78	0.79
	T_{rise}	0.90	0.87	0.89	0.87
	T_{decay}	0.90	0.86	0.91	0.85
$\mathbf{G_n}$		0.81	0.75	0.82	0.83

benefit lies in their interpretability, since the considered dictionary atoms are able to meaningfully capture the characteristics of input signals and can be used for knowledge-driven classification and inference. On the other hand, the exemplar signals learned from non-parametric methods are hardly interpretable and can blindly represent the input space. Since the functionality and scope of these methods is so different, meaningful comparison is challenging. In the case of synthetic data, where function ϕ is a perfect match to the signal (i.e., input signals are built with the same functions as the dictionary atoms), our proposed parametric Bayesian approach is more reliable than K-SVD in learning the hidden atom parameters. This means that given a perfectly constructed dictionary function, Bayesian parametric DL yields better results, even compared to non-parametric approaches. However, in the case of real signals, function ϕ cannot always perfectly selected, therefore parametric approaches seem to perform slightly worse than K-SVD. While more precise function might have yielded lower errors, the problem of finding the optimal mapping function is still active in research [82].

TABLE VI
METROPOLIS-HASTINGS ACCEPTANCE RATES (%)

Synthetic Data		Real Data	
w- rplcm	w-o rplcm	w- rplcm	w-o rplcm
23.22	31.52	25.48	38.15

Although our proposed approach is general for learning the parameters of the dictionary atoms, we need to have good knowledge about the appropriate mapping function ϕ between the parameters and the data. As discussed in Section V, the selection of ϕ is usually guided by the application of interest and can vary for different types of signals. In the case of 2D images, for example, the dictionary needs to be constructed using a function ϕ that captures the context of the image, such as a Gabor or wavelet-like function. In order to reduce complexity, we can convert the 2D image into a 1D vector with dimensionality equal to the number of pixels, as typically performed [14], [40], [44]. Given a function ϕ suitable for the image of interest, our Bayesian approach can learn the parameters that efficiently capture the shape of the data.

In this paper, for the sake of simplicity we considered a simple dictionary combination approach with K-means quantization (Section IV). However, there exist more sophisticated approaches of block-coordinate descent with warm restarts [17] and weighted batch averaging [43].

Convergence is a critical component of MCMC in the context of Bayesian inference. In Section VI, we discussed the uniform ergodicity of the corresponding MH-within-Gibbs sampler, ensuring that the MC converges to the invariant distribution. Convergence rate is another important issue, for which previous studies have proposed explicit theoretical bounds in simple scenarios [59], [83]. The computation of these bounds can be quite difficult in such high-dimensional problems [84], therefore it goes beyond the scope of our study.

IX. CONCLUSION

We propose a sparse Bayesian model for parametric DL, whose problem variables follow appropriately selected probabilistic distributions. We use MH-within-Gibbs to infer the corresponding variables, because of its ability to compensate for posterior distributions that cannot be analytically computed. We further show the uniform ergodicity of the proposed

MCMC through the minorization of the corresponding Gibbs sampler and the bounded conditional weights of the MH. Our experimental results performed on synthetic and real biomedical signals indicate that this approach offers benefits in terms of signal reconstruction compared to previously proposed SD and ETF methods and also provides a good tradeoff between learning the signal structure and avoiding overfitting.

APPENDIX A
MCMC ERGODICITY PROOFS

We provide the proofs for the theorems and lemmas of Section VI.

$$
\begin{aligned}
&P((\mathbf{Y_1},\mathbf{Y_2},\ldots,\mathbf{Y_{B-1}}),A_{-B})\\
&\quad=\int\limits_{A_1\ldots A_{B-1}} p(\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}}|\mathbf{y_1},\ldots,\mathbf{y_B})\mu\left(d(\mathbf{y'}_{-\mathbf{B}})\right)\\
&\quad=\int\limits_{A_1\ldots A_{B-1}} p(\mathbf{y'_1},\ldots,\mathbf{y'_{B-2}}|\mathbf{y}_{-\mathbf{y_1}},\ldots,\mathbf{y}_{-\mathbf{y_{B-2}}})\\
&\qquad\times p\left(\mathbf{y'_{B-1}}|\mathbf{y}_{-\mathbf{y_{B-1}}}\right)\mu_{B-1}\left(d\mathbf{y'_{B-1}}\right)\ldots\mu_1\left(d\mathbf{y'_1}\right)\\
&\quad=\int\limits_{A_1\ldots A_{B-2}} p\left(\mathbf{y'_1},\ldots,\mathbf{y'_{B-2}}|\mathbf{y}_{-\mathbf{y_1}},\ldots,\mathbf{y}_{-\mathbf{y_{B-2}}}\right)\\
&\qquad\times\left(\int\limits_{A_{B-1}} p\left(\mathbf{y'_{B-1}}|\mathbf{y}_{-\mathbf{y_{B-1}}}\right)\mu_{B-1}\left(d\mathbf{y'_{B-1}}\right)\right)\\
&\qquad\times\mu_{B-2}\left(d\mathbf{y'_{B-2}}\right)\ldots\mu_1\left(d\mathbf{y'_1}\right)\\
&\quad\geq\epsilon_{B-1}\nu_{B-1}(A_{B-1})\int\limits_{A_1\ldots A_{B-3}} p\left(\mathbf{y'_1},\ldots,\mathbf{y'_{B-3}}|\ldots\right)\\
&\qquad\times\left(\int\limits_{A_{B-2}} p\left(\mathbf{y'_{B-2}}|\mathbf{y}_{-\mathbf{y_{B-2}}}\right)\mu_{B-2}\left(d\mathbf{y'_{B-2}}\right)\right)\\
&\qquad\times\mu_{B-3}\left(d\mathbf{y'_{B-3}}\right)\ldots\mu_1\left(d\mathbf{y'_1}\right)\\
&\quad\geq\ldots\geq\prod_{b-1}^{B-1}\epsilon_b\nu_b(A_b)
\end{aligned}
\tag{30}
$$

$$
\begin{aligned}
&P\left((\mathbf{Y_1},\mathbf{Y_2},\ldots,\mathbf{Y_B}),A_1\times A_2\times\ldots\times A_B\right)\\
&\quad=\int\limits_{A_1\ldots A_B} p\left(\mathbf{y'_1},\ldots,\mathbf{y'_B}|\mathbf{y_1},\ldots,\mathbf{y_B}\right)\mu\left(d\left(\mathbf{y'_1},\ldots,\mathbf{y'_B}\right)\right)\\
&\quad=\int\limits_{A_B} p\left(\mathbf{y'_B}|\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}},\mathbf{y_B}\right)\\
&\qquad\times\int\limits_{A_1\ldots A_{B-1}} p\left(\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}}|\mathbf{y_1},\ldots,\mathbf{y_{B-1}},\mathbf{y_B}\right)\\
&\qquad\times\mu\left(d\left(\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}}\right)\right)\cdot\mu_B\left(d\mathbf{y'_B}\right)\\
&\quad\geq\epsilon_0\int\limits_{A_B} p\left(\mathbf{y'_B}|\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}},\mathbf{y_B}\right)\\
&\qquad\times\int\limits_{A_1\ldots A_{B-1}}\nu_0\left(d\left(\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}}\right)\right)\cdot\mu_B\left(d\mathbf{y'_B}\right)=\epsilon_0\\
&\qquad\times\int\limits_{A_1\ldots A_{B-1}} p\left(\mathbf{y}_{-\mathbf{y_B}},A_b\right)\nu\left(d(\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}}\right)
\end{aligned}
\tag{31}
$$

Proof of Lemma 6.1: We assume that the sampling order is $\mathbf{Y_1},\ldots,\mathbf{Y_{B-1}}$. Let $\{\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}}\}$ and $\{\mathbf{y_1},\ldots,\mathbf{y_B}\}$ be the block variables at the current and previous MCMC state, respectively. The conditional probability for sampling the first $B-1$ blocks is

$$
\begin{aligned}
&p\left(\mathbf{y'_1},\ldots,\mathbf{y'_{B-1}}|\mathbf{y_1},\ldots,\mathbf{y_B}\right)\\
&\quad=p\left(\mathbf{y'_{B-1}}|\mathbf{y'_1},\ldots,\mathbf{y'_{B-2}},\mathbf{y_{B-1}},\mathbf{y_B}\right)\\
&\qquad\cdot p\left(\mathbf{y'_1},\ldots,\mathbf{y'_{B-2}}|\mathbf{y_1},\ldots,\mathbf{y_B}\right)\\
&\quad=p\left(\mathbf{y'_{B-1}}|\mathbf{y'_1},\ldots,\mathbf{y'_{B-2}},\mathbf{y_{B-1}},\mathbf{y_B}\right)\\
&\qquad\cdot p\left(\mathbf{y'_{B-2}}|\mathbf{y'_1},\ldots,\mathbf{y'_{B-3}},\mathbf{y_{B-2}},\mathbf{y_{B-1}},\mathbf{y_B}\right)\\
&\qquad\cdot\ldots\cdot p\left(\mathbf{y'_1}|\mathbf{y_1},\ldots,\mathbf{y_B}\right)\\
&\quad=p\left(\mathbf{y'_{B-1}}|\mathbf{y}_{-\mathbf{y_{B-1}}}\right)\cdot p\left(\mathbf{y'_{B-2}}|\mathbf{y}_{-\mathbf{y_{B-2}}}\right)\cdot p\left(\mathbf{y'_1}|\mathbf{y}_{-\mathbf{y_1}}\right)
\end{aligned}
$$

where $\mathbf{y}_{-\mathbf{b}}$ contains all variables except $\mathbf{y_b}$. The Markov kernel for the first $B-1$ blocks can be written as in (30). The first and second inequalities in (30) occur from the minorization of the $(B-1)$th and $(B-2)$th blocks. Therefore $\exists\epsilon_0=\prod_{b=1}^{B-1}\epsilon_b>0$ and $\nu_0=\prod_{b-1}^{B-1}\nu_b(A_b)$ (a probability measure) that satisfy the desired inequality. ∎

Proof of Theorem 6.2: If we assume that the sampling order is $\mathbf{Y_1},\ldots,\mathbf{Y_B}$, the Markov kernel of the Gibbs sampler can be expressed as in (31). The first inequality in (31) results from Lemma 6.1 and can yield to a minorization condition for the entire Gibbs sampler. ∎

Proof of Lemma 6.2: From Def. 6.1, Table II, and (7)–(8)

$$
\begin{aligned}
w_{\mathbf{y}_{\tilde{\boldsymbol{\theta}}_n}}&\propto\exp\left(-\frac{\gamma_{\epsilon_n}}{2}||\boldsymbol{\epsilon_n}||^2\right)\frac{|\tilde{\mathbf{G}}_\mathbf{n}|^{\frac{1}{2}}}{|\boldsymbol{V}_{\tilde{\boldsymbol{\theta}}_n}|^{-1/2}}\\
&\cdot\frac{\exp\left[-\frac{1}{2}(\tilde{\boldsymbol{\theta}}_n-\tilde{\mathbf{g}}_\mathbf{n})^T\tilde{\mathbf{G}}_\mathbf{n}(\tilde{\boldsymbol{\theta}}_n-\tilde{\mathbf{g}}_\mathbf{n})\right]}{\left[1+\frac{1}{\nu_1+Q}\left(\tilde{\boldsymbol{\theta}}_n-\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_n}\right)^T\hat{\mathbf{V}}^{-1}_{\tilde{\boldsymbol{\theta}}_n}\left(\tilde{\boldsymbol{\theta}}_n-\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_n}\right)\right]^{-\frac{\nu_1+Q}{2}}}
\end{aligned}
\tag{32}
$$

Using (10), we have

$$
\frac{|\tilde{\mathbf{G}}_\mathbf{n}|^{\frac{1}{2}}}{|\boldsymbol{V}_{\tilde{\boldsymbol{\theta}}_n}|^{-\frac{1}{2}}}=\tau|\tilde{\mathbf{G}}_\mathbf{n}|^{\frac{1}{2}}<\infty\text{ and }\frac{1}{\prod_{k=1}^{K}|\boldsymbol{V}_{\boldsymbol{\theta}_k}|^{-\frac{1}{2}}}=\tau^K<\infty
$$

since $\tilde{\mathbf{G}}_\mathbf{n}$ is the precision matrix of Gaussian distribution, therefore has finite eigenvalues and determinants, and $\tau<\infty$. Moreover $\prod_{n=1}^{N}\exp\left(-\frac{\gamma_{\epsilon_n}}{2}||\boldsymbol{\epsilon_n}||^2\right)<1$ because $\frac{\gamma_{\epsilon_n}}{2}||\boldsymbol{\epsilon_n}||^2\geq 0$ for $n=1,\ldots,N$ and $\gamma_{\epsilon_n}>0$. Finally the function

$$
f(\tilde{\boldsymbol{\theta}}_n)=\frac{\left[1+\frac{1}{\nu_1+Q}\left(\tilde{\boldsymbol{\theta}}_n-\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_n}\right)^T\hat{\mathbf{V}}^{-1}_{\tilde{\boldsymbol{\theta}}_n}\left(\tilde{\boldsymbol{\theta}}_n-\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\theta}}_n}\right)\right]^{\frac{\nu_1+Q}{2}}}{\exp\left(\frac{1}{2}(\tilde{\boldsymbol{\theta}}_n-\tilde{\mathbf{g}}_n)^T\tilde{\mathbf{G}}_\mathbf{n}(\tilde{\boldsymbol{\theta}}_n-\tilde{\mathbf{g}}_n)\right)}
\tag{33}
$$

is bounded since $f(\tilde{\boldsymbol{\theta}}_n)\to 0$, $||\tilde{\boldsymbol{\theta}}_n||\to\infty$, and f is continuous. Function f remains bounded at $||\tilde{\boldsymbol{\theta}}_n||\to\infty$, since the quadratic forms $(\tilde{\boldsymbol{\theta}}_n-\tilde{\mathbf{g}}_\mathbf{n})^T\tilde{\mathbf{G}}_\mathbf{n}(\tilde{\boldsymbol{\theta}}_n-\tilde{\mathbf{g}}_\mathbf{n})$ and $(\tilde{\boldsymbol{\theta}}_n-\tilde{\mathbf{g}}_\mathbf{n})^T\hat{\mathbf{V}}^{-1}_{\tilde{\boldsymbol{\theta}}_n}(\tilde{\boldsymbol{\theta}}_n-\tilde{\mathbf{g}}_\mathbf{n})$ increase to infinity at the same rate, and the denominator increases exponentially fast, while the numerator with polynomial rate. ∎

TABLE VII
AVERAGE COMPUTATION TIME OF DICTIONARY LEARNING ALGORITHMS

Method	# Variables	Computation Time (sec / training iteration)
SD	438	0.1173
ETF	432	0.0616
K-SVD	438	3.9536
MCMC w- rplcm	1024	3.5014
MCMC w-o rplcm	1024	4.9638

Proof of Theorem 6.3: The first $B-1$ blocks of $\mathbf{y_n}$, that are generated with the Gibbs sampler, follow well-known distributions (Table I), therefore it is trivial to show that their updates are minorisable. From Theorem 6.2, since the partial updates for all blocks except the last one are minorisable, the entire Gibbs sampler is minorisable, therefore uniformly ergodic. From Lemma 6.2, the conditional weight of the Bth block $\mathbf{y}_{\tilde{\boldsymbol{\theta}}_n}$ generated with the MH is bounded. Thus the MH-within-Gibbs sampler meets the conditions of Theorem 6.1, therefore it is uniformly ergodic. ■

APPENDIX B
COMPUTATIONAL COMPLEXITY

We analyze the computational complexity of our proposed framework for each MCMC step (Algorithm 1, Table II) using the "$\mathcal{O}$" notation. For an input signal and MCMC iteration, the weight of the Multinomial distribution for each atom is computed with cost $\mathcal{O}(P)$, i.e., $\mathcal{O}(PK)$ for all atoms. Sampling L indices from K dictionary atoms with replacement results in $\mathcal{O}(LK)$, therefore the final cost yields $\mathcal{O}(PK+LK)$. For sampling without replacement, re-adjusting the atom weights requires $\mathcal{O}\left((K-1)+\ldots+(K-L+1)\right) \sim \mathcal{O}(LK)$. Since each of the L iterations takes into account the previously selected atoms [55], the cost of the Wallenius distribution is $\mathcal{O}\left(K+(K-1)+\ldots+(K-L+1)\right) \sim \mathcal{O}(LK)$.

Each of the L coefficients is generated from the normal distribution, whose mean requires $\mathcal{O}(P)$, therefore the entire complexity is $\mathcal{O}(LP)$. Regarding the dictionary atom parameters, their posterior (7) requires $\mathcal{O}\left(L(P+Q^2)\right)$, while the typical cost of the Nelder Mead's simplex is $\mathcal{O}\left(Q^2\right)$ [80]. Finally, complexity results in $\mathcal{O}(1)$ for the noise $\boldsymbol{\epsilon_n}$, $\mathcal{O}(K+L)$ for the Dirichlet prior $\boldsymbol{\pi_n}$, $\mathcal{O}(Q^2)$ for the mean $\mathbf{g_{nk}}$ and precision $\mathbf{G_n}$ of the dictionary parameters prior, and $\mathcal{O}(P)$ for the noise precision γ_{ϵ_n}.

Taking these into account, the total complexity when using sampling with and without replacement, respectively, yields:

$$\mathcal{O}(PK+LK+2LP+(L+3)Q^2+P+K+L+1) \sim \mathcal{O}(PK+LK+LP+LQ^2) \quad (34)$$

$$\mathcal{O}(2LK+2LP+(L+3)Q^2+P+K+L+1) \sim \mathcal{O}(LK+LP+LQ^2) \quad (35)$$

Our approach requires first-order polynomial time with the signal dimensionality P, the number of dictionary atoms K, and the number of selected atoms in the representation $L \ll K$, and second-order polynomial cost with respect to the number of dictionary parameters Q. We note that $Q \ll P, L$, therefore the latter is not too expensive.

We further compare run-time statistics of all the approaches. We report the average duration of one training iteration and the number of estimated variables for each approach (Table VII). Experiments were performed with an Intel Core i7 Processor with CPU at 2.93 Hz and RAM at 7.8 GB. SD and K-SVD require a decomposition step, therefore they compute slightly more variables than ETF. Our method estimates the highest number of variables, including the priors of the considered problem. Results indicate that SD and ETF are computationally less expensive than the proposed MCMC. Consistently with previous observations [19], K-SVD also has high computational cost. MCMC sampling with and without replacement yield computation times of the same order.

REFERENCES

[1] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York, NY, USA: Springer, 2010.

[2] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[3] Y. C. Pati, R. Ramin, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Conf. Signals, Syst., Comput.*, 1993, pp. 40–41.

[4] G. Davis, S. G. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Construct. Approx.*, vol. 13, no. 1, pp. 57–98, 1997.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientif. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[6] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, 1999.

[7] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 760–770, 2003.

[8] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[9] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, 2007.

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

[11] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vis. Res.*, vol. 20, no. 10, pp. 847–856, 1980.

[12] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961–1005, 1990.

[13] J. L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, 2002.

[14] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[15] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. ICASSP*, 1999, pp. 2443–2446.

[16] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," *Proc. ACM*, pp. 689–696, 2009.

[18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proc. NIPS*, 2008, pp. 1033–1040.

[19] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[20] K. Skretting and K. Engan, "Image compression using learned dictionaries by RLS-DLA and compared with K-SVD," in *Proc. ICASSP*, 2011, pp. 1517–1520.

[21] D. Thanou, D. Shuman, and P. Frossard, "Learning parametric dictionaries for signals on graphs," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3849–3862, 2014.

[22] M. Ataee, H. Zayyani, M. Babaie-Zadeh, and C. Jutten, "Parametric dictionary learning using steepest descent," in *Proc. ICASSP*, 2010, pp. 1987–1981.

[23] H. H. Szu, B. A. Telfer, and S. L. Kadambe, "Neural network adaptive wavelets for signal representation and classification," *Opt. Eng.*, vol. 21, no. 9, pp. 1907–1916, 1992.

[24] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4800–4810, 2009.

[25] S. Merlet, E. Caruyer, A. Ghosh, and R. Deriche, "A computational diffusion MRI and parametric dictionary learning framework for modeling the diffusion signal and its features," *Med. Image Anal.*, vol. 17, no. 9, pp. 830–843, 2013.

[26] Q. Barthélemy, C. Gouy-Pailler, Y. Isaac, A. Souloumiac, A. Larue, and J. I. Mars, "Multivariate temporal dictionary learning for EEG," *J. Neurosci. Methods*, vol. 215, no. 1, pp. 19–28, 2013.

[27] M. Reisert, H. Skibbe, and V. G. Kiselev, "The diffusion dictionary in the human brain is short: Rotation invariant learning of basis functions," *Proc. CDMRI*, pp. 47–55, 2014.

[28] T. Chaspari, A. Tsiartas, L. I. Stein, S. A. Cermak, and S. S. Narayanan, "Sparse representation of electrodermal activity with knowledge-driven dictionaries," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 3, pp. 960–971, 2015.

[29] M. E. Tipping, "Bayesian inference: An introduction to principles and practice in machine learning," in *Advanced Lectures on Machine Learning*. New York, NY, USA: Springer, 2004, pp. 41–62.

[30] B. Ophir, M. Elad, N. Bertin, and M. D. Plumbley, "Sequential minimal eigenvalues—An approach to analysis dictionary learning," *Proc. EUSIPCO*, 2011.

[31] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, 2005.

[32] M. Aharon and M. Elad, "Sparse and redundant modeling of image content using an image-signature-dictionary," *Soc. Indust. Appl. Math. (SIAM) J. Imaging Series*, vol. 1, no. 3, pp. 228–247, 2008.

[33] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1553–1564, 2010.

[34] J. A. Mazaheri, C. Guillemot, and C. Labit, "Learning a tree-structured dictionary for efficient image representation with adaptive sparse coding," in *Proc. ICASSP*, 2013, pp. 1320–1324.

[35] T. Blumensath and M. Davies, "Sparse and shift-invariant representations of music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 50–57, 2006.

[36] K. Engan, K. Skretting, and J. H. Husøy, "Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation," *Digit. Signal Process.*, vol. 17, no. 1, pp. 32–49, 2007.

[37] W. J. Jasper, S. J. Garnier, and H. Potlapalli, "Texture characterization and defect detection using adaptive wavelets," *Opt. Eng.*, vol. 35, no. 11, pp. 3140–3149, 1996.

[38] T. Blumensath, "Monte Carlo methods for compressed sensing," in *Proc. ICASSP*, 2014, pp. 1000–1004.

[39] M. Elad and I. Yavneh, "A plurality of sparse representations is better than the sparsest one alone," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4701–4714, 2009.

[40] M. S. Lewicki and B. A. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *J. Opt. Soc. Amer. A*, vol. 16, no. 7, pp. 1587–1601, 1999.

[41] M. S. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2000.

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

[43] L. Li, J. Silva, M. Zhou, and L. Carin, "Online Bayesian dictionary learning for large datasets," in *Proc. ICASSP*, 2012, pp. 2157–2160.

[44] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 130–44, 2012.

[45] L. He, H. Qi, and R. Zaretzki, "Non-parametric Bayesian dictionary learning for image super resolution," in *Proc. Future of Instrum. Int. Workshop (FIIW)*, 2011, pp. 122–125.

[46] R. Crandall, B. Dong, and A. Bilgin, "Randomized iterative hard thresholding: A fast approximate MMSE estimator for sparse approximations," 2013 [Online]. Available: http://math.arizona.edu/~dongbin/Publications/RandIHT

[47] P. Sallee and B. A. Olshausen, "Learning sparse multiscale image representations," in *Proc. NIPS*, 2002, pp. 1327–1334.

[48] M. B. I. Reaz, M. S. Hussain, and F. Mohd-Yasin, "Techniques of EMG signal analysis: Detection, processing, classification and applications," *Biological Procedures Online*, vol. 8, no. 1, pp. 11–35, 2006.

[49] K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapp.*, vol. 2, pp. 189–210, 1995.

[50] J. Grimmer, "An introduction to Bayesian inference via variational approximations," *Politic. Anal.*, vol. 19, no. 1, pp. 32–47, 2010.

[51] P. S. Meyn and L. R. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. New York, NY, USA: Cambridge Univ. Press, 2009.

[52] K. L. Mengersen and R. L. Tweedie, "Rates of convergence of the hastings and metropolis algorithms," *Ann. Statist.*, vol. 24, no. 1, pp. 101–121, 1996.

[53] M. Tsagkias, M. De Rijke, and W. Weerkamp, "Hypergeometric language models for republished article finding," in *Proc. SIGIR*, 2011, pp. 485–494.

[54] A. Fog, "Calculation methods for Wallenius' noncentral hypergeometric distribution," *Commun. Statist.—Simulat. Comput.*, vol. 37, no. 2, pp. 258–273, 2008.

[55] A. Fog, "Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions," *Commun. Statist.—Simulat. Comput.*, vol. 37, no. 2, pp. 241–257, 2008.

[56] K. T. Wallenius, "Biased sampling: The noncentral hypergeometric probability distribution," DTIC Document, Tech. Rep., 1963.

[57] S. Chib and I. Jeliazkov, "Marginal likelihood from the metropolis-hastings output," *J. Amer. Statist. Assoc.*, vol. 96, no. 453, pp. 270–281, 2001.

[58] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *Amer. Statist.*, vol. 49, no. 4, pp. 327–335, 1995.

[59] A. Johnson, "Geometric ergodicity of Gibbs samplers," Ph.D. dissertation, School of Statist., Univ. of Minnesota, Minnesota, MN, USA, 2009.

[60] A. A. Johnson, L. G. Jones, and C. R. Neath, "Component-wise Markov chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition," *Statist. Sci.*, vol. 28, no. 3, pp. 360–375, 2013.

[61] L. G. Jones, O. G. Roberts, and J. S. Rosenthal, "Convergence of conditional Metropolis-Hastings samplers," *Adv. Appl. Probab.*, vol. 46, no. 2, pp. 422–445, 2014.

[62] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[63] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[64] G. O. Roberts, A. Gelman, and W. R. Gilks, "Weak convergence and optimal scaling of random walk Metropolis algorithms," *Ann. Appl. Probab.*, vol. 7, no. 1, pp. 110–120, 1997.

[65] S. Chib, E. Greenberg, and R. Winkelmann, "Posterior simulation and Bayes factors in panel count data models," *J. Econometr.*, vol. 86, no. 1, pp. 33–54, 1986.

[66] M. Bédard and D. A. S. Fraser, "On a directionally adjusted Metropolis-Hastings algorithm," *Int. J. Statist. Sci.*, vol. 9, no. 1, pp. 33–57, 2008.

[67] S. Chib and S. Ramamurthy, "Tailored randomized block MCMC methods with application to DSGE models," *J. Econometr.*, vol. 155, no. 1, pp. 19–38, 2010.

[68] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, 1965.

[69] A. P. Lobo and P. C. Loizou, "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition," in *Proc. ICASSP*, 2003, pp. 817–820.

[70] I. Tošić and P. Frossard, "Dictionary learning for stereo image representation," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 921–934, 2011.

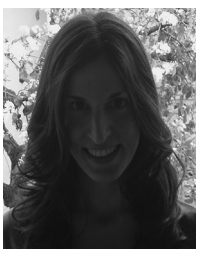
[71] L. Demanet and L. Ying, "Wave atoms and sparsity of oscillatory patterns," *Appl. Comput. Harmon. Anal.*, vol. 23, no. 3, pp. 368–387, 2007.

[72] Q. Barthélemy, C. Gouy-Pailler, Y. Isaac, A. Souloumiac, A. Larue, and J. I. Mars, "Multivariate temporal dictionary learning for EEG," *J. Neurosci. Methods*, vol. 215, no. 1, pp. 19–28, 2013.

[73] N. Ruiz-Reyes, P. Vera-Candeas, P. J. Reche-López, and F. Canadas-Quesada, "A time-frequency adaptive signal model-based approach for parametric ecg compression," in *Proc. EUSIPCO*, 2006, pp. 1–5.

[74] G. L. Jones and J. P. Hobert, "Honest exploration of intractable probability distributions via Markov chain Monte Carlo," *Statist. Sci.*, pp. 312–334, 2001.

[75] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.
[76] M. E. Dawson, A. M. Schell, and D. L. Filion, , J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds., "The Electrodermal System," in *Handbook of Psychophysiology*, 3rd ed. New York, NY, USA: Cambridge Univ. Press, 2007, pp. 159–181.
[77] J. Geweke, , J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, Eds., "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in *Bayesian Statistics*. New York, NY, USA: Oxford Univ. Press, 1991, vol. 4, pp. 169–193.
[78] M. K. Cowles and B. P. Carlin, "Markov chain Monte Carlo convergence diagnostics: A comparative review," *J. Amer. Statist. Assoc.*, vol. 91, no. 434, pp. 883–904, 1996.
[79] J. S. Rosenthal, "Optimal proposal distributions and adaptive MCMC," in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. L. Jones, and X. L. Meng, Eds. Boca Raton, FL, USA: Chapman & Hall/CRC, 2011, pp. 93–112.
[80] S. Singer and S. Singer, "Complexity analysis of Nelder-Mead search iterations," in *Proc. Conf. Appl. Math. Comput.*, 1999, pp. 185–196.
[81] C. M. Bishop, "Model-based machine learning," *Philos. Trans. Roy. Soc. A, Math., Phys., Eng. Sci.*, vol. 371, no. 1984, 2013.
[82] Q. Qiu, V. M. Patel, T. Pavan, and R. Chellappa, "Domain adaptive dictionary learning," in *Proc. ECCV*, 2012, pp. 631–645.
[83] J. S. Rosenthal, "Minorization conditions and convergence rates for Markov Chain Monte Carlo," *J. Amer. Statist. Assoc.*, vol. 90, no. 430, pp. 558–566, 1995.
[84] J. S. Rosenthal, "Theoretical rates of convergence for Markov Chain Monte Carlo," *Comput. Sci. Statist.*, pp. 486–486, 1994.



Theodora Chaspari (S'12) received the diploma in electrical and computer engineering from the National Technical University of Athens, Greece (2010) and the master's degree from the University of Southern California (2012), where she is currently pursuing a Ph.D. degree. Since 2010 she has been a member of the Signal Analysis and Interpretation Laboratory (SAIL). Her research interests lie in the area of biomedical signal processing, speech analysis and behavioral signal processing.

Ms. Chaspari is a recipient of the USC Annenberg Graduate Fellowship, USC WiSE Merit Fellowship, and the IEEE Signal Processing Society Travel Grant.

Andreas Tsiartas (S'10–M'14) received the B.Sc. degree in electronics and computer engineering from the Technical University of Crete, Crete, Greece, in 2006, and the M.Sc. and Ph.D. degrees in the Department of Electrical Engineering, University of Southern California (USC), Los Angeles, in 2014. He is currently a research engineer at SRI International. His main research direction focuses on speech-to-speech translation. Other research interests include acoustic and language modeling for automatic speech recognition (ASR) and voice activity detection.

Dr. Tsiartas received best teaching assistant awards for the years 2009 and 2010 in the Department of Electrical Engineering, USC. In 2006, he was awarded the Viterbi School Deans Doctoral Fellowship from USC.

Panagiotis Tsilifis was born in Aigio, Greece, in 1985. He received his Diploma and M.Sc. degrees in applied mathematical sciences from the National Technical University of Athens, Greece, in 2009 and 2011 respectively. In between, he also spent one year as an exchange student at the Royal Institute of Technology (KTH), in Stockholm, Sweden. He also received an M.A. in applied mathematics in 2014 from the University of Southern California (USC), Los Angeles, CA, USA. Currently he is pursuing his Ph.D. degree in applied mathematics at USC, working under the supervision of Prof. Roger G. Ghanem. His research focuses on Uncertainty Quantification methods including Bayesian approaches to optimal design and inference as well as construction of Polynomial Chaos surrogates with applications in geosciences, reservoir modeling and food microbiology.

Mr. Tsilifis received the Graduate Scholarship by the State Scholarship Foundation (IKY), Greece in 2010 and the Gerondelis Graduate Fellowship in 2013. He is also a SIAM student member and has received twice the SIAM student travel award, in 2015 and 2016.

Shrikanth S. Narayanan (S'88–M'95-SM'02–F'09) is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics, Psychology Neuroscience and Pediatrics and as the founding director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995–2000. At USC he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biomedical problems and applications with direct societal relevance. [http://sail.usc.edu]

Prof. Narayanan is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an Editor in Chief for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and an Editor for the *Computer Speech and Language Journal* and an Associate Editor for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, *APSIPA Transactions on Signal and Information Processing* and the *Journal of the Acoustical Society of America*. He was also previously an Associate Editor of the IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING (2000–2004), IEEE SIGNAL PROCESSING MAGAZINE (2005–2008) and the IEEE TRANSACTIONS ON MULTIMEDIA (2008–2011). He is a recipient of a number of honors including Best Transactions Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010–2011 and ISCA Distinguished Lecturer for 2015–2016. Papers co-authored with his students have won awards including the 2014 Ten-year Technical Impact Award from ACM ICMI and at Interspeech 2015 Nativeness Detection Challenge, 2014 Cognitive Load Challenge, 2013 Social Signal Challenge, Interspeech 2012 Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, InterSpeech 2013 and 2010, InterSpeech 2009 Emotion Challenge, IEE DCOSS 2009, IEEE MMSP 2007, IEEE MMSP 2006, ICASSP 2005 and ICSLP 2002. He has published over 650 papers and has been granted seventeen U.S. patents.