Bayesian Nonnegative Matrix Factorization With Dirichlet Process Mixtures

Caoyuan Li, Hong-Bo Xie, *Member, IEEE*, Kerrie Mengersen, Xuhui Fan, Richard Yi Da Xu, Scott A. Sisson, and Sabine Van Huffel, *Fellow, IEEE*

***Abstract*—Nonnegative Matrix Factorization (NMF) is valuable in many applications of blind source separation, signal processing and machine learning. A number of algorithms that can infer nonnegative latent factors have been developed, but most of these assume a specific noise kernel. This is insufficient to deal with complex noise in real scenarios. In this paper, we present a hierarchical Dirichlet process nonnegative matrix factorization (DPNMF) model in which the Gaussian mixture model is used to approximate the complex noise distribution. Moreover, the model is cast in the nonparametric Bayesian framework by using Dirichlet process mixture to infer the necessary number of Gaussian components. We derive a mean-field variational inference algorithm for the proposed nonparametric Bayesian model. We first test the model on synthetic data sets contaminated by Gaussian, sparse and mixed noise. We then apply it to extract muscle synergies from the electromyographic (EMG) signal and to select discriminative features for motor imagery single-trial electroencephalogram (EEG) classification. Experimental results demonstrate that DPNMF performs better in extracting the latent nonnegative factors in comparison with state-of-the-art methods.**

***Index Terms*—Dirichlet process, nonnegative matrix factorization, nonparametric Bayesian methods, Gaussian mixture model, variational Bayes.**

I. Introduction

MATRIX factorization aims to factorize a given matrix $\mathbf{Y}$ into two low-rank latent factor matrices $\mathbf{U}$ and $\mathbf{V}$ [1]. In many real-life applications, negative factors may contradict physical or physiological reality and lack intuitive meaning. Fortunately, nonnegative matrix factorization (NMF) overcomes the shortcomings of other methods by providing meaningful components with physical or physiological interpretations under nonnegative constraints. Due to the effectiveness of NMF in signal processing and machine learning, much theoretical and applied research effort has focused on NMF in order to solve challenging problems in e.g. signal blind separation, hyperspectral unmixing, audio spectra analysis, text mining, image restoration, spectral clustering, and source localization and neural information extraction in neuroscience [1]–[6].

Following Lee and Seung's seminal NMF paper [7], they later presented two further algorithms based on multiplicative updates to minimize the cost functions using the Frobenius norm and generalized KL-divergence, respectively [8]. However, in practice, the observed data inevitably contains noise and outliers. Several extensions of [8] either modifying the cost function or update rule, or imposing extra constraints, have been proposed to avoid performance degradation. Cichocki *et al.* [9] extended the generalized KL-divergence to generalized Alpha-Beta divergences to improve robustness with respect to noise and outliers. Considering the sparseness of considerable additive noise, imposing an ℓ_1-norm term into the objective function is a popular method to achieve robust NMF [10]. Kong *et al.* [11] proposed a robust formulation of NMF based on the mixed $\ell_{2,1}$-norm, which is trained using multiplicative updates. Shen *et al.* [12] explored a robust NMF which used the hyperbolic tangent ($tanh$) function as a robust loss to evaluate the reconstruction error.

While NMF was originally posited via optimizing a suitable cost function subject to non-negativity constraints, it is well-known that most popular NMF cost function estimates can be interpreted as the maximum likelihood (ML) estimator of a particular statistical model. For instance, the ℓ_2-norm distance measure is related to Gaussian error statistics, while KL- or IS-divergence can be approximated by alternative error statistics given by Poisson or Gamma noise model. Hence, constrained optimization of proper cost functions can be achieved within a statistical framework in terms of maximum likelihood estimation [13]. This results in the development of more conceptually principled approaches based on Bayesian probabilistic interpretations of NMF.

One of the most popular and straightforward Bayesian NMF models under nonnegative constraints uses the Poisson likelihood (noise) function with a conjugate Gamma prior for $\mathbf{U}$

Manuscript received August 28, 2019; revised January 24, 2020; accepted June 10, 2020. Date of publication June 19, 2020; date of current version June 30, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Justin Dauwels. This work was supported in part by the Australian Research Council (ARC) Laureate Program and in part by Centre of Excellence and Discovery Project Schemes. under Grants FL150100150, CE140100049, DP160102544, and FT170100079. *(Corresponding author: Hong-Bo Xie.)*

Caoyuan Li is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China, and also with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: caoyuan.li@student.uts.edu.au).

Hong-Bo Xie and Kerrie Mengersen are with the ARC Centre of Excellence for Mathematical & Statistical Frontiers, Queensland University of Technology, Brisbane, QLD 4001, Australia (e-mail: hongbo.xie@qut.edu.au; k.mengersen@qut.edu.au).

Xuhui Fan and Scott A. Sisson are with the School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2033, Australia (e-mail: xuhui.fan@unsw.edu.au; scott.sisson@unsw.edu.au).

Richard Yi Da Xu is with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: YiDa.Xu@uts.edu.au).

Sabine Van Huffel is with the Department of Electrical Engineering, ESAT-Stadius Division, KU Leuven, Leuven 3001, Belgium (e-mail: sabine.VanHuffel@esat.kuleuven.be).

Digital Object Identifier 10.1109/TSP.2020.3003120

1053-587X © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

and $\mathbf{V}$ [14]. Vincent and Hugo [15] replaced the Gamma prior with an exponential distribution to be coupled with the Poisson likelihood. However, the Poisson distribution is formally defined only for integers, which impairs the statistical interpretation of KL-NMF on uncountable data such as real-valued signals or images.

In addition, as noted by Chien and Yang [16], dependency of the variational lower bound on the model parameters was ignored in the original Bayesian Poisson-Gamma NMF model in [14], which resulted in the inferred parameters not reaching the true optimum of the variational objective. To circumvent these issues, Schachtner *et al.* [17] developed a variational Bayesian NMF model with a Gaussian likelihood and a truncated normal (TN) distribution for the factors prior, by truncating all negative entries and renormalizing the integral to unity. They later modified the prior of the one-factor matrix by deliberately adding a delta peak at the origin of the truncated normal distribution to accommodate both Gaussian and sparse noise [13]. However, this model slightly over-estimates the number of sources under low noise levels, while in the case of high noise levels, the method may fail completely [13]. In order to integrate prior knowledge on the factor matrices, suitable prior distributions, like Gaussian processes [18], Gamma chain priors [19], have also been incorporated into the models.

Although these NMF models, based on either regularization or Bayesian inference, have gained a certain degree of success in various signal processing and machine learning applications, a common shortcoming of most of these methods is that they consider only a single noise kernel or noise distribution. In practice, it is well known that most observations contain complex signal and noise components. For example, physiological signals including electromyogram (EMG) and electroencephalogram (EEG) are weak bioelectric recordings contaminated by white Gaussian noise, motion artefact, crosstalk, power line interference, as well as spurious background spikes [20], [21]. These existing models are insufficient to identify the nonnegative factors or extract signal components accurately. To address this issue, this study aims to develop a hierarchical Bayesian non-negative matrix factorization model. A Gaussian mixture model (GMM), a universal approximator for any continuous distribution [22], is employed to approximate the distribution of the complex noise components. For GMMs, choice of the number of mixture components is an important issue. Insufficient components results in underfitting, while an excessive number of components leads to overfitting. In this paper, we leverage the power of a nonparametric Bayesian technique, i.e., Dirichlet process mixtures, to determine the required number of Gaussian components. The model is thus termed Dirichlet process nonnegative matrix factorization (DPNMF).

Contribution of this work: Our main contributions are summarized as follows. (a) To deal with complex noise in real scenarios, we propose a hierarchical Dirichlet process nonnegative matrix factorization model. (b) We take advantage of the GMM as a universal approximator to fit various types of noise rather than a single noise kernel as with existing NMF models. (c) We employ a nonparametric Bayesian technique, i.e. Dirichlet process, to determine the number of mixture components needed, instead of performing heuristic pruning or ad-hoc methods. (d) We formulate the model into variational Bayesian update rules instead of the usual multiplicative updating rules for NMF. (e) We demonstrate that DPNMF significantly improves the performance of two real-world problems, i.e., muscle synergies extraction and movement imagery EEG classification, which heavily rely on NMF technique.

Related work: Our DPNMF model is somewhat related to both GMM and nonparametric Bayesian models [23], [24]. Penny *et al.* [25] proposed modelling noise in a general linear model (GLM) as a mixture of Gaussians (MoG) rather than using any specific distribution. However, their model is for time series regression, and the model order, i.e. the number of mixture components, has to be pre-specified. While there are many criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and minimum message length (MML) available to determine the model order, this kind of model selection approach could be highly computationally demanding [26]. In contrast, Dirichlet process mixture models (DPMMs) assume the data are generated from an infinite number of components, model selection and parameter learning are simultaneously performed within one training round [27].

Recently, Shao *et al.* [28] presented a Dirichlet process mixture of Gaussians to predict chemical processes. Ren *et al.* [29] introduced a Dirichlet process mixture of principal components to properly choose the latent dimension of the GLM problem in [25]. Lack of conjugacy due to the nonnegativity constraint makes inference for our model very different from the conjugate distributions in [28], [29]. Further, the Dirichlet process mixture of Gaussians is used to model the noise term in our hierarchical model, whereas a GMM and/or DPMM are employed to model the observation itself in [25], [28], [29]. Under the more general framework of nonparametric Bayesian NMF, Porteous *et al.* [30] assumed that there were latent classes for the elements of $\mathbf{Y}$ and regularization should be performed per class. To this end, they used a Dirichlet process mixture to automatically prune the clusters of latent vectors [30]. Similarly, Xuan *et al.* [31] proposed a doubly sparse nonparametric NMF framework where the dependent Indian buffet processes (dIBP) [32] were used to generate two stick weights. However, this nonparametric Bayesian modelling in [30], [31] is implemented on latent vectors rather than the noise term in our model.

Experimental results on both synthetic data sets and two real-world signals (EMG and EEG) demonstrate the proposed DPNMF model achieves improved performance compared to representative baseline methods, including Lee and Seung's NMF algorithm and three other sparse or robust counterparts. The rest of this paper is organized as follows. In Section II, we elaborate on the details of model specification and inference for the DPNMF method in the presence of different noise types. Results on both synthetic and two experimental signals, comparison with state-of-the-art methods, and objective assessments are presented in Section III. Finally, Section IV discusses and concludes the paper.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

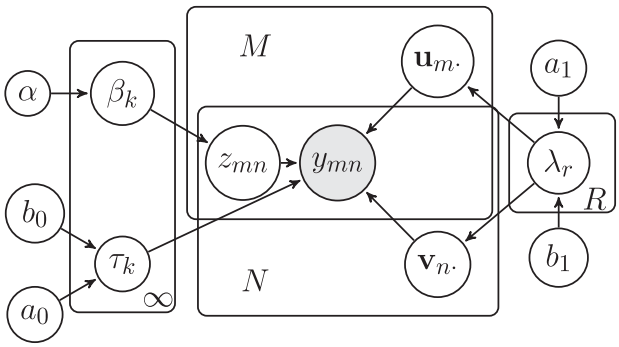


Fig. 1. Directed graphical representation of the DPNMF model.

II. Methods

A. Specification of the DPNMF Model

For the observation matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$, nonnegative matrix factorization can be formulated as decomposing $\mathbf{Y}$ into two latent matrices $\mathbf{U} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{V} \in \mathbb{R}_+^{n\times r}$, whose values are constrained to be positive. In other words, the task is to solve

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^\top + \mathbf{E}, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{m\times n}$ represents a noise term. We take the variational Bayesian approach to this problem. Fig. 1 shows the graphical model of the proposed hierarchical Bayesian Dirichlet process nonnegative matrix factorization with latent variables and their corresponding priors. In particular, we utilize the mixture Gaussian distributions to model the noise:

$$p(e_{mn}) = \sum_{k=1}^{\infty} \theta_k \mathcal{N}(0, \tau_k), \tag{2}$$

where $\theta_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l)$ and β_k is drawn independently from a Beta distribution $\mathcal{B}(1,\alpha)$ according to the stick-breaking construction [27], and τ_k is the precision of the k-th component. In the model, a_0, b_0, α, a_1 and b_1 are hyperparameters. Let z_{mn} be a latent variable that assigns the index of the parameter associated with the entry e_{mn}. The distribution of z_{mn} can be regarded as a multinomial distribution with parameters $\{\theta_1, \ldots, \theta_\infty\}$. In practice, we set a relatively large value, K, as the initial number of Gaussian components to model the noise term.

For the precision τ_k we use a Gamma distribution with shape $a_0 > 0$ and rate $b_0 > 0$,

$$p(\tau_k) \sim \mathcal{G}(a_0, b_0). \tag{3}$$

We choose an exponential prior over $\mathbf{U}$ and $\mathbf{V}$. In order to automatically prune the rank of $\mathbf{U}$ and $\mathbf{V}$, we assign rate parameters λ_r to the exponential prior for $\mathbf{u}_{\cdot r}$ and $\mathbf{v}_{\cdot r}$, the r-th columns of $\mathbf{U}$ and $\mathbf{V}$.

$$u_{mr} \sim f(u_{mr}|\lambda_r), \tag{4}$$

$$v_{nr} \sim f(v_{nr}|\lambda_r), \tag{5}$$

where $f(x|\lambda) = \lambda \exp(-\lambda x)s(x)$ is the density of the exponential distribution, and $s(x)$ is the unit step function, which means when $x < 0$, $s(x) = 0$, otherwise $s(x) = 1$. With the constraint of the same rate parameters λ_r across $\mathbf{u}_{\cdot r}$ and $\mathbf{v}_{\cdot r}$, most of the rate parameters λ_r will be iteratively updated to very large values due to the iterative update of its distribution parameters a_r^* and b_r^*. The corresponding columns of $\mathbf{U}$ and $\mathbf{V}$ are removed since they make little contribution to the approximation of $\mathbf{Y}$, and hence the rank of the latent factors $\mathbf{U}$ and $\mathbf{V}$ are automatically determined.

Combining the likelihood and the priors, we can formulate the joint distribution as

$$\begin{aligned} &p(\mathbf{Y}, \mathbf{U}, \mathbf{V}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}|a_0, b_0, a_1, b_1, \alpha) \\ &\quad = p(\mathbf{Y}|\mathbf{U}, \mathbf{V}, \mathbf{z}, \boldsymbol{\tau}) p(\mathbf{U}|\boldsymbol{\lambda}) p(\mathbf{V}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}|a_1, b_1) \\ &\quad \cdot p(\boldsymbol{\tau}|a_0, b_0) p(\mathbf{z}|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\alpha), \end{aligned} \tag{6}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_R)$.

Since full Bayesian inference using the above joint distribution is computationally intractable, we resort to variational Bayesian inference to deal with this problem [33]. We use $\mathbf{Z}$ to represent the set of all latent variables such that $\mathbf{Z} = (\mathbf{U}, \mathbf{V}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta})$. The approximate posterior distribution is therefore denoted by $q(\mathbf{Z})$. The principle is to define a parameterized family of distributions over the hidden variables and then update the parameters to minimize the Kullback-Leibler (KL) divergence between $q(\mathbf{Z})$ and the true distribution $p(\mathbf{Z}|\mathbf{Y})$, denoted by

$$\min_{q(\mathbf{Z})} KL(q(\mathbf{Z}) \parallel p(\mathbf{Z}|\mathbf{Y})) = \int q(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{Y})} d\mathbf{Z}. \tag{7}$$

Equivalently, this corresponds to the following maximum problem

$$\max_{q(\mathbf{Z})} -KL(q(\mathbf{Z}) \parallel p(\mathbf{Z}|\mathbf{Y})) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}, \mathbf{Y})}{q(\mathbf{Z})p(\mathbf{Y})} d\mathbf{Z}. \tag{8}$$

This can be viewed as estimation of the marginal likelihood $p(\mathbf{Y})$ with a maximal lower bound. With a mean field approximation, $q(\mathbf{Z})$ is factorized with respect to its partitions as [33]

$$q(\mathbf{Z}) = \prod_k q(\mathbf{Z}_k). \tag{9}$$

The expression of the optimal posterior approximation $q(\mathbf{Z}_k)$ with other variables fixed can be denoted as

$$\ln\ q(\mathbf{Z}_k) = \langle \ln\ p(\mathbf{Y}, \mathbf{Z}) \rangle_{\mathbf{Z} \setminus \mathbf{Z}_k} + const, \tag{10}$$

where $\langle \cdot \rangle$ denotes expectation and $const$ denotes a constant which is not dependent on the current variable. $\mathbf{Z} \setminus \mathbf{Z}_k$ means the set of $\mathbf{Z}$ with $\mathbf{Z}_k$ removed. Each variable is updated in turn while holding the others fixed. For this specific DPNMF model, we have

$$\begin{aligned} q(\mathbf{U}, \mathbf{V}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}) = &\prod_{m=1}^{M} \prod_{r=1}^{R} q_{u_{mr}}(\mu_{mr}^U, \tau_{mr}^U) \\ &\prod_{n=1}^{N} \prod_{r=1}^{R} q_{v_{nr}}(\mu_{nr}^V, \tau_{nr}^V) \prod_{r=1}^{R} q_{\lambda_r}(a_r^*, b_r^*) \prod_{k=1}^{K} q_{\beta_k}(\gamma_{k,1}, \gamma_{k,2}) \\ &\prod_{k=1}^{K} q_{\tau_k}(\rho_{k,1}, \rho_{k,2}) \prod_{m=1}^{M} \prod_{n=1}^{N} q_{z_{mn}}(\phi_{mn}), \end{aligned} \tag{11}$$

where the approximation of each entry of $\mathbf{U}$ ($q_{u_{mr}}$) follows a truncated normal distribution with mean μ_{mr}^U and covariance

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

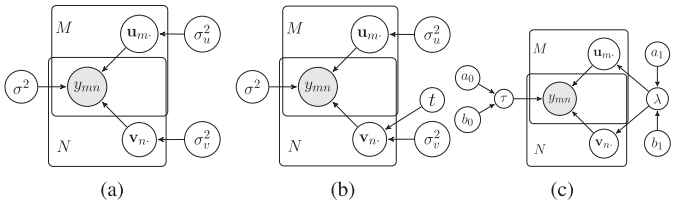


Fig. 2. The equivalent graphic representation of (a) MUNMF and MahNMF, (b) SCNMF, and (c) PSNMF.

τ_{mr}^U restricted to positive values, and similarly for $\mathbf{V}$, τ_k and λ_r follow gamma distribution parametrized by $\rho_{k,1}$, $\rho_{k,2}$ and a_r^*, b_r^*, respectively, $q_{\beta_k}(\gamma_{k,1}, \gamma_{k,2})$ is a beta distribution, and $q_{z_{mn}}(\phi_{mn})$ is a multinomial distribution (see next section for details).

B. Model Inference of DPNMF

By applying the variational Bayesian inference principle of Eq. (10), one can derive the mean-field posterior representation of each latent variable and parameter from the observed data and the assumed priors. We summarize the major update rules in this section while the details of posterior inference are available in the Appendix.

$q(\beta_k)$ follows a Bata distribution,

$$q(\beta_k) = \text{Beta}(\beta_k|\gamma_{k,1}, \gamma_{k,2}), \tag{12}$$

with

$$\gamma_{k,1} = 1 + \sum_{m=1}^{M}\sum_{n=1}^{N}\phi_{mnk}, \tag{13}$$

$$\gamma_{k,2} = \alpha + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=k+1}^{K}\phi_{mnt}, \tag{14}$$

and

$$\mathbb{E}_q(\ln\beta_k) = \psi(\gamma_{k,1}) - \psi(\gamma_{k,1}+\gamma_{k,2}), \tag{15}$$

$$\mathbb{E}_q(\ln(1-\beta_k)) = \psi(\gamma_{k,2}) - \psi(\gamma_{k,1}+\gamma_{k,2}), \tag{16}$$

where ψ denotes the digamma function.

$q(z_{mn})$ obeys a multinomial distribution with parameters ϕ_{mnk} for $k = 1, 2, \cdots K$ given by

$$\phi_{mnk} \propto \exp\left\{\frac{1}{2}(\psi(\rho_{k,1}) - \ln\rho_{k,2}) + \mathbb{E}_q\ln\beta_k - \frac{1}{2}\frac{\rho_{k,1}}{\rho_{k,2}}\mathbb{E}_q\{(y_{mn} - \mathbf{u}_{m\cdot}\mathbf{v}_{n\cdot}^\top)^2\}\right\} + \sum_{t=1}^{k-1}\mathbb{E}_q\ln(1-\beta_t)\}. \tag{17}$$

$q(u_{mr})$ follows a truncated normal distribution (defined on positive values), with mean and precision

$$\mu_{mr}^U = \frac{1}{\tau_{mr}^U}\left(-\langle\lambda_r\rangle + \sum_{n\in\Omega_m}\sum_{k=1}^{K}\phi_{mnk}\langle\tau_k\rangle \times\left(y_{mn} - \sum_{r'\neq r}\langle u_{mr'}\rangle\langle v_{nr'}\rangle\right)\langle v_{nr}\rangle\right), \tag{18}$$

TABLE I
NOISE PARAMETER SETTING FOR THE SYNTHETIC DATA SETS. U DENOTES THE UNIFORMLY DISTRIBUTED NOISE FOLLOWED BY ITS RANGE

	$\mathcal{N}(0, 0.5^2)$	$\mathcal{N}(0, 0.1^2)$	U[-5, 5]	SNR
Gaussian Noise	100%	0	0	5
Sparse Noise	0	0	30%	10
Mixture Noise	60%	20%	20%	15

$$\tau_{mr}^U = \sum_{n\in\Omega_m}\sum_{k=1}^{K}\phi_{mnk}\langle\tau_k\rangle\langle v_{nr}^2\rangle, \tag{19}$$

where $\Omega = \{(m,n) : m = 1, \ldots, M, n = 1, \ldots, N\}$ and $\Omega_m = \{m|(m,n) \in \Omega\}$.

Similarly, the element v_{nr} of factor matrix $\mathbf{V}$ follows truncated normal distribution with mean and precision

$$\mu_{nr}^V = \frac{1}{\tau_{nr}^V}\left(-\langle\lambda_r\rangle + \sum_{m\in\Omega_n}\sum_{k=1}^{K}\phi_{mnk}\langle\tau_k\rangle \times\left(y_{mn} - \sum_{r'\neq r}\langle u_{mr'}\rangle\langle v_{nr'}\rangle\right)\langle u_{mr}\rangle\right), \tag{20}$$

$$\tau_{nr}^V = \sum_{m\in\Omega_n}\sum_{k=1}^{K}\phi_{mnk}\langle\tau_k\rangle\langle u_{mr}^2\rangle. \tag{21}$$

where $\Omega_n = \{n|(m,n) \in \Omega\}$.

The precision parameter τ_k follows a Gamma distribution

$$q(\tau_k) = G(\tau_k|\rho_{k,1}, \rho_{k,2}), \tag{22}$$

where

$$\rho_{k,1} = a_0 + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\phi_{mnk},$$

$$\rho_{k,2} = b_0 + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\phi_{mnk}\mathbb{E}_q[(y_{mn} - \mathbf{u}_{m\cdot}\mathbf{v}_{n\cdot}^\top)^2]. \tag{23}$$

Here $\mathbb{E}_q[(y_{mn} - \mathbf{u}_m\mathbf{v}_n^\top)^2]$ is given by

$$\mathbb{E}_q[(y_{mn} - \mathbf{u}_m\mathbf{v}_n^\top)^2] = \left(y_{mn} - \sum_{r=1}^{R}\langle u_{mr}\rangle\langle v_{nr}\rangle\right)^2 + \sum_{r=1}^{R}(\langle u_{mr}^2\rangle\langle v_{nr}^2\rangle - \langle u_{mr}\rangle^2\langle v_{nr}\rangle^2). \tag{24}$$

After updating the parameters $\boldsymbol{\rho}$ via Eq. (23), the weight coefficients $\boldsymbol{\theta}$ of the K clusters are also updated. A θ_k smaller than a pre-defined threshold means that the probability of some entries to be represented by this cluster is very rare. We therefore prune those clusters, and the noise is represented by a limited number of Gaussians.

λ_r follows a Gamma distribution with shape parameter

$$a_r^* = a_1 + M + N, \tag{25}$$

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

TABLE II
MEAN RELATIVE ERRORS ± STD OF FIVE ALGORITHMS UNDER THREE NOISE TYPES (TABLE I) WITH THREE DIFFERENT INITIAL RANKS, r. BEST RESULTS ARE SHOWN IN BOLD

Noise	r	DPNMF	MUNMF	SCNMF	MahNMF	PSNMF
Gaussian	5	**0.0746**±0.0008	0.0897±0.0195	0.6360±0.0008	0.0934±0.0015	0.0766±0.0008
	10	**0.1009**±0.0004	0.1372±0.0043	0.4675±0.0030	0.1329±0.0020	0.1049±0.0004
	15	**0.1207**±0.0013	0.1767±0.0022	0.2853±0.0029	0.1688±0.0038	0.1249±0.0008
Sparse	5	**0.0212**±0.0069	0.0535±0.0161	0.6232±0.0143	0.0572±0.0042	0.0350±0.0058
	10	0.0468±0.0064	0.0815±0.0136	0.4585±0.0096	0.1348±0.0071	**0.0275**±0.0065
	15	0.0330±0.0053	0.0975±0.0046	0.2542±0.0084	0.1766±0.010	**0.0233**±0.0043
Mixture	5	**0.0109**±0.0040	0.0551±0.0095	0.6229±0.0120	0.0929±0.0074	0.0449±0.0039
	10	**0.0089**±0.0052	0.0856±0.0096	0.4578±0.0093	0.1623±0.0109	0.0355±0.0057
	15	**0.0061**±0.0064	0.0965±0.0062	0.2585±0.0072	0.2004±0.0142	0.0300±0.0054

and inverse scale parameter

$$b_r^* = b_1 + \sum_{m=1}^{M} \langle u_{mr} \rangle + \sum_{n=1}^{N} \langle v_{nr} \rangle. \tag{26}$$

We update each parameter in turn while holding others fixed. Convergence of the DPNMF algorithm to a local minimum can be guaranteed after a finite number of iterations following standard variational Bayes arguments [27].

III. RESULTS

In this Section, we empirically compare the proposed DPNMF model with several state-of-the-art methods including MUNMF in [8], a sparseness-constrained NMF (SCNMF) [6], a Bayesian NMF (PSNMF) [34], and the outlier-robust MahNMF [35]. MUNMF is the standard NMF with multiplicative update rules to minimize KL divergence. SCNMF is a sparseness constrained NMF method with a Gaussian noise distribution. We set the sparseness level at 0.5 across activation coefficients so as to be consistent with the neural sparse coding scheme [36]. MahNMF minimizes the Manhattan distance between $\mathbf{Y}$ and $\mathbf{UV}^\top$ for modelling the heavy tailed Laplacian noise. PSNMF is a variational Bayesian sparse NMF model which assumes the column-wise sparseness of two latent matrices with Gaussian noise distribution. The equivalent graphical representation of these models is shown in Fig. 2.

A. Results on Synthetic Data

We first compare the performance of each NMF method using synthetic data sets. The elements of the two latent matrices $\mathbf{U}$ and $\mathbf{V}$ are generated from unit mean exponential distributions, and the matrices have three different ranks $r = 5$, 10 and 15. The ground-truth $\mathbf{Y}_0$ is generated as the product of $\mathbf{U} \in \mathbb{R}_+^{500\times r}$ and $\mathbf{V} \in \mathbb{R}_+^{500\times r}$. Three types of noise are considered for $\mathbf{E}$, i.e. Gaussian, sparse, and mixed (see Table I). Gaussian noise is directly generated with 5 dB SNR. There is no direct way to generate sparse or mixed noise with a specific SNR. We first generate a sparse noise such that 30% of the elements are located uniformly at random and the amplitudes are drawn uniformly at random in the range $[-5, 5]$ [37], [38]. The mixed noise is generated from $0.6 \times \mathcal{N}(0, 0.5^2) + 0.2 \times \mathcal{N}(0, 0.1^2) + 0.2 \times U[-5, 5]$. The SNRs for these two cases are then normalized to 10 and 15 dB, respectively. The initial number of Gaussian components K is set to 300 for DPNMF, which is large enough to model a wide range of noise structures.

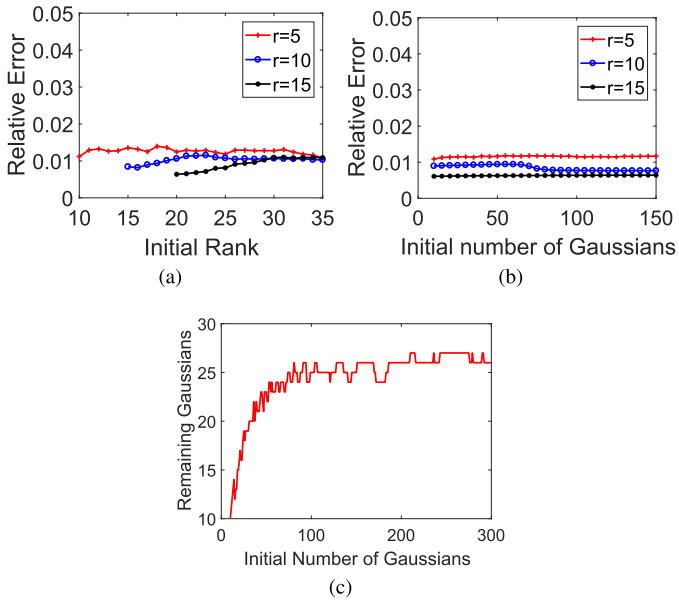


Fig. 3. Effect of initial number of Gaussians and rank, r on the relative error performance of DPNMF for mixture noise.

Because some methods for comparison are unable to automatically infer the rank r, we set the ground-truth rank r as a known model input parameter. We use the relative error of the Frobenius norm with respect to the ground truth, defined by [37]–[39]

$$Error = \frac{|\mathbf{Y}_0 - \overline{\mathbf{U}}\overline{\mathbf{V}}^\top|_F}{|\mathbf{Y}_0|_F}, \tag{27}$$

as the metric to quantify the performance of each algorithm, where $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ are the recovered latent matrices. For each noise setting, we run each method 20 times with different random input of $\mathbf{Y}_0$. The resulting mean relative errors are shown in Table II.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

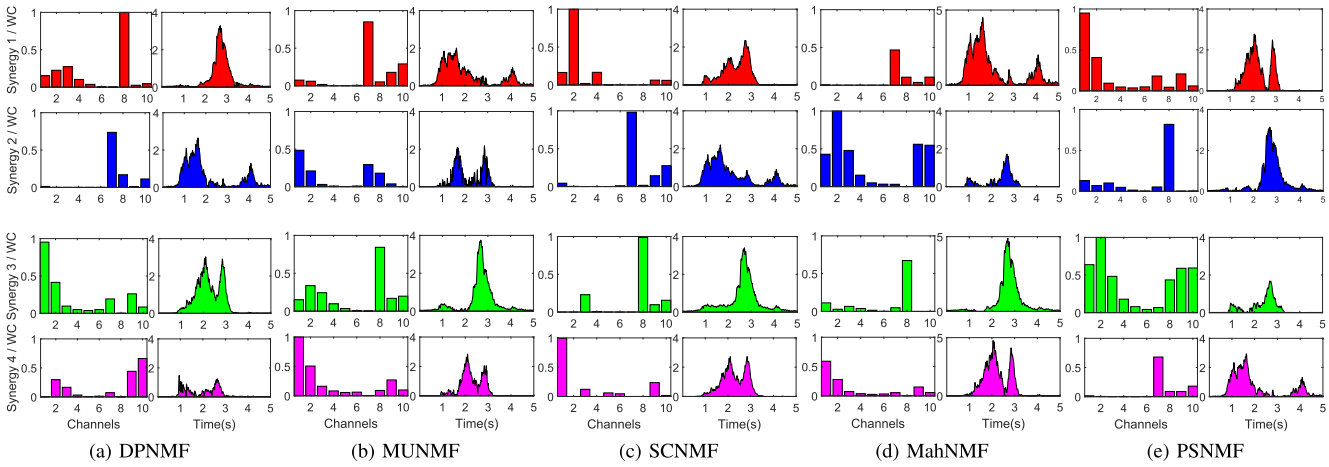


Fig. 4. Four muscle synergies (odd columns) and the weight coefficients (WC) (even columns) extracted via the five NMF models for a representative EMG recording of wrist extension movement.

It can be seen that DPNMF yields the lowest relative error under most noise and rank settings. For Gaussian noise, PSNMF achieves comparable performance to DPNMF, however the relative errors for the other methods are much higher than DPNMF and PSNMF, even given the correct rank input, which is difficult to estimate in practice. The relative error of these methods increases significantly when the initial rank deviates from the ground-truth rank. In the case of sparse noise, it is evident that the performance of MUNMF and MahNMF methods degrade. This is reasonable since only Gaussian noise is considered in these two models. The DPNMF algorithm performs better than PSNMF for rank 5, and is slightly inferior to DPNMF for rank 10 and 15. Finally, for the mixture noise, it is not surprising that DPNMF outperforms all methods with significantly lower errors. When the noise type is simple, PSNMF achieves comparable and even slightly better results than DPNMF. However, the performance of PSNMF is significantly inferior to DPNMF when the noise is more complicated, due to the ability of DPNMF to fit unknown complex noise, as well as automatically tune the rank.

We now empirically evaluate the stability of DPNMF in terms of the varying initial number of Gaussians components K and rank r under mixture noise. Fig. 3(a) shows the mean relative error in terms of the different initial rank for (true) $r = 5, 10$, and 15, respectively. In each case, the mean relative error has tiny fluctuation as the initial rank varies. However the relative error remains stable at low values as the initial rank increases. Fig. 3(b) shows the effect of the initial number of Gaussian components on the mean relative error for $r = 5, 10, 15$. Similar to the effect of fixing K in Fig. 3(a), the mean relative error is almost flat as the initial number of components varying from 10 to 150. Fig. 3(c) shows the estimated number of Gaussian components versus the initial number, for $r = 5$. As the initial number of components increases, the estimated number of components stabilise at around 25 components, which is adequate to model the mixture noise. Overall we may conclude from Fig. 3 that DPNMF is fairly robust to initialization and input parameters. The computational complexity of the factor matrices $\mathbf{U}$ and $\mathbf{V}$ in Eqs. (18)−(21) is $O(Kmnr)$ in each iteration, where K is the number of initial mixture components, mn represents the data size, and r the rank. Since this is the major computational cost of the model, DPNMF hence has linear complexity w.r.t. the data size. It should be noted that because latent factors are pruned out in the first few iterations, r reduces rapidly in practice. With a recent 3.6 GHz CPU, it costs around 10 mins to fit the model with $K = 300$ mixture components.

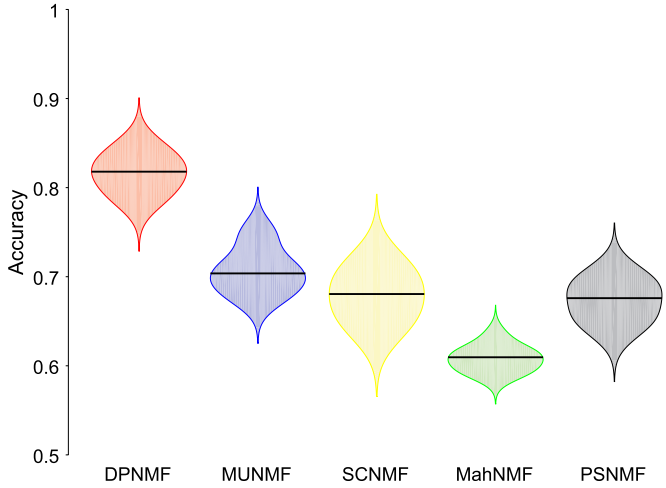


Fig. 5. Violin plots for the accuracy of the full synergy for wrist movements classification for five NMF models. Black line represents the mean accuracy.

B. *Extraction of Muscle Synergies*

In neuroscience, it is supposed that the central neural system controls muscle synergies, or groups of co-activated muscles, rather than individual muscles, to organize any simple or complex actions and movements. Muscle synergies, i.e. the nonnegative factor vectors of $\mathbf{U}$, extracted from multi-channel EMG signals using NMF have been widely applied in the human-machine interfaces, prosthetic controls, neural system disease

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

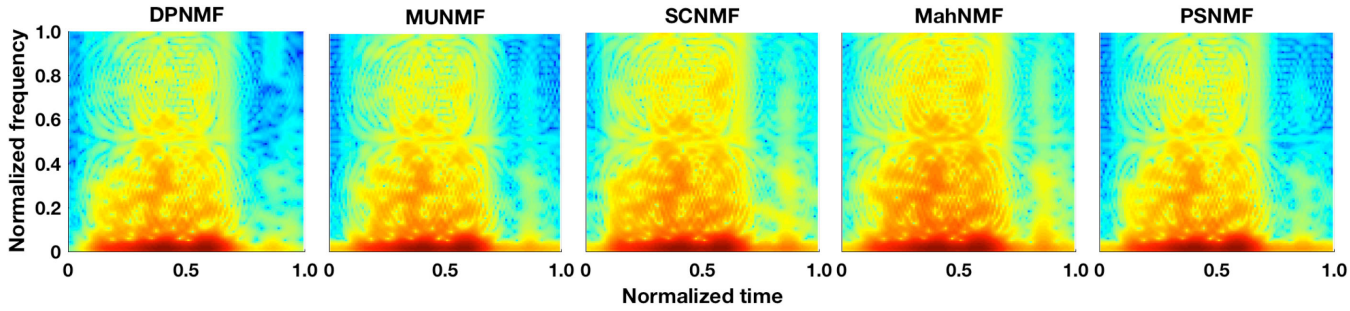


Fig. 6. The spectrograms of reconstructed muscle activity corresponding to channel 1 EMG signal of Fig. 4 by five algorithms.

diagnoses, and stroke rehabilitation [40], [41]. As discussed in the Introduction, EMG is a typical signal contaminated by multiple types of noise. Here we compare the performance of DPNMF with competitor methods to extract the muscle synergies. Since the ground truth synergies are not available, we follow the study in [40] to evaluate the classification accuracy of synergies extracted by DPNMF in terms of recognising six wrist motions. We used the Ninapro first dataset [41] which consists of EMG recordings for wrist, hand and finger movements. Each movement/task has 10 repetitions from 27 healthy subjects. To this end, we partition the Ninapro dataset into training (60%; 6 repetitions of each task) and testing (40%; 4 repetitions) sets for each subject. For each NMF technique, the synergies are estimated from training repetitions for each task independently. The synergies are then used to train support vector machines (SVM) to classify six movements: wrist flexion and extension, wrist radial and ulnar deviation, and wrist supination and pronation. DPNMF automatically selects the number of synergies (i.e. the rank r), while we assign $r = 4$ as the fixed number of synergies for other methods following the recommendations in [40]. The other four repetitions of each task are used to test those classifiers. We randomly select the training and test samples 10 times to produce a mean predictive accuracy.

Fig. 4 shows the four synergies and the corresponding weight coefficients of a representative wrist extension movement extracted via the five NMF models. From Fig. 4, Synergies 1 and 2 extracted by DPNMF, indicated on the first column of Fig. 4, have relatively large values for channels 7 and 8, respectively. This result demonstrates that DPNMF clearly indicates the primary drive of these movements coming from channels 7 and 8, i.e., the front arm. This characterisation with the physiological origin of muscles involved in wrist extension is consistent with other methods [41]. Fig. 5 is a violin graph of the classification accuracy for each method with DPNMF yielding the largest accuracy (mean 82.4%). The mean accuracy of the other methods is below the lowest decile of that of DPNMF.

To further ascertain the performance of DPNMF and the other competitive methods, we use the extracted muscle synergies to reconstruct the muscle activity. Fig. 6 shows the spectrograms of recovered EMG signal corresponding to channel 1 of Fig. 4. One can find from the left and right of the spectrogram that DPNMF removes Gaussian noise. In addition, the movement artifacts and sparse noise occupying the right bottom of the spectrogram, where the major muscle activation has completed, have been

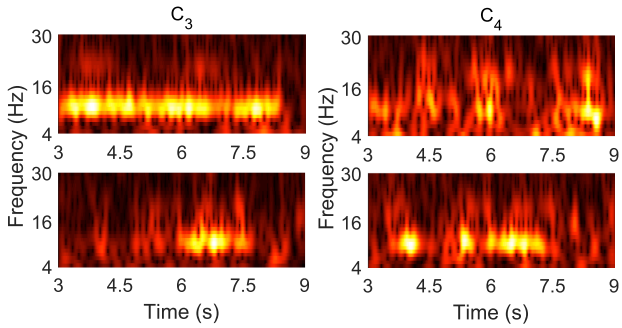


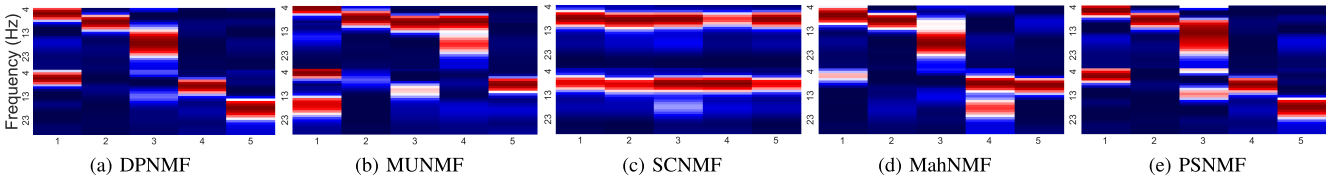
Fig. 7. Contours of Morlet wavelet coefficients of channels C_3 and C_4 for representative left imagery (upper panels) and a right imagery (bottom panels) EEGs.

significantly attenuated. MUNMF and PSNMF perform as well as DPNMF to remove the Gaussian noise, indicated by the blue stripe at the left and right part of each panel. However, movement artifacts and sparse noise are still visible at the left bottom of both panels. As for SCNMF and MahNMF, residual Gaussian noise are clearly observed in yellow in the right part of both panels. Fig. 6 further demonstrates the effectiveness of DPNMF over the other competitive noise models to model EMG signal contaminated by complex noise.

C. Classification of Motor Imagery EEG

We investigate the performance of DPNMF in a motor imagery EEG classification problem in the brain-machine interface, which heavily relies on the nonnegative matrix factorization techniques. A BCI competition data set is used [42] (provided by the Department of Medical Informatics, Institute for Biomedical Engineering, Graz University of Technology, Austria). This data set was recorded from a healthy subject (female, 25 yrs) during a feedback session. The experiment included seven runs with 40 trials each. The 140 labeled trials are used for training and 140 unlabeled trials for testing with subjects performing left/right imagery hand movements. Each trial has a duration of 9 seconds with the first 3 seconds as preparation period. We therefore only analyze the EEG in the final 6 seconds of the task. The aim is to dynamically identify the actual class with high probability as soon as possible during the trail, rather than to provide the class label using the entire data segment. We

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.



(a) DPNMF (b) MUNMF (c) SCNMF (d) MahNMF (e) PSNMF

Fig. 8. Basis vectors extracted from imagery movement two-channel EEG training samples using the five NMF models.

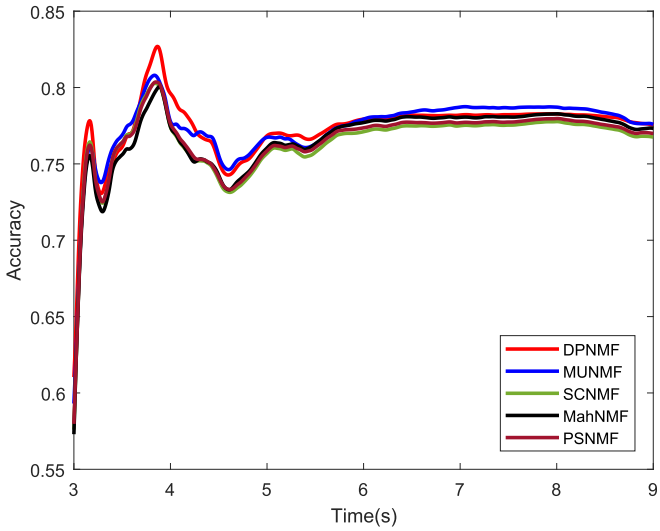


Fig. 9. Time course of the classification accuracy using the encoding variable matrix extracted from each NMF model.

follow [42] and only use signals in channels C_3 and C_4 since channel C_z contains little relevant information.

We first decompose the EEG into wavelet coefficients using complex Morlet wavelet over the frequency range 4–30 Hz with the scale step of 1 Hz. Fig. 7 shows the contours of wavelet coefficients for representative left and right imagery EEGs. One can visually identify the different μ rhythm (8–12 Hz) and β rhythm (18–25 Hz) during the two different (left and right imaginary) movements. For simplicity, most previous studies assumed that the noise distribution of the wavelet coefficients follows the same Gaussian distribution as in time domain [43]. We use DPNMF and the other four state-of-the-art methods to generate the factor matrices (which are specifically termed basis vectors ($\mathbf{U}$) and encoding variable matrix ($\mathbf{V}$) in BCI) using the training sampling matrix $\mathbf{Y} \in \mathbb{R}^{54\times TP}$, where T and P denote the number of trials and data points within the EEG.

For the test sample $\tilde{\mathbf{Y}}$, its encoding variable matrix $\tilde{\mathbf{V}}$ is recovered by the product of $pi(\mathbf{U})$ and $\tilde{\mathbf{Y}}$, where $pi(\mathbf{U})$ is the pseudo-inverse of $\mathbf{U}$. With the encoding variable matrices $\mathbf{V}$ and $\tilde{\mathbf{V}}$, we can make the decision to have the class label at any single point in time using the maximal posterior probability. Readers are referred to [44] for the training and testing details of this online Bayesian classifier. Fig. 8 shows the basis vectors of training samples obtained by each NMF model. Visual inspection of Fig. 8 shows that the basis vectors of DPNMF clearly reveal some useful EEG characteristics: (1) μ rhythm (8–12 Hz); (2) β rhythm (18–22 Hz); (3) sensori-motor rhythm (12–16 Hz). MUNMF, MahNMF, and PSNMF can also reveal these EEG rhythms, however, redundant features across columns are clearly visible for these three methods. As for SCNMF, the basis vectors do not seem cover all rhythms, perhaps due to the sparse constraint. Fig. 9 illustrates the time course of five methods to continuously classify a single-trial imagery movements on test data. Overall the five methods have similar classification accuracy time course profiles. Accuracy is initially low at the beginning of the EEG, peaks twice at around 3.2 and 3.8 seconds before levelling out from around 5.5 seconds. Of all methods, DPNMF and MUNMF perform the best, with one of these models providing the best classification accuracy at any time point. Table III shows the average maximal accuracy during the time course over the 140 test samples. DPNMF provides the highest average classification accuracy of 83.12% compared to the 79.36% − 81.22% maximum of other methods. Overall DPNMF provides a highly competitive classification accuracy through the duration of the time course.

TABLE III
MEAN AND STANDARD DEVIATION OF THE AVERAGE MAXIMAL ACCURACY DURING THE TIME COURSE TO CLASSIFY EEG TEST SAMPLES BY FIVE ALGORITHMS

	DPNMF	MUNMF	SCNMF	MahNMF	PSNMF
Mean	83.12%	81.09%	79.36%	81.17%	81.22%
STD	1.23%	1.47%	1.34%	1.56%	1.21%

IV. DISCUSSION AND CONCLUSION

We have proposed a new NMF method (DPNMF) by using a Dirichlet process to model noise as an infinite mixture of Gaussian distributions, under the variational Bayesian framework. To make variational inference feasible, a stick-breaking representation of the Dirichlet process and a factorization assumption for the posterior distribution were used. Besides, the order of latent matrices is automatically pruned by placing an ARD prior. Compared with the existing, state-of-the-art NMF methods, which assume a certain noise distribution (e.g. Gaussian or sparse noise) on the data, our method can extract the latent factor matrices under more complex noise distributions. The effectiveness of the DPNMF model has been demonstrated by synthetic data with artificial noise and by muscle synergies extraction and imagery movement classification problems with real noise. The proposed DPNMF model yields much better performance result over existing methods in its capability to accurately extract the latent structure and elaborately model the multimodal noise configurations from observed signals. The model is implemented by variational mean-field inference in

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

this study. In cases without computational cost restrictions, it is straightforward to estimate the parameters of this model using Gibbs sampling. The Dirichlet process tends to produce slightly more Gaussian components than needed, which can be avoided by setting a slightly larger threshold for λ. This can also be addressed by modifying the model using the powered Chinese restaurant process [45]. The dimensionality of $\mathbf{U}$ and $\mathbf{V}$ is relatively small in the two real applications in this study. Future work will further validate the model on datasets with large-scale matrices, for example, the speech spectrogram. Although we have focused on two applications in biomedical engineering and neuroscience, we anticipate that the DPNMF model will have a broader range of applications in signal processing and machine learning. A number of NMF algorithms have been extended to nonnegative tensor decomposition. To leverage the capacity of Dirichlet processes to model complex noise and tensors to jointly represent multiway data, we are currently working to develop a Dirichlet process-based Bayesian nonnegative tensor decomposition model, which is expected to improve the performance of tensor signal processing tasks. This includes, for example, the tensor EEG involving both temporal and trial coordinates, or tensor EMG in the temporal-spatial-spectral domain.

APPENDIX

We first list some statistics used in updating the parameters of the DPNMF algorithm. For any variable a we have

$$\langle a^2\rangle = \langle a\rangle^2 + \mathrm{Var}_q[a]. \tag{28}$$

For a Gamma distribution

$$x \sim \mathrm{Gamma}(a,b), \qquad \langle x\rangle = \frac{a}{b}.$$

For a truncated (on $\mathbb{R}_+$) normal distribution $c \sim \mathcal{TN}(\mu,\tau)$ where

$$\mathcal{TN}(c|\mu,\tau) = \begin{cases} \frac{\sqrt{\frac{\tau}{2\pi}}\exp\left\{-\frac{\tau}{2}(c-\mu)^2\right\}}{1-\Phi(-\mu\sqrt{\tau})} & \text{if } c \geq 0 \\ 0 & \text{if } c < 0 \end{cases},$$

$$\langle c\rangle = \mu + \frac{1}{\sqrt{\tau}} g(-\mu\sqrt{\tau}),$$

$$\mathrm{Var}[c] = \frac{1}{\tau}(1-\delta(-\mu\sqrt{\tau})), \tag{29}$$

where $\Phi(x)$ is the cumulative distribution function of $\mathcal{N}(0,1)$, $g(x) = \phi(x)/(1-\Phi(x))$, and $\delta(x) = g(x)(g(x)-x)$. $\phi(x) = \frac{1}{2\pi}\exp\{-\frac{1}{2}x^2\}$ is the density function of $\mathcal{N}(0,1)$.

We then derive the mean-field posterior approximation for each latent variable.

$$\ln q(\beta_k) = \mathbb{E}_q[\ln p(\beta_k|\alpha)] + \sum_{m=1}^{M}\sum_{n=1}^{N} \mathbb{E}_q[\ln p(z_{mn}=k|\boldsymbol{\theta})] + \text{const.} \tag{30}$$

Since $p(z_{mn}|\boldsymbol{\theta})$ is a multinomial distribution, expanding $\boldsymbol{\theta}$ and using the assumption that $q(\beta_K = 1) = 1$, we can infer that $1-\beta_K = 0$, and $q(z_{mn} > K) = 0$. Then we have

$$\mathbb{E}_q[\ln p(z_{mn}|\boldsymbol{\theta})] = \mathbb{E}_q\left[\ln\left(\prod_{k=1}^{\infty}(1-\beta_k)^{\mathbb{I}(z_{mn}>k)}\beta_k^{\mathbb{I}(z_{mn}=k)}\right)\right]$$
$$= \sum_{k=1}^{K} q(z_{mn}>k)\mathbb{E}_q\ln(1-\beta_k) + q(z_{mn}=k)\mathbb{E}_q\ln\beta_k. \tag{31}$$

Since $q(z_{mn} > K) = 0$, then $q(z_{mn} > k) = \sum_{t=k+1}^{K} q(z_{mn}=t)$. Substitute Eq. (31) into Eq. (30) and simplifying gives

$$\ln q(\beta_k) = \text{const} + \left(\sum_{m=1}^{M}\sum_{n=1}^{N} q(z_{mn}=k)\right)\ln\beta_k + \left(\alpha - 1 + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=k+1}^{K} q(z_{mn}=t)\right)\ln(1-\beta_k). \tag{32}$$

To update the parameters related to z_{mn} based on the stick-breaking process, first note that

$$\ln q(z_{mn}=k) = \mathbb{E}_{\mathbf{Z}/z_{mn}=k}[\ln p(\mathbf{Y}|\mathbf{U},\mathbf{V},\mathbf{z},\boldsymbol{\tau}) + \ln p(\mathbf{z}|\boldsymbol{\beta})]$$
$$= \mathbb{E}_q \ln\sqrt{\frac{\tau_{z_{mn}=k}}{2\pi}}\exp\left\{-\frac{\tau_{z_{mn}=k}}{2}(y_{mn}-\mathbf{u}_{m\cdot}\mathbf{v}_{n\cdot}^{\top})^2 + \mathbb{I}(z_{mn}>k)\mathbb{E}_q\ln(1-\beta_k) + \mathbb{I}(z_{mn}=k)\mathbb{E}_q\ln\beta_k\right\}$$
$$= \frac{1}{2}[\psi(\rho_{k,1}) - \ln\rho_{k,2}] - \frac{\rho_{k,1}}{2\rho_{k,2}}\mathbb{E}_q\left\{(y_{mn}-\mathbf{u}_{m\cdot}\mathbf{v}_{n\cdot}^{\top})^2 + \sum_{t=1}^{k-1}\mathbb{E}_q\ln(1-\beta_t) + \mathbb{I}(z_{mn}=k)\mathbb{E}_q\ln\beta_k\right\}. \tag{33}$$

From Eq. (33), it is clear that $q(z_{mn})$ follows a multinomial distribution.

The variational posterior for u_{mr} can then be expressed as

$$q(u_{mr}) \propto \exp\{\mathbb{E}_q[\ln p(\mathbf{Y}|\mathbf{U},\mathbf{V},\mathbf{z},\boldsymbol{\tau}) + \ln p(u_{mr}|\lambda_r)]\}$$
$$\propto s(u)\exp\left\{\ln(\lambda_r\exp\{-\lambda_r u_{mr}\}) + \mathbb{E}_q\sum_{n\in\Omega_m}\sum_{k=1}^{K} \times\left(p(z_{mn}=k)\ln\left[\sqrt{\frac{\tau_k}{2\pi}}\exp\left\{-\frac{\tau_k}{2}(y_{mn}-\mathbf{u}_m\mathbf{v}_n^{\top})^2\right\}\right]\right)\right\}$$
$$\propto \exp\left\{-\frac{u_{mr}^2}{2}\left[\sum_{n\in\Omega_m}\sum_{k=1}^{K}\phi_{mnk}\langle\tau_k\rangle\langle v_{nr}^2\rangle\right] + u_{mr}\left[-\langle\lambda_r\rangle + \sum_{n\in\Omega_m}\sum_{k=1}^{K}\phi_{mnk}\langle\tau_k\rangle(y_{mn} - \sum_{r'\neq r}\langle u_{mr'}\rangle\langle v_{nr'}\rangle)\langle v_{nr}\rangle\right]\right\}s(u)$$
$$\propto \exp\left\{-\frac{\tau_{mr}^u}{2}(u_{mr}-\mu_{mr}^u)^2\right\}\times s(u)$$
$$\propto \mathcal{TN}(u_{mr}|\mu_{mr}^U,\tau_{mr}^U), \tag{34}$$

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

Eq. (34) demonstrates that $q(u_{mr})$ follows a truncated normal distribution.

For the precisions τ_k, we have

$$\begin{aligned}\ln q(\tau_k) &= \mathbb{E}_q[\ln p(\tau_k|a_0,b_0)] + \text{const} \\ &+ \prod_{m=1}^{M}\prod_{n=1}^{N} q(z_{mn}=k)\mathbb{E}_q[\ln p(y_{mn}|\mathbf{U},\mathbf{V},z_{mn}=k,\boldsymbol{\tau})] \\ &= \text{const} + \left(a_0 + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} q(z_{mn}=k)\right)\ln\tau_k \\ &- \left(b_0 + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} q(z_{mn}=k)\mathbb{E}_q\left[(y_{mn}-\mathbf{u}_{m\cdot}\mathbf{v}_{n\cdot}^{\top})^2\right]\right)\tau_k. \end{aligned} \tag{35}$$

This is a Gamma distribution.

Finally, we update the rate parameter of the exponential prior related to the factor matrices $\mathbf{U}$ and $\mathbf{V}$. We have

$$\begin{aligned} q(\lambda_r) &\propto \exp\left\{\mathbb{E}_q\left(\sum_{m=1}^{M}\ln p(u_{mr}|\lambda_r)\right)\right. \\ &\left. + \mathbb{E}_q\left(\sum_{n=1}^{N}\ln p(v_{nr}|\lambda_r)\right) + \mathbb{E}_q(\ln p(\lambda_r|a_1,b_1))\right\} \\ &\propto \exp\left\{(M+N+a_1-1)\ln\lambda_r\right. \\ &\left. - \left(\sum_{m=1}^{M}\langle u_{mr}\rangle + \sum_{n=1}^{N}\langle v_{nr}\rangle + b_1\right)\lambda_r\right\}. \end{aligned} \tag{36}$$

Obviously, λ_r follows a Gamma distribution,

$$q(\lambda_r) = \text{Gamma}(\lambda_r|a_r^*,b_r^*). \tag{37}$$

ACKNOWLEDGMENT

The code associated with this paper will be released on GitHub. Hong-Bo Xie would like to thank Prof Bin Gao, University of Electronic Science and Technology of China, for his helpful advice on Bayesian parameter inference in this paper.

REFERENCES

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

[2] K. Huang and N. D. Sidiropoulos, "Putting nonnegative matrix factorization to the test: A tutorial derivation of pertinent Cramer–Rao bounds and performance benchmarking," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 76–86, May 2014.

[3] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.

[4] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optim., Kernels, Support Vector Mach.*, vol. 12, no. 257, pp. 257–291, 2014.

[5] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie, "Nonnegative matrix and tensor factorizations: An algorithmic perspective," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 54–65, May 2014.

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

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

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[9] A. Cichocki, S. Cruces, and S.-I. Amari, "Generalized Alpha-Beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, 2011.

[10] B. Shen, B.-D. Liu, Q. Wang, and R. Ji, "Robust nonnegative matrix factorization via l_1 norm regularization by multiplicative updating rules," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 5282–5286.

[11] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using $l_{2,1}$ norm," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 673–682.

[12] X. Shen, X. Zhang, L. Lan, Q. Liao, and Z. Luo, "Another robust NMF: Rethinking the hyperbolic tangent function and locality constraint," *IEEE Access*, vol. 7, pp. 31 089–31 102, 2019.

[13] R. Schachtner *et al.*, "A new Bayesian approach to nonnegative matrix factorization: Uniqueness and model order selection," *Neurocomputing*, vol. 138, pp. 142–156, 2014.

[14] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, vol. 2009, 2009, Art. no. 785152.

[15] V. Renkens *et al.*, "Automatic relevance determination for nonnegative dictionary learning in the Gamma-Poisson model," *Signal Process.*, vol. 132, pp. 121–133, 2017.

[16] J.-T. Chien and P.-K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 1, pp. 185–195, Jan. 2016.

[17] R. Schachtner, G. Poeppel, A. M. Tomé, and E. W. Lang, "A Bayesian approach to the Lee–Seung update rules for NMF," *Pattern Recognit. Lett.*, vol. 45, pp. 251–256, 2014.

[18] M. N. Schmidt and H. Laurberg, "Nonnegative matrix factorization with Gaussian process priors," *Comput. Intell. Neurosci.*, vol. 2008, pp. 361705–361705, 2008.

[19] M. Schmidt and H. Laurberg, "Nonnegative matrix factorization with gaussian process priors," *Comput. Intell. Neurosci.*, 2008, pp. 1825–1828.

[20] X. Chen *et al.*, "Removal of muscle artifacts from the EEG: A review and recommendations," *IEEE Sensors J.*, vol. 19, no. 14, pp. 5353–5368, Jul. 2019.

[21] S. Thongpanja, A. Phinyomark, F. Quaine, Y. Laurillau, C. Limsakul, and P. Phukpattaranont, "Probability density functions of stationary surface EMG signals in noisy environments," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 7, pp. 1547–1557, Jul. 2016.

[22] V. Maz'ya and G. Schmidt, "On approximate approximations using Gaussian kernels," *IMA J. Numer. Anal.*, vol. 16, no. 1, pp. 13–29, 1996.

[23] D. M. Titterington, A. F. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. Hoboken, NJ, USA: Wiley, 1985.

[24] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, 1984.

[25] W. Penny, J. Kilner, and F. Blankenburg, "Robust Bayesian general linear models," *Neuroimage*, vol. 36, no. 3, pp. 661–671, 2007.

[26] S. A. Sisson and Y. Fan, "Towards automating model selection for a mark-recapture-recovery analysis," *J. Roy. Stat. Soc., Ser. C*, vol. 58, pp. 247–266, 2009.

[27] D. M. Blei *et al.*, "Variational inference for Dirichlet process mixtures," *Bayesian Anal.*, vol. 1, no. 1, pp. 121–143, 2006.

[28] W. Shao, Z. Ge, and Z. Song, "Quality variable prediction for chemical processes based on semisupervised Dirichlet process mixture of Gaussians," *Chem. Eng. Sci.*, vol. 193, pp. 394–410, 2019.

[29] J. Ren, K. Li, and C. Chen, "Supervised Dirichlet process mixtures of principal component analysis," *Neurocomputing*, vol. 305, pp. 15–26, 2018.

[30] I. Porteous, A. Asuncion, and M. Welling, "Bayesian matrix factorization with side information and Dirichlet process mixtures," in *Proc. 24th AAAI Conf. Artif. Intell.*, 2010, pp. 563–568.

[31] J. Xuan, J. Lu, G. Zhang, R. Y. D. Xu, and X. Luo, "Doubly nonparametric sparse nonnegative matrix factorization based on dependent Indian buffet processes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1835–1849, May 2018.

[32] S. Williamson, P. Orbanz, and Z. Ghahramani, "Dependent Indian buffet processes," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 924–931.

[33] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

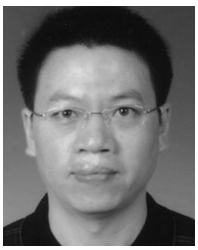
[34] J. L. Hinrich and M. Mørup, "Probabilistic sparse non-negative matrix factorization," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 488–498.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.

[35] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor, "MahNMF: Manhattan non-negative matrix factorization," 2012, *arXiv:1207.3438*.
[36] M. Beyeler, E. L. Rounds, K. D. Carlson, N. Dutt, and J. L. Krichmar, "Neural correlates of sparse coding and dimensionality reduction," *PLOS Comput. Biol.*, vol. 15, no. 6, 2019, Art. no. e1006908.
[37] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, Dec. 2011.
[38] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *J. ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.
[39] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, Aug. 2012.
[40] A. Ebied, E. Kinney-Lang, L. Spyrou, and J. Escudero, "Evaluation of matrix factorisation approaches for muscle synergy extraction," *Med. Eng. Phys.*, vol. 57, pp. 51–60, 2018.
[41] M. Atzori *et al.*, "Characterization of a benchmark database for myoelectric movement classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 1, pp. 73–83, Jan. 2015.
[42] B. Blankertz *et al.*, "The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1044–1051, Jun. 2004.
[43] W. Zeng, X. Fu, C. Hu, and Y. Du, "Wavelet denoising with generalized bivariate prior model," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 20 863–20 887, 2018.
[44] S. Lemm, C. Schafer, and G. Curio, "BCI competition 2003-data set III: Probabilistic modeling of sensorimotor μ rhythms for classification of imaginary hand movements," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1077–1080, Jun. 2004.
[45] J. Lu, M. Li, and D. Dunson, "Reducing over-clustering via the powered Chinese restaurant process," 2018, *arXiv:1802.05392*.



Caoyuan Li received the bachelor's degree in software engineering from Tongji University, Shanghai, China, in 2013. Currently, he is working towards the dual doctoral degree in both Beijing Institute of Technology and University of Technology Sydney. His main interests include machine learning, image processing and data mining.



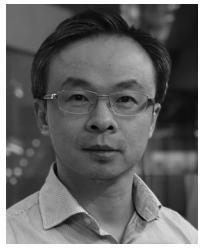
Hong-Bo Xie (Member, IEEE) received the Ph.D. degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China. He is currently a Vice-Chancellor's Senior Research Fellow/Lecturer at the ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology, Brisbane, Australia. His current research interests include machine learning, signal and image processing, nonlinear time series analysis, and their applications in biomedical engineering. He has published over 70 peer-reviewed journal and conference papers.



Kerrie Mengersen is a Professor at the School of Mathematical Sciences, Queensland University of Technology (QUT), Australia. She was elected as a Fellow of Australian Academy of Science in 2018. She was the National President of the Statistical Society of Australia from 2011 to 2012, and the President of International Society for Bayesian Analysis from 2015 to 2017. In 2016, QUT awarded the title of Distinguished Professor to Professor Mengersen in recognition of her outstanding achievements, both nationally and internationally, in mathematics and statistical research. She has over 300 refereed journal publications, over 40 keynote and invited international conference presentations, and attraction of over 30 large research grants.



Xuhui Fan received the bachelor's degree in mathematical statistics from the University of Science and Technology of China, Hefei, China, in 2010, and the Ph.D. degree in computer science from the University of Technology Sydney, Chippendale, NSW, Australia, in 2015. His current research interests include stochastic random partition and Bayesian nonparametrics.



Richard Yi Da Xu is currently an Associate Professor with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia. He has authored about 50 papers, including IEEE TRANSACTIONS NO IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, Pattern Recognition, ACM Transactions on Knowledge Discovery from Data, Association for the Advancement of Artificial Intelligence, and The International Conference on Image Processing. His current research interests include machine learning, computer vision, and statistical data mining.



Scott A. Sisson is currently a Professor and Head of Statistics in the School of Mathematics and Statistics at the University of New South Wales. He is internationally recognized for his work in computational and Bayesian statistics, and in particular for developing inferential techniques for computationally intractable models and challenging data. He is a past President of the Statistical Society of Australia, and is currently Chair of the Australian Chapter of the International Society of Bayesian Analysis (ISBA) and a member of the Australian National Committee for Mathematical Sciences. Scott is a previous winner of the Moran Medal (Australian Academy of Science), the G. N. Alexander Medal (Engineers Australia) and the J. G. Russell Award (Australian Academy of Science), and is a previous Australian Research Council Queen Elizabeth II Research Fellow. His current research interests include Bayesian statistics and computation, big data, extreme value theory and applied science.



Sabine Van Huffel (Fellow, 2009) is a Professor at the Department of Electrical Engineering, KU Leuven, since 2002. She was elected as a Fellow of Royal Flemish Academy of Belgium for Sciences and the Arts in 2017. The research topics of Sabine Van Huffel are fundamental/theoretical as well as application oriented and are performed in the domain of (multi)linear algebra, (non)linear signal analysis, classification and system identification with special focus to the development of numerically reliable and robust algorithms for improving medical diagnostics. Sabine Van Huffel has the following publications record: two monographs, about 304 articles in peer reviewed international journals, about 257 conference papers, four edited books, seven edited journal special issues, 21 book chapters. She was a Guest Professor at Stanford University, CA, USA and at Uppsala University, Sweden, and a Visiting Fellow and Scientist at the University of Minnesota, MN, USA.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 03,2020 at 08:53:22 UTC from IEEE Xplore. Restrictions apply.