A Statistically Principled and Computationally Efficient Approach to Speech Enhancement using Variational Autoencoders : Supporting Document

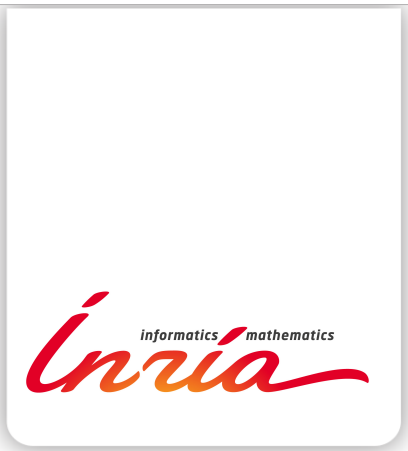
Manuel Pariente, Antoine Deleforge, Emmanuel Vincent

A Statistically Principled and Computationally Efficient Approach to Speech Enhancement using Variational Autoencoders : Supporting document

Manuel Pariente, Antoine Deleforge, Emmanuel Vincent

RESEARCH REPORT
N° 9268
April 2019
Project-Team MULTISPEECH

ISSN 0249-6399 ISRN INRIA/RR--9268--FR+ENG



A Statistically Principled and Computationally Efficient Approach to Speech Enhancement using Variational Autoencoders : Supporting document

Manuel Pariente, Antoine Deleforge, Emmanuel Vincent

Project-Team MULTISPEECH

Research Report n° 9268 — April 2019 — 8 pages

Abstract: Recent studies have explored the use of deep generative models of speech spectra based of variational autoencoders (VAEs), combined with unsupervised noise models, to perform speech enhancement. These studies developed iterative algorithms involving either Gibbs sampling or gradient descent at each step, making them computationally expensive. This paper proposes a variational inference method to iteratively estimate the power spectrogram of the clean speech. Our main contribution is the analytical derivation of the variational steps in which the encoder of the pre-learned VAE can be used to estimate the variational approximation of the true posterior distribution, using the very same assumption made to train VAEs. Experiments show that the proposed method produces results on par with the aforementioned iterative methods using sampling, while decreasing the computational cost by a factor 36 to reach a given performance.

Key-words: Speech enhancement, variational autoencoders, variational Bayes, non-negative matrix factorization

RESEARCH CENTRE
NANCY – GRAND EST

615 rue du Jardin Botanique
CS20101
54603 Villers-lès-Nancy Cedex

1 Model

This document provides the details of the analytical derivation of the algorithm presented in [1]. We first remind the statistical model used in [1].

We use f to denote the frequency index and t to denote the time frame index. Independently for $(f,t) \in \{0,...,F-1\} \times \{0,...,N-1\}$, the single channel observation of the mixture is modeled by

$$x_{ft} = s_{ft} + n_{ft} + \epsilon_{ft}, \tag{1}$$

where x_{ft}, s_{ft} and n_{ft} are the short term Fourier transform (STFT) coefficients of the mixture, the speech source and the noise source respectively, and $\epsilon_{ft} \sim \mathcal{N}_c(0, \sigma_\epsilon^2)$ is introduce to prevent $p(x_{ft}|s_{ft}, b_{ft})$ from being a Dirac, we have

$$x_{ft}|s_{ft}, b_{ft} \sim \mathcal{N}_c(s_{ft} + b_{ft}, \sigma_\epsilon^2), \tag{2}$$

where $\mathcal{N}_c$ denotes the univariate proper complex Gaussian defined in Section 3. σ_ϵ^2 will be set to 0, once the variational updates are obtained.

The noise STFT coefficients are modelled using Non-Negative Matrix Factorization (NMF) [2] :

$$n_{ft} \sim \mathcal{N}_c(0, (\boldsymbol{W}\boldsymbol{H})_{ft}). \tag{3}$$

with $\boldsymbol{W} \in \mathbb{R}_+^{F \times K}$, $\boldsymbol{H} \in \mathbb{R}_+^{K \times N}$, K being the rank of the NMF model.

We define $\boldsymbol{s}_t = [s_{1t}, ..., s_{Ft}]^T$, $\boldsymbol{s} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_N]$, $\boldsymbol{n} = [n_{1t}, ..., n_{Ft}]^T$, $\boldsymbol{n}_t = [\boldsymbol{n}_1, ..., \boldsymbol{n}_N]$, $\boldsymbol{x}_t = [x_{1t}, ..., x_{Ft}]^T$, $\boldsymbol{x} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$ and $\boldsymbol{H}_t = [H_{1t}, ..., H_{Kt}]^T$, where $(.)^T$ denotes the transpose operator.

The speech STFT coefficients are modeled using a variational autoencoder (VAE) [3], we have

$$\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1}), \tag{4}$$

$$s_{ft}|\boldsymbol{z}_t; \theta \sim \mathcal{N}_c(0, \sigma_f^2(\boldsymbol{z}_t)), \tag{5}$$

$$z_{l,t}|\boldsymbol{s}_t; \phi \sim \mathcal{N}(\tilde{\mu}_l(|\boldsymbol{s}_t|^2), \tilde{\sigma}_l^2(|\boldsymbol{s}_t|^2)), \tag{6}$$

where $\boldsymbol{z}_t = [z_{1t}, ..., z_{Lt}]^T$ is the latent variable and $L < F$.

We remind that σ_f^2; $\tilde{\mu}_l$ and $\tilde{\sigma}_l^2$, are non-linear functions implemented by deep neural networks (DNNs) with parameters θ and ϕ respectively. They are learned by maximizing the marginal log-likelihood $\log p_\theta(\boldsymbol{s})$. All the STFT frames are considered to be independent, we have

$$\log p_\theta(\boldsymbol{s}) = \sum_t \log p_\theta(\boldsymbol{s}_t). \tag{7}$$

We can then maximize the marginal likelihoods of individual STFT frames, we write

$$\log p_\theta(\boldsymbol{s}_t) = \mathcal{D}_{KL}(q_\phi(\boldsymbol{z}_t|\boldsymbol{s}_t)||p_\theta(\boldsymbol{z}_t|\boldsymbol{s}_t)) + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{s}_t), \tag{8}$$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{s}_t) = \mathbb{E}_{q_\phi(\boldsymbol{z}_t|\boldsymbol{s}_t)}[\log p_\theta(\boldsymbol{s}_t|\boldsymbol{z}_t)] - \mathcal{D}_{KL}(q_\phi(\boldsymbol{z}_t|\boldsymbol{s}_t)||p(\boldsymbol{z}_t)), \tag{9}$$

where $\mathcal{D}_{KL}(q||p) = -\mathbb{E}_q[\log(p/q)]$ denotes the Kullback-Leibler (KL) divergence. Based on (4), (5) and (6), we rewrite $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{s}_t)$:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \overset{c}{=} \sum_f \mathbb{E}_{q_\phi(\boldsymbol{z}_t|s_{ft})} \left[d_{IS}(|s_{ft}|^2, \sigma_f^2(\boldsymbol{z}_t)) \right] + \frac{1}{2} \sum_l \log(\tilde{\sigma}_l^2(|\boldsymbol{s}_t|^2) - \tilde{\mu}_l^2(|\boldsymbol{s}_t|^2) - \tilde{\sigma}_l^2(|\boldsymbol{s}_t|^2),$$

where $d_{IS}(x; y) = x/y - \log(x/y) - 1$ denotes the Itakura-Saito divergence, and $\overset{c}{=}$ denotes equality up to a constant.

2 Inference

Given an observation $\boldsymbol{X} = \{x_{ft}\}_{(ft)\in(B)}$, our goal is now to maximize the log-likelihood of $\boldsymbol{X}$ given the mixture model (1), the generative model of speech ((4) and (5)) and the noise's NMF model (3).
We consider $\boldsymbol{y}_t = \{\boldsymbol{s}_t, \boldsymbol{n}_t, \boldsymbol{z}_t\}$ to be the set of latent variables, $\boldsymbol{\Theta}_t = \{\boldsymbol{W}, \boldsymbol{H}_t\}$ the parameters of the model. We introduce a variational distribution $r(\boldsymbol{y}_t)$ and write the following decomposition:

$$\log p(\boldsymbol{x}_t; \boldsymbol{\Theta}_t) - \mathcal{D}_{KL}(r(\boldsymbol{y}_t)||p(\boldsymbol{y}_t|\boldsymbol{x}_t; \boldsymbol{\Theta}_t)) = \mathcal{L}(r, \boldsymbol{\Theta}_t), \tag{10}$$

where $\mathcal{L}(r, \boldsymbol{\Theta}_t)$ follows

$$\mathcal{L}(r, \boldsymbol{\Theta}_t) = \mathbb{E}_{r(\boldsymbol{y}_t; \boldsymbol{\Theta}_t)} \left[\log \frac{p(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\Theta}_t)}{r(\boldsymbol{y}_t; \boldsymbol{\Theta}_t)} \right]. \tag{11}$$

We suppose that the variation distribution $r(\boldsymbol{y}_t)$ factorizes as

$$r(\boldsymbol{s}_t, \boldsymbol{n}_t, \boldsymbol{z}_t) = r(\boldsymbol{s}_t, \boldsymbol{n}_t) r(\boldsymbol{z}_t) = \prod_f r(s_{ft}, n_{ft}) \prod_l r(z_{lt}). \tag{12}$$

Given the independence of s_t and n_t, and n_t and z_t, we can write the complete data likelihood as

$$p(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\Theta}_t) = p(\boldsymbol{x}_t|\boldsymbol{s}_t, \boldsymbol{n}_t) p(\boldsymbol{s}_t|\boldsymbol{z}_t) p(\boldsymbol{z}_t) p(\boldsymbol{n}_t; \boldsymbol{\Theta}_t). \tag{13}$$

We can then iteratively maximize $\mathcal{L}(r, \boldsymbol{\Theta}_t)$ with respect to the factorized distributions $r(\boldsymbol{s}_t, \boldsymbol{b}_t)$ and $r(\boldsymbol{z}_t)$, and the NMF parameters $\boldsymbol{\Theta}_t = \{\boldsymbol{W}, \boldsymbol{H}_t\}$. As given by the equation (10.9) in [4], the variational distributions $r(s_{ft}, n_{ft})$ and $r(z_{lt})$ can be updated using

$$\log r(s_{ft}, n_{ft}) \overset{c}{=} \mathbb{E}_{r(\boldsymbol{z}_t)}[\log p(x_{ft}, s_{ft}, n_{ft}, \boldsymbol{z}_t; \boldsymbol{\Theta}_t)], \tag{14}$$

$$\log r(z_{lt}) \overset{c}{=} \mathbb{E}_{r(\boldsymbol{s}_t, \boldsymbol{n}_t)}[\log p(\boldsymbol{x}_t, \boldsymbol{s}_t, \boldsymbol{n}_t, z_{lt}; \boldsymbol{\Theta}_t)]. \tag{15}$$

$\boldsymbol{W}$ and $\boldsymbol{H}$ can be updated by maximizing the following

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{old}) &\overset{c}{=} \mathbb{E}_{r(\boldsymbol{s}, \boldsymbol{b}, \boldsymbol{z}; \boldsymbol{\Theta}^{old})}[\log p(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{b}, \boldsymbol{z}; \boldsymbol{\Theta})] \\ &\overset{c}{=} \mathbb{E}_{r(\boldsymbol{s}, \boldsymbol{b}; \boldsymbol{\Theta}^{old})}[\log p(\boldsymbol{b}; \boldsymbol{\Theta})], \end{aligned} \tag{16}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{H}\}$.

2.1 E-(s,n) step

We define $\sigma^2_{n,ft} = (\boldsymbol{WH})_{ft}$ to make the notation less cluttered. Removing the terms independent from s_{ft} and n_{ft} in (14), we have

$$\begin{aligned}\log r(s_{ft}, n_{ft}) &\stackrel{c}{=} \log p(x_{ft}|s_{ft}, n_{ft}) + \mathbb{E}_{r(\boldsymbol{z}_t)}[\log p(s_{ft}|\boldsymbol{z}_t)] + \log p(n_{ft}; \sigma^2_{n,ft}) \\ &\stackrel{c}{=} \log p(x_{ft}|s_{ft}, n_{ft}) - \mathbb{E}_{r(\boldsymbol{z}_t)}\left[\frac{1}{\sigma^2_f(\boldsymbol{z}_t)}\right]|s_{ft}|^2 + \log p(n_{ft}; \sigma^2_{n,ft}). \end{aligned} \tag{17}$$

We can define γ^2_{ft} as

$$\frac{1}{\gamma^2_{ft}} = \mathbb{E}_{r(\boldsymbol{z}_t)}\left[\frac{1}{\sigma^2_f(\boldsymbol{z}_t)}\right] \approx \sum_{d=1}^{D}\left[1/\sigma^2_f(\boldsymbol{z}_t^{(d)})\right], \tag{18}$$

where $\{\boldsymbol{z}_n^{(d)}\}_{d=1,..,D}$ are randomly drawn from $r(\boldsymbol{z}_t)$.

We recognize $r(s_{ft}, n_{ft})$ to be a product of Gaussian, up to the normalization constant. We have

$$r(s_{ft}, n_{ft}) \sim \mathcal{N}_c(x_{ft}; s_{ft} + n_{ft}, \sigma^2_\epsilon)\mathcal{N}_c(n_{ft}; 0, \sigma^2_{n,ft})\mathcal{N}_c(s_{ft}; 0, \gamma^2_{ft}) \tag{19}$$

$$\sim \mathcal{N}_c(x_{ft}; s_{ft} + n_{ft}, \sigma^2_\epsilon)\mathcal{N}_c([s_{ft}, n_{ft}]; \mathbf{0}, \boldsymbol{\Psi}), \tag{20}$$

with $\boldsymbol{\Psi}$ defined as

$$\boldsymbol{\Psi}_{ft} = \begin{bmatrix}\gamma^2_{ft} & 0 \\ 0 & \sigma^2_{n,ft}\end{bmatrix}. \tag{21}$$

Finally, with $\boldsymbol{S}_{ft} = [s_{ft}, n_{ft}]^T$ and $(.)^H$ denoting the Hermitian transpose, we can rewrite $r(s_{ft}, n_{ft})$ using Bayes' theorem, as in [5]

$$r(\boldsymbol{S}_{ft}) = \frac{1}{|\pi\boldsymbol{\Sigma}|}\exp\left(-(\boldsymbol{S}_{ft} - \boldsymbol{\mu}_{ft})^H\boldsymbol{\Sigma}^{-1}_{ft}(\boldsymbol{S}_{ft} - \boldsymbol{\mu}_{ft})\right), \tag{22}$$

where, using $\mathbf{M} = [1, 1]$, $\boldsymbol{\mu}_{ft}$ and $\boldsymbol{\Sigma}_{ft}$ are defined as

$$\boldsymbol{\Sigma}_{ft} = \left(\boldsymbol{\Psi}^{-1}_{ft} + \frac{\mathbf{M}^T\mathbf{M}}{\sigma^2_\epsilon}\right)^{-1} \tag{23}$$

$$\boldsymbol{\mu}_{ft} = \frac{\boldsymbol{\Sigma}_{ft}R^T}{\sigma^2_\epsilon}x_{ft}. \tag{24}$$

With $A = \boldsymbol{\Psi}^{-1}_{ft}$, $U = \mathbf{M}^T$, $C = 1/\sigma^2_\epsilon$ and $V = \mathbf{M}$, we can use the Woodbury matrix identity :

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1} \tag{25}$$

to rewrite $\boldsymbol{\Sigma}_{ft}$ as

$$\boldsymbol{\Sigma}_{ft} = \frac{\gamma^2_{ft}\sigma^2_{n,ft}}{\gamma^2_{ft} + \sigma^2_{n,ft} + \sigma^2_\epsilon}\begin{bmatrix}\frac{\sigma^2_{n,ft}}{\sigma^2_{n,ft}+\sigma^2_\epsilon} & -1 \\ -1 & \frac{\gamma^2_{ft}}{\gamma^2_{ft}\sigma^2_\epsilon}\end{bmatrix} \xrightarrow[\sigma^2_\epsilon \to 0]{} \frac{\gamma^2_{ft}\sigma^2_{n,ft}}{\gamma^2_{ft} + \sigma^2_{n,ft}}\begin{bmatrix}1 & -1 \\ -1 & 1\end{bmatrix}. \tag{26}$$

Using (24), we have

$$\boldsymbol{\mu}_{ft} = \begin{bmatrix} \frac{\gamma_{ft}^2}{\gamma_{ft}^2+\sigma_{n,ft}^2+\sigma_\epsilon^2} \\ \frac{\sigma_{n,ft}^2}{\gamma_{ft}^2+\sigma_{n,ft}^2+\sigma_\epsilon^2} \end{bmatrix} x_{ft} \xrightarrow[\sigma_\epsilon^2 \to 0]{} \begin{bmatrix} \frac{\gamma_{ft}^2}{\gamma_{ft}^2+\sigma_{n,ft}^2} \\ \frac{\sigma_{n,ft}^2}{\gamma_{ft}^2+\sigma_{n,ft}^2} \end{bmatrix} x_{ft}, \tag{27}$$

in which we recognize the Wiener filter applied to x_{ft}.

We define $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ with

$$r(\boldsymbol{s}_t, \boldsymbol{n}_t) \sim \prod_f \mathcal{N}_c(\boldsymbol{\mu}_{ft}, \boldsymbol{\Sigma}_{ft}) = \mathcal{N}_c(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t). \tag{28}$$

We note $[\boldsymbol{\mu}_{t,s}, \boldsymbol{\mu}_{t,n}] = \boldsymbol{\mu}_t$ and define $\boldsymbol{\Sigma}_{t,ss}$ and $\boldsymbol{\Sigma}_{t,nn}$ to be the diagonal terms of $\boldsymbol{\Sigma}_t$.

2.2 E-z step

After removing the terms independent from $\boldsymbol{z}_t$ from (15), we can update $r(\boldsymbol{z}_t)$ using

$$\log r(\boldsymbol{z}_t) \overset{c}{=} \log p(\boldsymbol{z}_t) + \mathbb{E}_{r(\boldsymbol{s}_t, \boldsymbol{n}_t)} \left[\log p(\boldsymbol{s}_t|\boldsymbol{z}_t)\right] \tag{29}$$

$$\overset{c}{=} \log p(\boldsymbol{z}_t) - \frac{1}{2} \sum_f \left[\log(2\pi\sigma_f^2(\boldsymbol{z}_t)) + \frac{|\boldsymbol{\mu}_{ft,ss}|^2 + \boldsymbol{\Sigma}_{ft,ss}}{\sigma_f^2(\boldsymbol{z}_t)} \right] \tag{30}$$

$$\overset{c}{=} \log p(\boldsymbol{z}_t) + p\left(\boldsymbol{s}_t = \sqrt{|\boldsymbol{\mu}_{t,s}|^2 + \boldsymbol{\Sigma}_{t,ss}}\, |\boldsymbol{z}_t\right), \tag{31}$$

where $p(\boldsymbol{s}_t|\boldsymbol{z}_t) = \prod_f p(s_{ft}|\boldsymbol{z}_t)$.
We can invert (31) using Bayes' theorem to obtain

$$r(\boldsymbol{z}_t) = p\left(\boldsymbol{z}_t|\boldsymbol{s}_t = \sqrt{|\boldsymbol{\mu}_{t,s}|^2 + \boldsymbol{\Sigma}_{t,ss}}\right). \tag{32}$$

As maximizing (8) minimizes $\mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{s})||p_\theta(\boldsymbol{z}|\boldsymbol{s}))$, we can use the probabilistic encoder $q_\phi(\boldsymbol{z}|\boldsymbol{s})$ as an approximation of the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{s})$. We further assume that this still holds for $\boldsymbol{s}$ of the form $\boldsymbol{s}_t = \sqrt{|\boldsymbol{\mu}_{t,s}|^2 + \boldsymbol{\Sigma}_{t,ss}}$. Finally, we have

$$r(\boldsymbol{z}_t) \approx q_\phi(\boldsymbol{z}_t|\boldsymbol{s}_t = \sqrt{|\boldsymbol{\mu}_{t,s}|^2 + \boldsymbol{\Sigma}_{t,ss}}) \tag{33}$$

$$\sim \mathcal{N}(\tilde{\mu}_l(|\boldsymbol{\mu}_{t,s}|^2 + \boldsymbol{\Sigma}_{t,ss}), \tilde{\sigma}_l^2(|\boldsymbol{\mu}_{t,s}|^2 + \boldsymbol{\Sigma}_{t,ss})). \tag{34}$$

2.3 M-step

We now maximize $\mathcal{L}(r, \boldsymbol{\Theta}_t)$ with respect to $\boldsymbol{\Theta}_t$ with fixed $r(\boldsymbol{y}_t)$ using (16) :

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{old}) \overset{c}{=} \mathbb{E}_{r(\boldsymbol{s}, \boldsymbol{b}; \boldsymbol{\Theta}^{old})}[\log p(\boldsymbol{b}; \boldsymbol{\Theta})]$$

$$\overset{c}{=} \sum_{ft} d_{IS}\left(|\boldsymbol{\mu}_{ft,n}|^2 + \boldsymbol{\Sigma}_{ft,nn}, (\boldsymbol{W}\boldsymbol{H})_{ft}\right) \tag{35}$$

We obtain classic NMF's multiplicative updates as in [6] :

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W}^T\Big((\boldsymbol{W}\boldsymbol{H})^{\odot -2} \odot \mathbf{V}\Big)}{\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{H})^{\odot -1}}, \tag{36}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\Big((\boldsymbol{W}\boldsymbol{H})^{\odot -2} \odot \mathbf{V}\Big)\boldsymbol{H}^T}{(\boldsymbol{W}\boldsymbol{H})^{\odot -1}\boldsymbol{H}^T}, \tag{37}$$

where $\boldsymbol{V} = \{|\boldsymbol{\mu}_{ft,n}|^2 + \boldsymbol{\Sigma}_{ft,nn}\}_{(ft)}$.

3 Distribution definitions

Proper complex Gaussian: The complex proper Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, noted $\mathcal{N}_c((\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined as:

$$\mathcal{N}_c(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\pi \det \boldsymbol{\Sigma}} \exp\left(-(\boldsymbol{x} - \boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{38}$$

where H denotes the Hermitian transpose. In the univariate case, it simplifies to

$$\mathcal{N}_c(x; \mu, \sigma) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x - \mu|^2}{\sigma^2}\right). \tag{39}$$

Real-valued Gaussian The real-valued multivariate distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, noted $\mathcal{N}((\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined as:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{40}$$

which, in the univariate case, simplifies to

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \tag{41}$$

References

[1] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders," *Submitted to INTERSPEECH*, 2019.

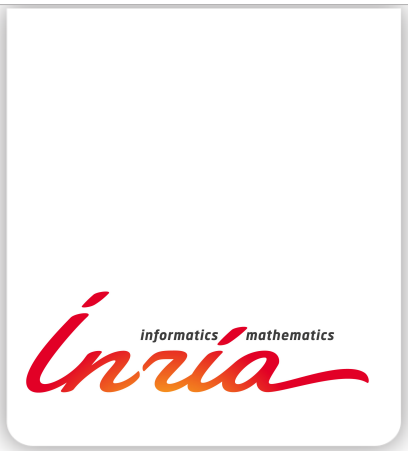
[2] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press, 2002.

[6] C. Fevotte, N. Bertin, and J. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.



RESEARCH CENTRE
NANCY – GRAND EST

615 rue du Jardin Botanique
CS20101
54603 Villers-lès-Nancy Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399