Unsupervised Blind Source Separation with Variational Auto-Encoders

Julian Neri, Roland Badeau, Philippe Depalle

Unsupervised Blind Source Separation with Variational Auto-Encoders

Julian Neri
CIRMMT, McGill University
Montréal, Canada
julian.neri@mcgill.ca

Roland Badeau
LTCI, Télécom Paris
Institut Polytechnique de Paris, France
roland.badeau@telecom-paris.fr

Philippe Depalle
CIRMMT, McGill University
Montréal, Canada
philippe.depalle@mcgill.ca

***Abstract*—Supervised source separation requires expensive synthetic datasets containing clean, ground truth-source signals, while unsupervised separation requires only data mixtures. Existing unsupervised methods still use supervision to avoid over-separation and compete with fully supervised methods. We present a new method of completely unsupervised single-channel blind source separation, based on variational auto-encoding, that automatically learns the correct number of sources in data mixtures and quantitatively outperforms the existing methods. A deep inference network disentangles (separates) data mixtures into low-dimensional latent source variables. A deep generative network individually decodes each latent source into its source signal, such that their sum represents the given mixture. Qualitative and quantitative results from separation experiments on pairs of randomly mixed MNIST handwritten digits and mixed audio spectrograms demonstrate that our method outperforms state-of-the-art unsupervised and semi-supervised methods, showing promise as a solution to this long-standing problem in computer vision and audition.**

***Index Terms*—blind source separation, Bayesian inference, unmixing, latent variable model, universal sound separation**

I. Introduction

Unsupervised blind source separation (BSS) has attracted much attention over the last 30 years and, because of its many exciting applications in computer vision and audition, remains as a key research problem today [1]. The problem can be stated as follows: estimate the underlying sources from a given single-channel (monophonic) mixture without knowing the true number of sources and without ever having access to clean target source signals for model training. An example unsupervised system for sounds would be able to learn from the vast collection of existing recorded sounds including music and films and generalize to separate unheard sounds with ideal accuracy. This goal is motivated, in part, by human ability; we are able to effortlessly disentangle an acoustic scene, focus on a single speaker in a busy room (the cocktail party effect) and perceive multiple objects superimposed on a single image [2].

Unsupervised methods that learn BSS solely from mixed data are valuable because datasets that include both the target mixture and the clean sources are rare and expensive to create. Supervised methods that take advantage of these tailored datasets have matured to the point of being deployable for practical applications [3], [4]. Despite growing research interest, there are currently much fewer works on unsupervised BSS. In contrast to supervised separation, this highly under-determined problem requires strong prior information about the sources to enable separation, limiting its application. Previous work has suggested a variety of models and prior information, as detailed in Section II.

A variational auto-encoder (VAE) [5] is a powerful generative model that leverages fast, amortized inference to encode high-dimensional data in a structured, low-dimensional latent space. VAEs have proven successful in interesting image and audio applications, including audio synthesis [6], interpolation, de-noising, and disentangling a musical note's timbre from pitch [7].

This paper presents a VAE for unsupervised blind source separation of high-dimensional data. Given a mixture signal, our VAE infers separated latent source encodings, then individually generates source signals from them with a decoder. Further, the VAE inherits Bayesian automatic relevancy determination [8] to infer the correct number of mixed sources in a totally unsupervised way, contrasting existing techniques that tend to over-separate mixtures when the number of assumed sources are more than in reality. To this point, we show that a regular auto-encoder is not suitable for source separation. Our memory-efficient VAE involves a single encoding and decoding neural network and assumes a Laplace likelihood to improve the quality of separated sources. Qualitative and quantitative results on handwritten digit and audio spectrogram data exemplify our method's high quality in comparison to the state-of-the-art and ideal masks.

II. Overview of Previous Work

The BSS problem has attracted signal processing research for several decades [1]. Classic BSS algorithms include independent component analysis (ICA) [9] and robust principal component analysis (RPCA) [10]. Non-negative matrix factorization (NMF) is a powerful BSS framework that assumes non-negative data and decomposes it into activations and templates, which for spectrograms correspond to spectral templates and temporal activations [11], [12].

Machine learning methods learn prior information about sources by fitting (training) a model to example data. Deep neural networks have been successful at supervised source separation due to their capacity for pattern recognition. Supervised separation is now mature enough to be used for some real-world musical source separation applications [1], [3]. Recent

methods combine supervised deep learning and auditory scene analysis [13].

Latent variable models, like VAEs, have been applied to supervised [14] and semi-supervised source separation. A supervised VAE-NMF hybrid method was designed for multi-channel signal separation [15]. Semi-supervised training with source class labels in [16] performed decently compared to supervised training on source signals. Existing VAE methods use separate decoder or encoders for each source and require some supervision during training. Deep clustering [17] uses supervised learning with ideal binary masks to cluster (separate) latent variables that correspond to different source signals.

Unsupervised BSS has garnered significant interest in the last several years. In [18], a generative adversarial network (GAN) separated two images including particular combinations of handwritten digits, yet it was not successful at separating audio spectrograms. Mixtures of generative latent optimization (GLO) [19] models were applied to semi-supervised separation of two images and speech signals [20]. Mixture invariant training (MixIT) [21] is capable of unsupervised separation of speech signals in the time-domain. MixIT uses supervision to avoid over-separation and datasets containing 1 and 2 source mixtures (*i.e.* semi-supervision) to perform similarly to supervised neural network-based methods.

Unlike previous work, we present a completely unsupervised single-channel BSS method that automatically estimates the number of sources in the mixture thanks to Bayesian automatic relevancy determination and out-performs existing unsupervised methods at separating mixtures of images and audio spectrograms when the dataset contains only mixtures of multiple sources. Contrasting iterative methods like NMF, the proposed VAE benefits from fast, deterministic inference that can separate in real-time. The new method can handle difficult single-channel, general unsupervised separation of handwritten image digits from the same distribution, for example a pair of superimposed "3"s, as well as sounds generated from the same instrument, like two frequency modulated (vibrato) violins.

III. Methodology

This section presents a variational auto-encoder (VAE) [5] for amortized inference of latent source encodings from mixed data and independent generation of high-dimensional source signals with deep neural networks.

A. Problem statement

The source separation problem involves a D_x-dimensional vector $\mathbf{x}$ made from a sum of M sources $\mathbf{s}_m$ plus noise $\boldsymbol{\varepsilon}$,

$$\mathbf{x} = \sum_{m=1}^{M} \mathbf{s}_m + \boldsymbol{\varepsilon}\,. \tag{1}$$

Given $\mathbf{x}$, the goal is to infer (estimate) an assumed number K of source estimates $\widehat{\mathbf{s}}_k$. In unsupervised BSS, we do not have access to the true sources $\mathbf{s}_m$ to train the model, nor do we know the true number of underlying sources M. Indeed, for complete separation, $M \leqslant K$. Therefore, we assume that a mixture is comprised of at most K sources. If $M < K$, then $\mathbf{s}_k = \mathbf{0}$, for $M < k \leqslant K$. Prior assumptions are required to solve this highly under-determined problem.

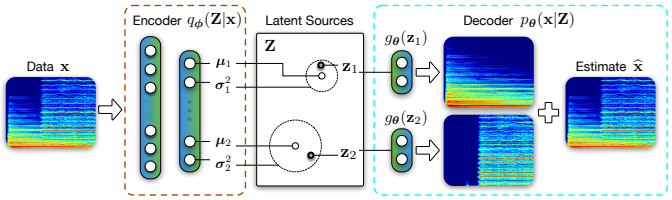

Fig. 1. *VAE for unsupervised blind source separation.* An encoder separates (disentangles) the data mixture into latent sources. Then, a decoder independently generates a signal from each latent source. The source signals are added together to provide an estimate of the data mixture.

Our data-driven approach learns prior information from a training dataset of mixtures, through the joint optimization of a deep generative model (decoder) and an inference model (encoder), as illustrated in Figure 1.

B. Generative model

Source signal $\mathbf{s}_k$ is generated according to a random process involving a D_z-dimensional continuous latent source variable $\mathbf{z}_k$, where $D_z \ll D_x$. Generator function $g_{\boldsymbol{\theta}}$ is a neural network with parameters $\boldsymbol{\theta}$ that individually decodes each latent source $\mathbf{z}_k$ into its higher-dimensional signal $\widehat{\mathbf{s}}_k$,

$$\widehat{\mathbf{s}}_k = g_{\boldsymbol{\theta}}(\mathbf{z}_k)\,. \tag{2}$$

Data noise $\boldsymbol{\varepsilon}$ is assumed to follow a zero-mean Laplace distribution with scale $b = \sqrt{0.5}$ for unit-variance [22]. Let $\mathbf{Z} = [\mathbf{z}_k]_{k=1}^K$ be the concatenation of the latent source variables with $D_Z = D_z K$ dimensions. Then, the likelihood of data $\mathbf{x}$ given $\mathbf{Z}$ is

$$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}) = \prod_{i=1}^{D_x} \text{Lap}\,(x_i|\widehat{x}_i, b) = \prod_{i=1}^{D_x} \frac{1}{2b} \exp\left(-\frac{|x_i - \widehat{x}_i|}{b}\right), \tag{3}$$

where the estimate of the mixed data is given by the sum of source signals,

$$\widehat{\mathbf{x}} = \sum_{k=1}^{K} \widehat{\mathbf{s}}_k = \sum_{k=1}^{K} g_{\boldsymbol{\theta}}(\mathbf{z}_k)\,. \tag{4}$$

Assuming a Gaussian likelihood (ℓ_2 loss) is common for VAEs but permits small deviations around the mean and leads to blurry reconstructions. Instead, the sharp peak of the Laplace likelihood (ℓ_1 loss) [23], [24] evenly penalizes deviations around the mean and affords precise reconstructions. To the best of our knowledge, this simple remedy to the established blurry VAE problem was absent from existing literature.

Isotropic Gaussian priors are defined over each source's latent variable,

$$p(\mathbf{Z}) = \prod_{k=1}^{K} p(\mathbf{z}_k) = \prod_{k=1}^{K} \mathcal{N}\,(\mathbf{z}_k|\mathbf{0}, \mathbf{I})\,. \tag{5}$$

This prior assumes that each element varies independently and helps to separate factors of variation in the data.

C. Inference model

Inferring $\mathbf{Z}$ from data $\mathbf{x}$ provides estimates of latent sources $\mathbf{z}_k$, $\forall k$, from which we can generate source signals $\widehat{\mathbf{s}}_k$, $\forall k$, with (2) and thus achieve source separation.

We use variational inference to approximate the posterior distribution over the latent variables given the data [25]. Approximate posterior $q_{\boldsymbol{\phi}}$ is mean-field factorized such that the elements of $\mathbf{Z}$ are independent and Gaussian distributed,

$$q_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{Z}|\mu_{\boldsymbol{\phi}}(\mathbf{x}), \sigma^2_{\boldsymbol{\phi}}(\mathbf{x})\mathbf{I}\right) . \tag{6}$$

An encoding neural network with parameters $\boldsymbol{\phi}$ outputs the mean $\boldsymbol{\mu} = \mu_{\boldsymbol{\phi}}(\mathbf{x})$ and variance $\boldsymbol{\sigma}^2 = \sigma^2_{\boldsymbol{\phi}}(\mathbf{x})$.

D. Variational lower bound

Variational inference turns approximate inference into an optimization problem [26], maximizing the variational lower bound on model evidence given by

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}) = \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(n)}) , \tag{7}$$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(n)}) = \left\langle \ln p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}|\mathbf{Z}) \right\rangle_{q_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{x}^{(n)})} - D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{x}^{(n)}) \| p(\mathbf{Z})\right) . \tag{8}$$

where dataset $\mathbf{X} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ consists of N *i.i.d.* samples.

The first term is the expected log-likelihood under the approximate posterior that minimizes the *reconstruction error*. The second term is the negative KLD between the approximate posterior and the prior that minimizes the difference between the two distributions. The KLD contributes a regularization that, along with the stochastic sampling of the latent space, is crucial as it promotes disentanglement (separation).

Assuming a Gaussian approximate posterior is common for variational auto-encoding as it enables simple Monte-Carlo estimation of the expected log-likelihood with the re-parametrization trick:

$$\mathbf{Z}^{(n)} \sim q_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{x}^{(n)}), \qquad \mathbf{Z}^{(n)} = \boldsymbol{\mu}^{(n)} + \boldsymbol{\sigma}^{(n)} \odot \boldsymbol{\epsilon} , \tag{9}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ denotes an element-wise product.

IV. Architecture

A. Neural networks

The encoder and decoder each consisted of five fully connected feed-forward neural network layers. Each linear layer was followed by a ReLU activation function and batch-normalization. The encoder's hidden units progressively decreased after each layer, starting with the high-dimensional, vectorized input to low-dimensional latent variable. A final layer output the D_Z-dimensional approximate posterior mean $\boldsymbol{\mu}$ and log-variance $\ln \boldsymbol{\sigma}^2$ of $\mathbf{Z}$. Table I shows the number of hidden units in each layer. The decoder's hidden units increased after each layer, starting with D_z dimensions for the sampled latent source $\mathbf{z}_k$ and progressing in reverse order. The last fully connected layer was followed by a sigmoid activation function to output a D_x-dimensional source signal $\widehat{\mathbf{s}}_k$ with non-negative values in the range 0-1. Crucially, the same decoder individually generated each source's latent vector $\mathbf{z}_k$ into its signal $\widehat{\mathbf{s}}_k$. Generating the source signals can be done efficiently with parallel processing. Source signals were summed to produce the expected value of the data mixture, $\widehat{\mathbf{x}}$.

TABLE I
Encoder input, hidden, and output layer units (dimensions).

Dataset	Input	Hidden Layers					Output
	D_x	L1	L2	L3	L4	L5	$2 \times D_Z$
MNIST	784	700	600	500	400	300	$2 \times 20K$
MUMS	32768	2560	2048	1536	1024	512	$2 \times 64K$

B. VAE mask (VAEM)

After training the VAE, an inferred source $\widehat{\mathbf{s}}_k$ can be converted to a mask-based source signal estimate $\breve{\mathbf{s}}_k$ (VAEM) that captures fine details from the data signal,

$$\breve{\mathbf{s}}_k = \widehat{\mathbf{s}}_k \odot (\mathbf{x} \oslash \widehat{\mathbf{x}}) , \tag{10}$$

where $\oslash$ denotes element-wise division. Unlike the VAE itself, VAEM constrains the sum of estimated sources to exactly match the data. We thus get the best of both worlds: automatic determination of relevant sources and sharp source estimates.

V. Evaluation

VAE, VAEM, baseline unsupervised BSS methods, and ideal masks were evaluated on mixtures of $M = 2$ true underlying sources. Evaluations were completed for $K = (2, 3, 4)$ assumed model sources to see whether source estimation quality degrades for $K \neq M$. Source code, audio examples, and additional results are available at *https://www.music.mcgill.ca/~julian/vae-bss*.

A. Datasets

Image and audio datasets were used to evaluate the proposed method. Individual source examples were mixed together randomly during each epoch of training. Due to this random mixing process, the number of unique combinations, and therefore the effective number of dataset examples, is given by the binomial coefficient $\binom{N}{M}$.

Handwritten Digits: The MNIST dataset [27] contains 60000 training and 10000 testing images of handwritten digits 0-9. Mixed image data was generated by adding $M = 2$ randomly sampled images and normalizing the result to the range 0-1. The training and testing images were never mixed with one-another.

Spectrograms: The McGill University master samples (MUMS) dataset [28] contains 6545 musical instrument recordings sampled at 44.1 kHz. Audio files were transformed into 2048-point complex-valued short-time Fourier transforms (STFT) using a Hann window and a hop length of 512 samples. STFTs were cropped to include only the first 256 frequency bins (up to 5.5 kHz) and the first 128 time frames (about 1.5 seconds of audio). Data was randomly organized into a collection of 5236 (80 %) training and 1309 (20 %) testing

datasets. After random mixing of $M = 2$ source spectra, the number of unique training examples is on the order of 10^7. Training and testing sets were never mixed with one-another. Mixed STFTs were transformed to magnitude spectrograms, normalized to the range 0-1, and used as the model's batched input data. This nonlinear transformation (absolute value of a complex number) means that the data mixture is not simply a sum of individual sources, as phase mis-match between the signals alters the resulting magnitude.

B. Training procedure

Models were trained on an NVIDIA V100 GPU using an ADAM optimizer [29] and a batch size of 128, where the learning rate was initially 1e-4 and decayed exponentially by 0.01% per epoch. To avoid posterior collapse early in training, the β-VAE formulation was used [30], [31] where the KLD (8) was multiplied by a factor β that increased linearly from 0 to 0.5 over the first 100 epochs. We found experimentally that using 0.5 improved the reconstruction error and reduced model over-pruning (again, improving quality), as it gave slightly more weight to the reconstruction error than the KLD. This was appropriate since we were primarily interested in inference and reconstruction, not sampling. Completing one epoch over the MNIST training set took 3 seconds (5e-5 seconds per example). Training was stopped once the negative variational lower bound converged, after about 5000 epochs.

C. Baseline methods and ideal masks

Baseline unsupervised BSS methods include NMF [32] with 8 templates per source, an auto-encoder (AE) that has the same architecture as VAE but with a deterministic latent variable and no KLD term, GLO [20], a *semi*-supervised method that has access to one of two clean target sources during training, and mixture invariant training (MixIT) [21], an unsupervised discriminative separation approach. Originally, MixIT was applied to waveform separation using a time-domain separation architecture. Here, we used MixIT with a separation architecture taken from the proposed method for image and spectrogram data, with the five encoding and decoding layers described in Sec. IV-A.

Ideal masks provide an upper-bound on performance. We compared these methods against ideal binary masks (IBM) and ideal ratio masks (IRM) as defined in [4].

D. Quantitative results

Image separation performance was evaluated according to the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [33]. For spectrogram data, source audio waveforms were synthesized from their estimated magnitude spectrograms using the inverse STFT and the data mixture's phase spectrum. Audio separation performance was quantified with bss_eval measures [34]: scale-invariant signal-to-distortion ratio (SI-SDR) [35], signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR).

Quantitative results in Tables II and III show the median values computed over the entire testing dataset. Our method outperformed the baseline unsupervised methods and semi-supervised GLO. The baseline methods over-separated the mixture when $K > M$, which can be seen by MixIT's lower performance for $K = 3$ in Table II. In contrast, VAE(M) automatically trimmed superfluous sources from the model for $K > M$, such that two of the K source signals contributed non-zero (relevant) values after training. Indeed, VAE(M)'s quality was consistent for $K \geqslant M$.

TABLE II
BENCHMARK PSNR AND SSIM RESULTS FROM THE UNSUPERVISED BSS TASK ON MIXED PAIRS OF MNIST HANDWRITTEN DIGITS.

		NMF	AE	GLO	MixIT	VAEM
$K = 2$	PSNR	17.22	15.96	24.63	25.69	**26.69**
	SSIM	0.50	0.43	0.86	0.88	**0.93**
$K = 3$	PSNR	18.27	16.30	n/a	25.41	**27.68**
	SSIM	0.44	0.43	n/a	0.69	**0.94**

TABLE III
BENCHMARK BSS_EVAL RESULTS FROM THE UNSUPERVISED BSS TASK ON MIXED PAIRS OF MUMS AUDIO SPECTROGRAMS.

	NMF	MixIT	VAE	VAEM	IBM	IRM
SI-SDR	5.40	6.64	14.33	**17.10**	23.97	22.66
SIR	16.93	17.59	**29.92**	29.55	48.89	34.39
SAR	7.96	8.26	14.87	**18.20**	24.06	23.23

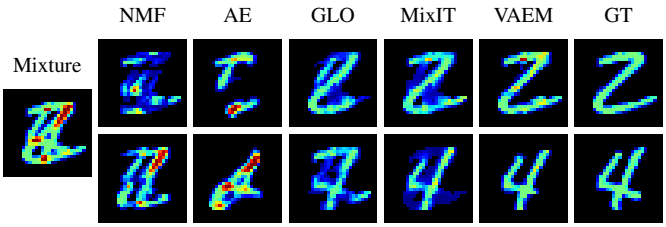

Fig. 2. *MNIST separation.* Example test data (Mixture) is composed of two unknown ground truth (GT) handwritten digit sources.

VAEM also outperformed existing methods at separating mixtures of spectrograms. We notice that the SI-SDR and SAR are improved with VAEM, while VAE has a better SIR than VAEM and the IBM. VAEM has a slightly lower SIR than VAE because the unit-sum condition from masking may re-introduce some energy from another source.

E. Qualitative results

Qualitative comparisons are presented in Figures 2 and 3. Figure 2 shows an ambiguous mixture of digits "2" and "4", where AE and NMF are poor, GLO and MixIT are good, and VAEM is excellent. Figure 3 shows a particularly challenging spectrogram example involving a low-pitched (A#1) bassoon sound mixed with a frequency modulated (E4) violin sound. We observe that VAE provides clean, smooth separated sources that are sharpened with masking.

VI. CONCLUSION

Our variational auto-encoding approach offers a solution to unsupervised blind source separation. Disentangling sources

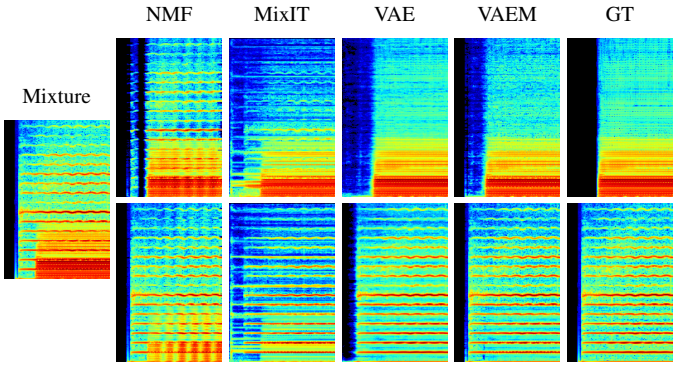

Fig. 3. *Spectrogram separation.* Example test data (Mixture) is composed of unknown sound sources (GT): bassoon (top) and violin (bottom).

in a low-dimensional probabilistic latent space is practically effective and aligns with intuition about how humans perceive separate sounds in an acoustic scene [2]. Generating sources independently using the same decoding deep neural network means that the size, and therefore memory consumption, of the network remains consistent regardless of the number of model sources, which is an advantage over existing methods that use different networks for each source. While a simplified version of the model with a few hidden layers and Gaussian assumptions is sufficient for baseline separation, the more precise model we developed yields improved qualitative and quantitative results. Extending the presented framework to blindly separate universal audio waveforms and videos of arbitrary durations is left to future work.

REFERENCES

[1] E. Vincent, T. Virtanen, and S. Gannot, eds., *Audio Source Separation and Speech Enhancement*. John Wiley and Sons, 2018.

[2] A. Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[3] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and state-of-the art music source separation tool with pre-trained models." ISMIR Late-Breaking/Demo, Nov. 2019. Deezer Research.

[4] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*, pp. 293–305, Springer Int. Publishing, 2018.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Int. Conf. Learning Representations (ICLR)*, 2014.

[6] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos, "Universal audio synthesizer control with normalizing flows," in *Proc. 22nd Int. Conf. Digital Audio Effects (DAFx-19)*, Sep. 2019.

[7] Y.-J. Luo, K. Agres, and D. Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders," in *Proc. 20th International Music Information Retrieval (ISMIR) Conf.*, pp. 746–753, Nov. 2019.

[8] J. Neri, R. Badeau, and P. Depalle, "Probabilistic filter and smoother for variational inference of Bayesian linear dynamical systems," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5885–5889, May 2020.

[9] J. Cardoso, "Blind signal separation: statistical principles," *Proc. of the IEEE*, vol. 86, pp. 2009–2025, Oct. 1998.

[10] P. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 57–60, 2012.

[11] C. Fevotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[12] R. Badeau and A. Drémeau, "Variational Bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6171–6175, May 2013.

[13] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, Dec 2019.

[14] L. Pandey, A. Kumar, and V. Namboodiri, "Monaural audio source separation using variational autoencoders," in *Interspeech*, pp. 3489–3493, 2018.

[15] S. Seki, H. Kameoka, T. Toda, and K. Takeda, "Underdetermined source separation based on generalized multichannel variational autoencoder," in *IEEE Access*, vol. 7, Nov. 2019.

[16] E. Karamatli, A. Cemgil, and S. Kirbiz, "Audio source separation using variational autoencoders and weak class supervision," *IEEE Signal Processing Letters*, vol. 26, pp. 1349–1353, Sep 2019.

[17] J. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 31–35, 2016.

[18] Y. Hoshen, "Towards unsupervised single-channel blind source separation using adversarial pair unmix-and-remix," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019.

[19] P. Bojanowski, A. Joulin, D. Lopez-Pas, and A. Szlam, "Optimizing the latent space of generative networks," in *Proc. Machine Learning Research*, vol. 80, pp. 600–609, 2018.

[20] T. Halperin, A. Ephrat, and Y. Hoshen, "Neural separation of observed and unobserved distributions," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, 2019.

[21] S. Wisdom, E. Tzinis, H. Erdogan, J. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, 2020.

[22] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*. John Wiley & Sons, Inc., 4th ed., 2011.

[23] J. Neri, P. Depalle, and R. Badeau, "Laplace state space filter with exact inference and moment matching," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5880–5884, May 2020.

[24] J. Neri, P. Depalle, and R. Badeau, "Approximate inference and learning of state space models with Laplace noise," *IEEE Transactions on Signal Processing*, doi: 10.1109/TSP.2021.3075146, April 2021.

[25] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "Introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 1999.

[26] D. Blei, A. Kucukelbir, and J. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[27] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[28] F. Opolko and J. Wapnick, "McGill university master samples (MUMS)," *Faculty of Music, McGill University, Montreal, Canada*, 1989.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learning Representations (ICLR)*, 2015.

[30] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Int. Conf. Learning Representations (ICLR)*, 2016.

[31] C. Sonderby, T. Raiko, L. Maaloe, S. Sonderby, and O. Winther, "How to train deep variational autoencoders and probabilistic ladder networks," in *Proc. 33rd Int. Conf. Machine Learning (ICML)*, 2016.

[32] P. López-Serrano, C. Dittmar, and Y. Özer, "NMF toolbox: Music processing applications of nonnegative matrix factorization," in *Proc. 22nd Int. Conf. Digital Audio Effects (DAFx-19)*, pp. 1–8, Sep 2019.

[33] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *20th Int. Conf. Pattern Recognition*, pp. 2366–2369, 2010.

[34] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[35] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 626–630, 2019.