

Maurice Frank

Unsupervised variational source separation with deep pri-  
ors



# Contents

<i>Proposal</i>	7
<i>Abstract</i>	7
<i>Research Question</i>	7
<i>Related works</i>	7
<i>Methodology</i>	11
<i>Planning</i>	11



## *Todo list*

add other supervision: knowing $k$ . . . . .	7
explain free energy . . . . .	9



# Proposal

## Abstract

### Research Question

Source separation is the task of finding a set of latent sources  $\mathbf{s} = [s_1, \dots, s_k, \dots, s_n]$  to an observed mix of those sources  $\mathbf{m}$ . The induced model proposes a mixing function  $\mathbf{m} = f(\mathbf{s})$ . The task is to find an approximate inverse model  $g(\cdot)$  which retrieves  $\mathbf{s}$ :

$$\mathbf{m} = f(\mathbf{s}) \quad (1)$$

$$g(\mathbf{m}) \cong \mathbf{s} \quad (2)$$

In this learning setting *supervision* can happen in two ways: First the source signals are identified as being from class  $k$ <sup>1</sup>. Second the tuples  $(\mathbf{m}, \mathbf{s})$  are supervised giving us examples of mixes and their corresponding sources.

<sup>1</sup> For the setting of music think of the classes being  $\{\text{guitar, piano, voice, } \dots\}$

1. Can we learn an source separation model  $g(\cdot)$  by learning deep priors for the different source classes.
2. Can we reduce this to an unsupervised setting. Unsupervised relating to the missing pairings of sources and mixes.

add other supervision: knowing  $k$

## Related works

In this chapter we discuss previous research in supervised and semi-supervised source separation.

### ICA

### Deep Latent-Variable Models

For our process we have observations from the data space  $\mathbf{x} \in \mathcal{D}$  for which there exists an unknown data probability distribution  $p^*(\mathcal{D})$ .

We collect a data set  $\{x_1 \dots x_N\}$  with  $N$  samples. We introduce an approximate model with density<sup>2</sup>  $p_\theta(\mathcal{D})$  and model parameters  $\theta$ . Learning or modelling means finding the values for  $\theta$  which will give the closest approximation of the true underlying process:

$$p_\theta(\mathcal{D}) \approx p^*(\mathcal{D}) \quad (3)$$

The model  $p_\theta$  has to be complex enough to be able to fit the data density while little enough parameters to be learnable. Every choice for the form of the model will *induce* biases<sup>3</sup> about what density we can model, even before we maximize a learning objective using the parameters  $\theta$ .

In the following described models we assume the sampled data points  $x$  to be drawn from  $\mathcal{D}$  *independent and identically distributed*<sup>4</sup>. Therefore we can write the data log-likelihood as:

$$p_\theta(\mathcal{D}) = \prod_{x \in \mathcal{D}} p_\theta(x) \quad (4)$$

$$\log p_\theta(\mathcal{D}) = \sum_{x \in \mathcal{D}} \log p_\theta(x) \quad (5)$$

The maximum likelihood estimation of our model parameters maximizes this objective.

To form a latent-variable model we introduce a *latent variable*<sup>5</sup>. The data likelihood now is the marginal density of the joint latent density:

$$p_\theta(x) = \int p_\theta(x, z) dz \quad (6)$$

Typically we introduce a factorization of the joint. Most commonly and simplest:

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz \quad (7)$$

This corresponds to the graphical model in which  $z$  is generative parent node of the observed  $x$ , see Figure 1. The density  $p(z)$  is called the *prior distribution*.

If the latent is small, discrete, it might be possible to directly marginalize over it. If for example  $z$  is a discrete random variable and the conditional  $p_\theta(x|z)$  is a Gaussian distribution than the data model density  $p_\theta(x)$  becomes a mixture-of-Gaussians, which we can directly estimate by maximum likelihood estimation of the data likelihood.

For more complicated models the data likelihood  $p_\theta(x)$  as well as the model posterior  $p_\theta(z|x)$  are intractable because of the integration over the latent  $z$  in Equation (7).

To formalize the search for an intractable posterior into a tractable optimization problem we follow the *variational principle*<sup>6</sup> which intro-

<sup>2</sup> We write density and distribution interchangeably to denote a probability function.

<sup>3</sup> called *inductive biases*

<sup>4</sup> meaning the sample of one datum does not depend on the other data points

<sup>5</sup> Latent variables are part of the directed graphical model but not observed.

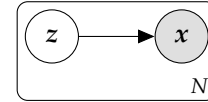


Figure 1: The graphical model with a introduced latent variable  $z$ . Observed variables are shaded.

<sup>6</sup> Michael I. Jordan et al. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233.



duces an approximate posterior distribution  $q_\phi(z|x)$ , also called the *inference model*. Again the choice of model here carries inductive biases as such that even in asymptotic expectation we can not obtain the true posterior.

Following the derivation in<sup>7</sup> we introduce the inference model into the data likelihood<sup>8</sup>:

$$\log p_\theta(x) = \mathbb{E}_{q_\theta(z|x)} [\log p_\theta(x)] \quad (8)$$

$$= \mathbb{E}_{q_\theta(z|x)} \left[ \log \frac{p_\theta(x, z)}{p_\theta(z|x)} \right] \quad (9)$$

$$= \mathbb{E}_{q_\theta(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \quad (10)$$

$$= \mathbb{E}_{q_\theta(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] + \mathbb{E}_{q_\theta(z|x)} \left[ \log \frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \quad (11)$$

$$= \mathbb{E}_{q_\theta(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] + \mathbb{D}_{\text{KL}}[q_\phi(z|x) \| p_\theta(z|x)] \quad (12)$$

Note that we separated the likelihood into two parts. The second part is the (positive) Kullback-Leibler divergence of the approximate posterior from the true intractable posterior. This unknown divergence states the ‘correctness’ of our approximation<sup>9</sup>.

The first term is the *variational free energy* or *evidence lower bound* (ELBO):

$$\text{ELBO}_{\theta, \phi}(x) = \mathbb{E}_{q_\theta(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] \quad (13)$$

We can introduce the same factorization as in Equation (7):

$$\text{ELBO}_{\theta, \phi}(x) = \mathbb{E}_{q_\theta(z|x)} \left[ \log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \right] \quad (14)$$

$$= \mathbb{E}_{q_\theta(z|x)} \left[ \log \frac{p(z)}{q_\phi(z|x)} \right] + \mathbb{E}_{q_\theta(z|x)} [\log p_\theta(x|z)] \quad (15)$$

$$= -\mathbb{D}_{\text{KL}}[q_\phi(z|x) \| p(z)] + \mathbb{E}_{q_\theta(z|x)} [\log p_\theta(x|z)] \quad (16)$$

Under this factorization we separated the lower bound into two parts. First the divergence of the approximate posterior from the latent prior distribution and second the data posterior likelihood from the latent<sup>10</sup>.

The optimization of the  $\text{ELBO}_{\theta, \phi}$  allows us to jointly optimize the parameter sets  $\theta$  and  $\phi$ . The gradient with respect to  $\theta$  can be estimated with an unbiased Monte Carlo estimate using data samples<sup>11</sup>. We can *not* though do the same for the variational parameters  $\phi$ , as the expectation of the ELBO is over the approximate posterior which

<sup>7</sup> Diederik P. Kingma and Max Welling. “An Introduction to Variational Autoencoders”. In: (2019). arXiv: [1906.02691](#) [cs, stat], p. 20.

<sup>8</sup> The first step is valid as  $q_\theta$  is a valid density function and thus integrates to one.

<sup>9</sup> More specifically the divergence marries two errors of our approximate model. First it gives the error of our posterior estimation from the true posterior, by definition of divergence. Second it specifies the error of our complete model likelihood from the marginal likelihood. This is called the *tightness* of the bound.

explain free energy

<sup>10</sup> this will later be the reconstruction error. How well can we return to the data density from latent space

<sup>11</sup>  $\nabla_\theta \text{ELBO}_{\theta, \phi} \cong \nabla_\theta \log p_\theta(x, z)$

depends on  $\phi$ . By a change of variable of the latent variable we can make this gradient tractable, the so called *reparameterization trick*.<sup>12</sup> We express the  $z \sim q_\theta$  as a random sample from a unparametrized source of entropy  $\epsilon$  and a parametrized transformation:

$$z = f_\eta(\epsilon) \quad (17)$$

For example for a Gaussian distribution we can express  $z \sim \mathcal{N}(\mu, \sigma)$  as  $z = \mu + \sigma \cdot \epsilon$  with  $\epsilon \sim \mathcal{N}(0, 1)$  and  $\eta = \{\mu, \sigma\}$ .

### The VAE framework

VAE<sup>12,13</sup>

The  $\beta$ -VAE<sup>14</sup> extends the VAE objective with an  $\beta$  hyperparameter in front of the KL divergence. The value  $\beta$  gives a constraint on the latent space controlling the capacity of it. Adapting  $\beta$  gives a trade-off between reconstruction quality of the autoencoder and the simplicity of the latent representations<sup>14</sup>. Using such a constraint is similar to the use of in the information bottleneck.<sup>15</sup>

### Flow based models

Another class of common deep latent models are based on *normalizing flows*.<sup>16</sup> A normalizing flow is a function  $f(x)$  that maps the input density to a fixed, prescribed density  $p(\epsilon) = p(f(x))$ , in that normalizing the density<sup>17</sup>. They use a flow for the approximate posterior  $q_\phi(z|x)$ . Again this is commonly set to be a factorized Gaussian distribution.

For a finite normalizing flow we consider a chain of invertible, smooth mappings.

NICE<sup>18</sup> - volume preserving transformations - coupling layer - triangular shape

Normalizing Flow<sup>19</sup>

RealNVP<sup>20</sup> - non-volume preserving

Glow<sup>21</sup> - invertible 1x1 convs - ActNorm - zero init

<sup>22</sup> introduced WaveNet an autoregressive generative model for raw (*time-domain*) audio. WaveNet closely similar to the earlier PixelCNN<sup>23</sup> but adapted for the audio domain. Unmodified Cnns are unsuitable to the application to raw audio because of the form of data. as digital audio is sampled at a extremely high sample rate commonly 16kHz up to 44kHz the features of interest lie at scale of stringly different magnitudes. On the one hand recognizing phase, frequency of a wave might require features at those ms scales on the other hand the modelling of speech or music audio happens at the scale of seconds or minutes. As such a generative model for this domain has to capture those different

<sup>12</sup> Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (2014). arXiv: [1312.6114 \[cs, stat\]](#).

<sup>13</sup> Danilo Jimenez Rezende et al. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: (2014). arXiv: [1401.4082 \[cs, stat\]](#).

<sup>14</sup> Irina Higgins et al. "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: (2016).

<sup>15</sup> Christopher P. Burgess et al. "Understanding Disentangling in Beta-VAE". In: (2018). arXiv: [1804.03599 \[cs, stat\]](#).

<sup>16</sup> Esteban Tabak and Cristina V. Turner. "A family of nonparametric density estimation algorithms". In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 145–164.

<sup>17</sup> The extreme of this idea is, of course, an infinitesimal, continuous-time flow with a velocity field.

<sup>18</sup> Laurent Dinh et al. "NICE: Non-Linear Independent Components Estimation". In: (2015). arXiv: [1410.8516 \[cs\]](#).

<sup>19</sup> Danilo Jimenez Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: (2016). arXiv: [1505.05770 \[cs, stat\]](#).

<sup>20</sup> Laurent Dinh et al. "Density Estimation Using Real NVP". In: (2017). arXiv: [1605.08803 \[cs, stat\]](#).

<sup>21</sup> Diederik P. Kingma and Prafulla Dhariwal. "Glow: Generative Flow with Invertible 1x1 Convolutions". In: (2018). arXiv: [1807.03039 \[cs, stat\]](#).

<sup>22</sup> Aaron van den Oord et al. "WaveNet: A Generative Model for Raw Audio". In: (2016). arXiv: [1609.03499 \[cs\]](#).

<sup>23</sup> Aaron van den Oord et al. "Conditional Image Generation with PixelCNN Decoders". In: (2016). arXiv: [1606.05328 \[cs\]](#).

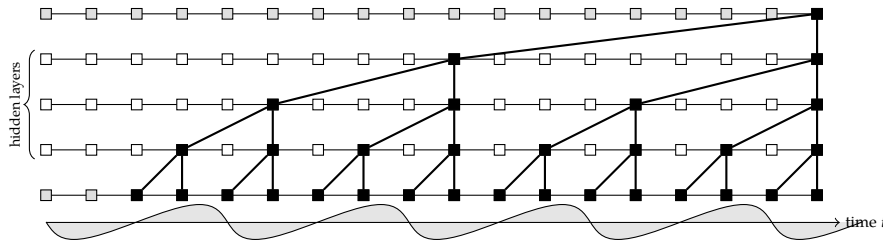


Figure 2: Dilated convolutions

time scales. The wvaenet accomplishes this by using dilated convolutions a common tool in signal processing.<sup>24</sup> A dilated convolutions uses a kernel with an inner stride. Using a stack of dialted convolutions increases the recpetive field of the features without increasing the comutional complexity.

- gated convs -pixelcnn -lstm<sup>25</sup> - dilated convs - global conditioning  
 -  $\mu$ -law encoding<sup>26</sup> - slow cause autoreg (better with<sup>27</sup>) -  
 PixelCNN++<sup>28</sup>

## Sound

NSynth<sup>29</sup>

In<sup>30</sup>

FloWaveNet<sup>31</sup>

## Source separation

WaveNet for Speech denoising<sup>32</sup>

WaveNet-VAE unsupervised speech rep learning<sup>33</sup>

<sup>34</sup> were the first to use an U-Net architecture for source separation.

Wave-U-Net<sup>35</sup>

DeMucs<sup>36</sup>

Source Sep in Time Domain<sup>37</sup>

## Methodology

### Datasets

### ToyData

### MusDB

### Planning

<sup>24</sup> P. Dutilleux. "An Implementation of the Algorithm à Trous to Compute the Wavelet Transform". In: *Wavelets*. Ed. by Jean-Michel Combes et al. Inverse Problems and Theoretical Imaging. Berlin, Heidelberg: Springer, 1990, pp. 298–304.

<sup>25</sup> Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

<sup>26</sup> Recommendation G. 711. Pulse Code Modulation (PCM) of Voice Frequencies. 1988.

<sup>27</sup> Tom Le Paine et al. "Fast Wavenet Generation Algorithm". In: (2016). arXiv: [1611.09482 \[cs\]](#).

<sup>28</sup> Tim Salimans et al. "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications". In: (2017). arXiv: [1701.05517 \[cs, stat\]](#).

<sup>29</sup> Nal Kalchbrenner et al. "Efficient Neural Audio Synthesis". In: (2018). arXiv: [1802.08435 \[cs, eess\]](#).

<sup>30</sup> Ryan Prenger et al. "WaveGlow: A Flow-Based Generative Network for Speech Synthesis". In: (2018). arXiv: [1811.00002 \[cs, eess, stat\]](#).

<sup>31</sup> Sungwon Kim et al. "FloWaveNet : A Generative Flow for Raw Audio". In: (2019). arXiv: [1811.02155 \[cs, eess\]](#).

<sup>32</sup> Dario Rethage et al. "A Wavenet for Speech Denoising". In: (2018). arXiv: [1706.07162 \[cs\]](#).

<sup>33</sup> Jan Chorowski et al. "Unsupervised Speech Representation Learning Using WaveNet Autoencoders". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12 (2019), pp. 2041–2053. arXiv: [1901.08810](#).

<sup>34</sup> Andreas Jansson et al. "Singing Voice Separation with Deep U-Net Convolutional Networks". In: *ISMIR*. 2017.

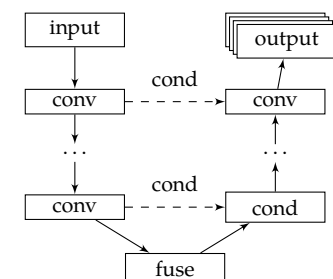


Figure 3: The U-Net



# Bibliography

- Burgess, Christopher P., Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. "Understanding Disentangling in Beta-VAE". In: (2018). arXiv: [1804.03599 \[cs, stat\]](#) (cit. on p. 10).
- Chorowski, Jan, Ron J. Weiss, Samy Bengio, and Aäron van den Oord. "Unsupervised Speech Representation Learning Using WaveNet Autoencoders". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12 (2019), pp. 2041–2053. arXiv: [1901.08810](#) (cit. on p. 11).
- Défossez, Alexandre, Nicolas Usunier, Léon Bottou, and Francis Bach. "Demucs: Deep Extractor for Music Sources with Extra Unlabeled Data Remixed". In: (2019). arXiv: [1909.01174 \[cs, eess, stat\]](#) (cit. on p. 11).
- Dinh, Laurent, David Krueger, and Yoshua Bengio. "NICE: Non-Linear Independent Components Estimation". In: (2015). arXiv: [1410.8516 \[cs\]](#) (cit. on p. 10).
- Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. "Density Estimation Using Real NVP". In: (2017). arXiv: [1605.08803 \[cs, stat\]](#) (cit. on p. 10).
- Dutilleul, P. "An Implementation of the Algorithme à Trous to Compute the Wavelet Transform". In: *Wavelets*. Ed. by Jean-Michel Combes, Alexander Grossmann, and Philippe Tchamitchian. Inverse Problems and Theoretical Imaging. Berlin, Heidelberg: Springer, 1990, pp. 298–304 (cit. on p. 11).
- Higgins, Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: (2016) (cit. on p. 10).
- Hochreiter, Sepp and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 11).
- Jansson, Andreas, Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Aparna Kumar, and Tillman Weyde. "Singing Voice Separation with Deep U-Net Convolutional Networks". In: *ISMIR*. 2017.
- Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233 (cit. on p. 8).
- Kalchbrenner, Nal, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. "Efficient Neural Audio Synthesis". In: (2018). arXiv: [1802.08435 \[cs, eess\]](#) (cit. on p. 11).
- Kim, Sungwon, Sang-gil Lee, Jongyoon Song, Jaehyeon Kim, and Sungroh Yoon. "FloWaveNet: A Generative Flow for Raw Audio". In: (2019). arXiv: [1811.02155 \[cs, eess\]](#) (cit. on p. 11).

- Kingma, Diederik P. and Prafulla Dhariwal. "Glow: Generative Flow with Invertible 1x1 Convolutions". In: (2018). arXiv: [1807.03039 \[cs, stat\]](#) (cit. on p. 10).
- Kingma, Diederik P. and Max Welling. "Auto-Encoding Variational Bayes". In: (2014). arXiv: [1312.6114 \[cs, stat\]](#) (cit. on p. 10).
- Kingma, Diederik P. and Max Welling. "An Introduction to Variational Autoencoders". In: (2019). arXiv: [1906.02691 \[cs, stat\]](#) (cit. on p. 9).
- Lluís, Francesc, Jordi Pons, and Xavier Serra. "End-to-End Music Source Separation: Is It Possible in the Waveform Domain?" In: (2019). arXiv: [1810.12187 \[cs, eess\]](#) (cit. on p. 11).
- Paine, Tom Le, Pooya Khorrami, Shiyu Chang, Yang Zhang, Prajit Ramachandran, Mark A. Hasegawa-Johnson, and Thomas S. Huang. "Fast Wavenet Generation Algorithm". In: (2016). arXiv: [1611.09482 \[cs\]](#) (cit. on p. 11).
- Prenger, Ryan, Rafael Valle, and Bryan Catanzaro. "WaveGlow: A Flow-Based Generative Network for Speech Synthesis". In: (2018). arXiv: [1811.00002 \[cs, eess, stat\]](#) (cit. on p. 11).
- Recommendation G. 711. *Pulse Code Modulation (PCM) of Voice Frequencies*. 1988 (cit. on p. 11).
- Rethage, Dario, Jordi Pons, and Xavier Serra. "A Wavenet for Speech Denoising". In: (2018). arXiv: [1706.07162 \[cs\]](#) (cit. on p. 11).
- Rezende, Danilo Jimenez and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: (2016). arXiv: [1505.05770 \[cs, stat\]](#) (cit. on p. 10).
- Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: (2014). arXiv: [1401.4082 \[cs, stat\]](#) (cit. on p. 10).
- Salimans, Tim, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications". In: (2017). arXiv: [1701.05517 \[cs, stat\]](#) (cit. on p. 11).
- Stoller, Daniel, Sebastian Ewert, and Simon Dixon. "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation" (Paris, France). 2018 (cit. on p. 11).
- Tabak, Esteban and Cristina V. Turner. "A family of nonparametric density estimation algorithms". In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 145–164 (cit. on p. 10).
- Van den Oord, Aaron, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: (2016). arXiv: [1609.03499 \[cs\]](#) (cit. on p. 10).
- Van den Oord, Aaron, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. "Conditional Image Generation with PixelCNN Decoders". In: (2016). arXiv: [1606.05328 \[cs\]](#) (cit. on p. 10).