# MLASS: Multi Latent Autoregressive Source Separation

September 3, 2024

**Domiziano Scarcelli**

## Abstract

Building upon the foundation established by LASS (Latent Autoregressive Source Separation) (Postolache et al., 2023), this project aims to extend its capabilities to allow the separation of three or more sources. LASS demonstrated an innovative method for separating mixed sources into their original components without the need for additional gradient-based optimization or modifications to pre-existing models. However, due to the high spatial complexity, the original work was limited to mixtures of two sources. In this project, we introduce two different techniques to enable effective separation of more complex mixtures involving three or more sources. One technique is based on message passing via belief propagation, while in the other one we use a probabilistic extractor to extract each source independently. Both of these techniques avoid using the joint likelihood, which is the cause of the high space complexity. We validate our extended method on both image (MNIST) and audio (SLAKH) datasets.

## 1. Introduction

Autoregressive models have demonstrated remarkable performance across a wide array of domains, from natural language processing to densely-valued fields like audio and vision. A crucial element in their success is the use of vector-quantized latent spaces, which enable dimensionality reduction and faster inference times. The LASS (Latent Autoregressive Source Separation) (Postolache et al., 2023) framework leverages these autoregressive models and quantized latent spaces to perform source separation without requiring additional gradient-based optimization or modifications to existing models.

LASS employs a Bayesian approach where autoregressive models are used to compute priors and a discrete, non-parametric likelihood function is constructed through frequency counts over latent sums of addend tokens. This method has shown competitive results in separating mixed sources into their original components, particularly in image and audio domains. However, LASS is currently limited to handling mixtures of only two sources.

In this work, we extend the LASS framework to handle mixtures of three or more sources, introducing Multi Latent Autoregressive Source Separation (MLASS). Our approach maintains the original benefits of LASS, including significant speedups in inference time and scalability to higher-dimensional data, while improving its capability to separate multiple sources from a single mixture.

## 2. Related Work

The problem of source separation has tackled in the literature by the usage of *generative* (Jayaram & Thickstun, 2020) or *regression* (Halperin et al., 2019) based methods. Some methods make the use of multimodality, such as language, in order to make the separation possible (Liu et al., 2022).

While multiple source separation has been heavily discussed in the audio domain (Stoller et al., 2018), (Takahashi & Mitsufuji, 2017), the common setup for image separation consists in separating two sources. While these methods provide good results, they often make use of ad-hoc models that require to be trained from scratch.

This paper aims to extend the LASS method to multiple sources for both images and audio, without requiring additional gradient-based optimization or modifications of existing models.

## 3. Method

Let $z^i$ be the latent vector of the $i$-th source, and $m$ the latent vector of the mixture, obtained by passing the source or mixture into the VQVAE (Van Den Oord et al., 2017); We will use the subscript $z_t^i$ to denote the $t$-th value of the $z^i$ latent vector. Let $\mathcal{L}(m|z^1, \dots, z^n)$ be the likelihood, which models the probability of having the mix-

Email: Domiziano Scarcelli <scarcelli.1872664@studenti.uniroma1.it>.

*Deep Learning and Applied AI 2021, Sapienza University of Rome*, 2nd semester a.y. 2020/2021.

ture latent vector, given the $n$ sources' latent vectors; Let $P(z_t^i|z_{s<t}^i)$ be the autoregressive prior of the $i$-th source; Let $P(z^1, z^2, \dots, z^n|m)$ be the joint posterior distribution. In the LASS approach, the sources are sampled from the joint posterior:

$$\underbrace{P(z^1, z^2, \dots, z^n|m)}_{\text{Joint posterior}} \propto \underbrace{\mathcal{L}(m|z^1, z^2, \dots, z^n)}_{\text{Likelihood}} \underbrace{\prod_{i=1}^{n} P(z_t^i|z_{s<t}^i)}_{\text{Joint prior}}$$

Since the joint likelihood $\mathcal{L}$ for $n$ sources is modeled as a tensor $\mathcal{L} \in \mathbb{R}^{K^{n+1}}$, its spatial complexity is $O(K^{n+1})$, where $K$ is the number of discrete latent codes of the VQ-VAE. Even if represented using sparse notation, using a small $K = 256$ and $n = 3$ sources, the space complexity is still too high, resulting into the impossibility of representing the likelihood of more than 2 sources, hence not being able to perform source separation.

### 3.1. Belief Propagation

The first technique utilizes message passing via belief propagation to compute the marginal distributions of the latent variables. Let $G$ be a graphical model, and let $m^i$ be the latent vector of the mixture between the sources $z^1, z^2, \dots, z^i$. Let us define $\mu_\alpha$ and $\mu_\beta$ the forward and backward messages passed between the nodes in the graphical model:

$$\mu_\alpha(m^i, z^i) = \sum_{m^{i-1}} p(m^i|m^{i-1}, z^i) \sum_{z^{i-1}} \mu_\alpha(m^{i-1}, z^{i-1})$$

$$\mu_\beta(m^i) = \sum_{z^{i+1}} p(z^{i+1}) \sum_{m^{i+1}} p(m^{i+1}|m^i, z^{i+1}) \mu_\beta(m^{i+1})$$

Then it's possible to compute the sources' marginals using those messages:

$$p(z^i|m^n) = p(z^i) \sum_{m^i} \mu_\alpha(m^i, z^i) \mu_\beta(m^i)$$

Sampling from $p(z^i|m^n)$ for each of the sources $z^i$ gives the solution, and it requires $O((n-1)K^3)$ memory for $n$ sources, due to $p(m^i|m^{i-1}, z^i)$, modeled with a $F \in \mathbb{R}^{(n-1)\times K\times K\times K}$ tensor, that is once again computed with frequency count. This is a major improvement over the $O(K^{n+1})$ memory requirement of LASS.

### 3.2. Probabilistic Extractor

The other technique involves sampling the sources independently in order to avoid the use of the joint likelihood $\mathcal{L}$. The *marginal likelihood* of $z^i$ can be computed summing over all the sources, $z^j$ where $j \neq i$.

$$\mathcal{L}(m|z^i) = \textstyle\sum_{j\neq i} \mathcal{L}(m|z^1, z^2, \dots, z^i, \dots, z^n)$$

Instead of computing the *joint likelihood* with frequency counts over latent sums of addend token, we compute the *joint probability* $p(m, z^i)$ for each source $z^i$. We do that by initializing a tensor $F \in \mathbb{R}^{n\times K\times K}$, and while iterating over a dataset, we increment by 1 each coordinate $F[i, z^i, m]$. The joint probability is used to compute the marginal likelihood

$$\mathcal{L}(m|z^i) \propto \frac{P(m, z^i)}{P(z^i)}$$

which is used along with the autoregressive priors to compute the *marginal posterior* for each source $z^i$:

$$\underbrace{P(z^i|m)}_{\text{Marginal posterior}} \propto \underbrace{\mathcal{L}(z_m|z^i)}_{\text{Marginal likelihood}} + \underbrace{P(z_t^i|z_{s<t}^i)}_{\text{Prior}}$$

The solution is given by sampling each source from the marginal posterior.

## 4. Results

We validate our methods on both image (MNIST) and audio (SLAKH) datasets.

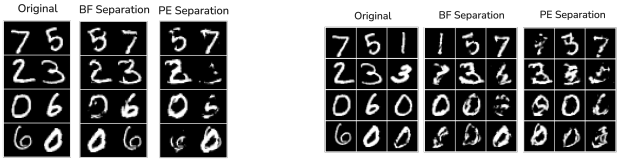


*Figure 1.* Examples of the MNIST separation with two sources (left) and three sources (right)

*Table 1.* Results on MNIST dataset (PSNR)

| Method | 2 sources | 3 sources |
|---|---|---|
| LASS | $24.23 \pm 6.23$ | N/A |
| MLASS-PE | $16.87 \pm 3.77$ | $13.64 \pm 1.76$ |
| MLASS-BP | $19.30 \pm 5.68$ | $14.19 \pm 2.23$ |

*Table 2.* Results on SLAKH dataset (SDR)

| Method | 2 sources | 3 sources |
|---|---|---|
| LASS | $5.01 \pm 2.39$ | N/A |
| MLASS-PE | $3.09 \pm 3.23$ | $-0.44 \pm 2.96$ |

The results are satisfactory regarding the separation on the MNIST set, but the quality of the separation heavily degrades for more complex tasks like audio separation on the SLAKH dataset.

## 5. Conclusions

MLASS extends the LASS framework to allow the separation of two or more sources, decoupling the spatial complexity required for the separation to the number of separated sources, maintaining the balance between separation quality and computational efficiency. In the paper, the two approaches *belief propagation* and *probabilistic extractor* have been compared, both of them addressing the limitations of the original LASS framework.

The code and other resources are available in the GitHub repository.[1]

## References

Halperin, T., Ephrat, A., and Hoshen, Y. Neural separation of observed and unobserved distributions. In *International Conference on Machine Learning*, pp. 2566–2575. PMLR, 2019.

Jayaram, V. and Thickstun, J. Source separation with deep generative priors. In *International Conference on Machine Learning*, pp. 4724–4735. PMLR, 2020.

Liu, X., Liu, H., Kong, Q., Mei, X., Zhao, J., Huang, Q., Plumbley, M. D., and Wang, W. Separate what you describe: Language-queried audio source separation, 2022. URL https://arxiv.org/abs/2203.15147.

Postolache, Mariani, Mancusi, Santilli, Cosmo, and Rodolà. Latent autoregressive source separation. *Proceedings of the AAAI Conference on Artificial Intelligence 37*, 8, 2023.

Stoller, D., Ewert, S., and Dixon, S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation, 2018. URL https://arxiv.org/abs/1806.03185.

Takahashi, N. and Mitsufuji, Y. Multi-scale multi-band densenets for audio source separation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 21–25, 2017. doi: 10.1109/WASPAA.2017.8169987.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[1] https://github.com/DomizianoScarcelli/multi-latent-autoregressive-source-separation