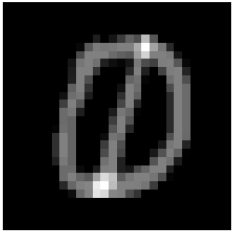
Semi-Supervised Source Separation by Learning What to Not Learn

NNPM – 18th August 2023

Martin Ludvigsen

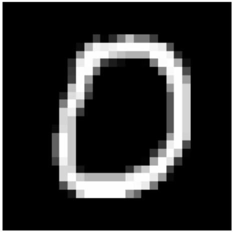
Department of Mathematical Sciences, NTNU.

Source Separation



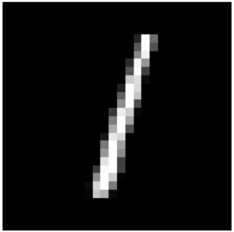
v

=



u_0

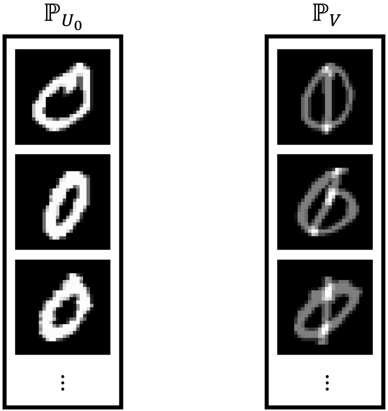
+



u_1

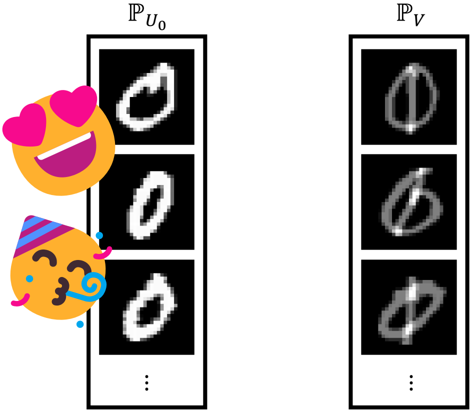
Given a mixed image $v = u_0 + u_1$, can we recover the individual images u_0 (and u_1)?

Semi-Supervision



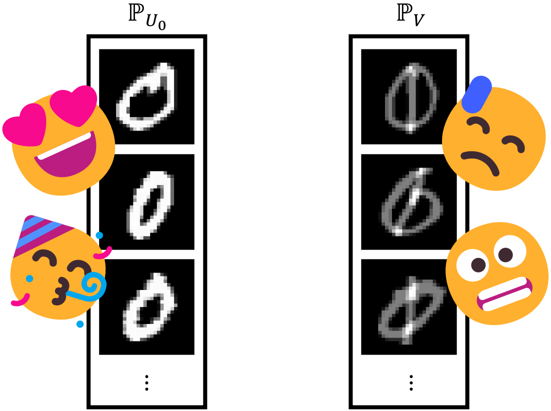
We have two datasets.

Semi Supervision



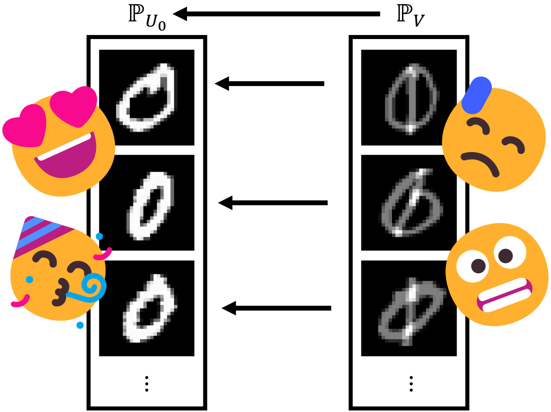
One dataset with clean data that we love!

Semi Supervision



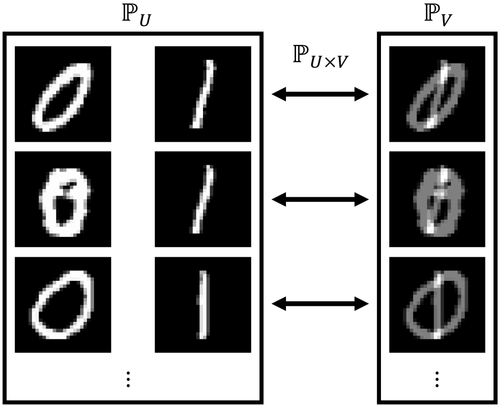
One dataset with corrupted data that we don't want...

Semi Supervision



Can we move the bad data to the distribution of good data? Even if we don't know the noise?

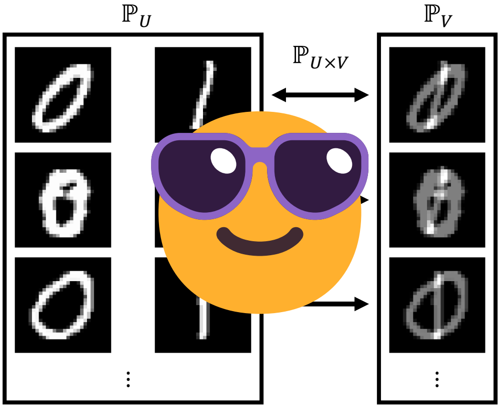
Strong Supervision



In standard learning approaches: Assume we have all data available!

NTNU
Norwegian University of Science and Technology

Strong supervision



Can learn mapping from noisy to good data...
...but data setting is unrealistic...

Singular Value Decomposition as prior/regularization

Question: What is the optimal (orthogonal) basis of dimension d for representing the data stored in $U \in \mathbb{R}^{m \times n}$:

$$\min_{W,H,W^TW=I} \|U - WH\|_F^2 = \min_{W,W^TW=I} \|(I - WW^T)U\|_F^2 ?$$

Singular Value Decomposition as prior/regularization

Question: What is the optimal (orthogonal) basis of dimension d for representing the data stored in $U \in \mathbb{R}^{m \times n}$:

$$\min_{W,H,W^TW=I} \|U - WH\|_F^2 = \min_{W,W^TW=I} \|(I - WW^T)U\|_F^2?$$

- **Answer:** Utilize SVD: $U = X\Sigma Y^T$.
- **Eckart–Young–Mirsky Theorem:** The best rank-d approximation to U is $U_d = X_d \Sigma_d Y_d^T$, keeping only the largest d singular values.
- **Optimal basis:** $W = X_d \in \mathbb{R}^{m \times d}$, usually called a dictionary.

Singular Value Decomposition as prior/regularization

Question: What is the optimal (orthogonal) basis of dimension d for representing the data stored in $U \in \mathbb{R}^{m \times n}$:

$$\min_{W,H,W^TW=I} \|U - WH\|_F^2 = \min_{W,W^TW=I} \|(I - WW^T)U\|_F^2?$$

- **Answer:** Utilize SVD: $U = X\Sigma Y^T$.
- **Eckart–Young–Mirsky Theorem:** The best rank-d approximation to U is $U_d = X_d \Sigma_d Y_d^T$, keeping only the largest d singular values.
- **Optimal basis:** $W = X_d \in \mathbb{R}^{m \times d}$, usually called a dictionary.
- **Prior distribution:** Can model negative log-likelihood of data as $-\ln \mathbb{P}_U(u) = \lambda \|(I - WW^T)u\|^2$, $\lambda \geq 0$.
- In practice it is convenient to let $\lambda \to \infty$. Certainty that the signals lie in the span of W.

Source separation with pre-learnt bases

- Want to separate signal $v = u_0 + u_1$ where $u_0 \sim \mathbb{P}_{U_0}$ and $u_1 \sim \mathbb{P}_{U_1}$.
- Assume we have pre-learnt bases W_0 and W_1 for the two data distributions $\mathbb{P}_{U_0}$ and $\mathbb{P}_{U_1}$. Solve Maximum A Posteriori problem:

$$\min_{h_0, h_1} \| v - W_0 h_0 - W_1 h_1 \|^2$$

Source separation with pre-learnt bases

NTNU

Norwegian University of Science and Technology

- Want to separate signal $v = u_0 + u_1$ where $u_0 \sim \mathbb{P}_{U_0}$ and $u_1 \sim \mathbb{P}_{U_1}$.
- Assume we have pre-learnt bases W_0 and W_1 for the two data distributions $\mathbb{P}_{U_0}$ and $\mathbb{P}_{U_1}$. Solve Maximum A Posteriori problem:

$$\min_{h_0, h_1} \|v - W_0 h_0 - W_1 h_1\|^2$$

- Assumption that signals lie in a (finite) compact, convex space $\mathcal{X}$. For images $\mathcal{X} = [0,1]^m$. Define projection onto this space as $P_{\mathcal{X}}$.
- Recover signals by projection and Wiener filter:

$$u_i = v \odot \frac{P_{\mathcal{X}}(W_i h_i)}{\sum_j P_{\mathcal{X}}(W_j h_j)}$$

What about the unseen source?

Need basis for unseen data $u_1 \sim \mathbb{P}_{U_1}$!

NTNU

Norwegian University of Science and Technology

Source separation with unseen source

- **Solution:** Given mixed data stored in matrix V, fit the basis for the unseen signal while separating:

$$\min_{H_0, H_1, W_1, W_1^T W_1 = I} \|V - W_0 H_0 - W_1 H_1\|_F^2$$

- Can do this by alternatively minimizing H_0, H_1 and W_1, a tiny bit more involved than just fitting the SVD...
- Can you see any issues with this?

Results with low complexity model

Mix

True

Source 0

Source 1

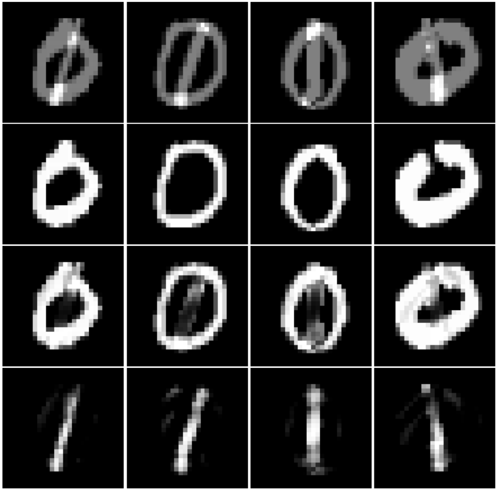


Figure: SVD approach with $d_0 = 16$ and $d_1 = 8$. Mean PSNR: 21.55.

Results with higher complexity model

Mix

True

Source 0

Source 1

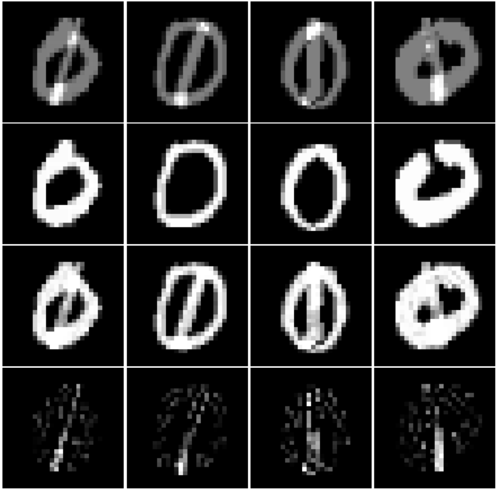


Figure: SVD approach with $d_0 = 128$ and $d_1 = 64$. Mean PSNR: 15.41.

Learning what to not learn with the Wasserstein distance

NTNU
Norwegian University of Science and Technology

- Several authors have proposed using the Wasserstein distance from optimal transport for adversarial learning[1]:

$$\mathbb{W}(\mathbb{P}_V, \mathbb{P}_U) = \min_{\|f\|_L \leq 1} \mathbb{E}_{u \sim \mathbb{P}_U}[f(u)] - \mathbb{E}_{v \sim \mathbb{P}_V}[f(v)],$$

  where $\|.\|_L$ is the Lipschitz norm.
- f is usually parameterized by a neural network.
- Describes the "cost" of transporting one distribution to the other.
- Need to not just fit certain data well, but fit other data poorly!
- Need to learn what not to learn!

[1] *Wasserstein GAN*, M. Arjovsky et al. *Adversarial Regularizers in Inverse Problems*, S. Lunz et al.

Adversarial SVD (ASVD)

- Motivated by this, we propose a modified Wasserstein Distance and look for functions on the form $f(u; W) = \|(I - WW^T)u\|^2$ to obtain:

$$\min_{W, W^T W = I} \underbrace{\frac{1}{N_U}\|(I - WW^T)U\|_F^2}_{\text{Good data}} - \underbrace{\frac{\tau}{N_V}\|(I - WW^T)V\|_F^2}_{\text{Bad data}}$$

- $0 \leq \tau \leq 1$ is the proposed adversarial weight, controlling how concerned we are with not fitting adversarial data.
- $\tau = 0$ is just standard SVD.
- Can show that fitting is equivalent to an eigenvalue problem, and can do some tricks so it is overall faster than standard SVD.

Results with low complexity adversarial model

Mix

True

Source 0

Source 1

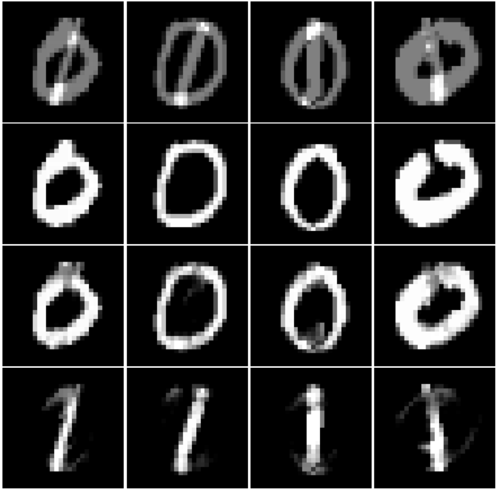


Figure: ASVD approach with $d_0 = 16$ and $d_1 = 8$. Mean PSNR: 24.00.

Results with higher complexity adversarial model

Mix

True

Source 0

Source 1

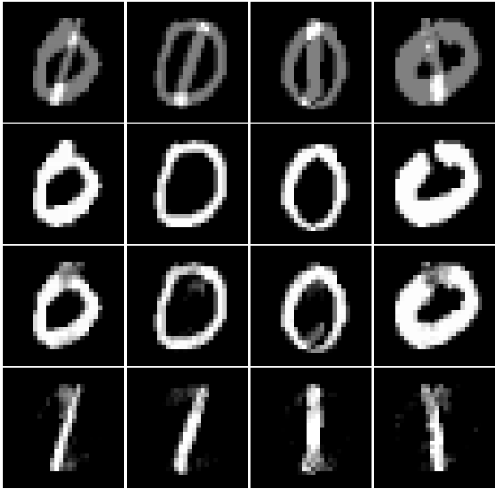


Figure: ASVD approach with $d_0 = 128$ and $d_1 = 64$. Mean PSNR: 25.04.

Shameless Self-Promotion

NTNU

Norwegian University of Science and Technology

Thank you for listening!

Questions?

Some natural questions you can ask if you feel like:

- Is this not the same as just saying the dictionaries for the two sources have to be orthogonal?
- How do you select the adversarial weight τ?
- How does the results depend on the choice of τ?
- Does this extend to more advanced generative methods like GANs?
- Can the learnt bases be reused for different problems?
- Sparsity? Extra regularization?
- Anything else?