Where prior learning can and can't work in unsupervised inverse problems

Florent Michel, Benoît Malézieux, Thomas Moreau, Matthieu Kowalski

Where Prior Learning Can and Can't Work in Unsupervised Inverse Problems

Benoît Malézieux, Florent Michel, Thomas Moreau and Matthieu Kowalski

***Abstract*—Linear inverse problems consist of recovering a signal from its noisy and incomplete (or compressed) observation in a lower dimensional space. Many popular resolution methods rely on data-driven algorithms that learn a prior from pairs of signals and observations to overcome the loss of information. However, these approaches are difficult, if not impossible, to adapt to unsupervised contexts – where no ground truth data are available – due to the need for learning from clean signals. This paper studies necessary and sufficient conditions that do or do not allow learning a prior in unsupervised inverse problems. First, we focus on dictionary learning and point out that recovering the dictionary is unfeasible without constraints when the signal is observed through only one measurement operator. It can, however, be learned with multiple operators, given that they are diverse enough to span the whole signal space. Then, we study methods where weak priors are made available either through optimization constraints or deep learning architectures. We empirically emphasize that they perform better than hand-crafted priors only if they are adapted to the inverse problem.**

***Index Terms*—Inverse problems, unsupervised learning, dictionary learning**

I. Introduction

Linear inverse problems are ubiquitous in observational science such as imaging [1], neurosciences [2], or astrophysics [3]. They consist in reconstructing signals $X \in \mathbb{R}^{n \times N}$ from remote and noisy measurements $Y \in \mathbb{R}^{m \times N}$ which are obtained as a linear transformation $A \in \mathbb{R}^{m \times n}$ of X, corrupted with noise $B \in \mathbb{R}^{m \times N}$: $Y = AX + B$. Here, m and n denote the dimension of the measurements and the signals respectively, and N is the number of problems that need to be solved. As the dimension m of Y is usually much smaller than the dimension n of X, these problems are ill-posed, and several signals could lead to a given set of observations. The measurement uncertainty due to noise also increases the number of signals that could generate some given measurements. To select the inverse problem's most plausible solution among all possible ones, practitioners rely on prior knowledge of the data.

Hand-crafted priors relying on sparsity in a basis produce satisfying results on specific data, such as wavelets in imaging or Gaborlets in audio [4]. However, the complexity and variability of the signals often make *ad hoc* priors inadequate. Alternatively, methods leveraging sparsity also allow summarizing the signal's structure [5]. In particular, dictionary learning [6]–[8] is efficient on pattern learning tasks such as blood cell detection [9] or MEG signals analysis [10]. Finally, the prior can be learned from ground truth data when available. For instance, frameworks based Deep Learning [11]–[13] propose to integrate a pre-trained denoiser in an iterative algorithm to solve the problem using the Plug-and-Play [14] (PnP) approach.

Recently, there has been a notable surge in interest towards the unrolling approach for dictionary learning techniques [15], [16]. Nevertheless, these methods require to have access to clean signals, which are sometime available in audio and imaging but often not accessible in fields like neuroimaging or astrophysics.

While data-driven methods have been extensively studied in the context of supervised inverse problems, recent works have focused on unsupervised scenarios and provided new algorithms to learn from corrupted data only [17]–[19]. The authors of [20] and [21] demonstrate that a necessary condition for the identifiability of the signals from their measurements is either to measure them with multiple operators spanning the whole space or to introduce weak prior knowledge such as group structures and equivariance in the model when only one operator is available. Other works based on Deep Learning have leveraged successful architectures to recover images without access to any ground truth data. In particular, Deep Image Prior shows that CNNs contain enough prior information to recover an image in several inverse problems, such as denoising or inpainting [22]. Finally, a few works have demonstrated that it is possible to learn dictionaries from incomplete data, especially in the context of missing values or inpainting in imaging [23]–[25]. Another line of work studied online factorization of large matrices by aggregating partial information randomly selected from the data at each iteration [26], [27]. This is equivalent to learning a dictionary from incomplete data, except that one sample can be looked at multiple times from different angles, which is hardly possible in an inverse problem context.

Contributions. This article explores unsupervised prior learning for solving inverse problems and identifies practical limitations in this domain. More specifically, assuming that the signal is sparse in some dictionaries and building on the idea that "seeing the whole space" is necessary for accurate signal identification [21], we examine the specific challenge of unsupervised dictionary learning for signal recovery in inverse problems. To our knowledge, only sufficient conditions have been discussed in [28]. The proposed necessary conditions complement how to blindly learn a prior in inverse problems.

We first show in Section II-A that, in an unsupervised

B. Malézieux, Florent Michel and T. Moreau are with Inria, Université Paris-Saclay, CEA, Palaiseau, France (e-mail: forname.name@inria.fr).

M. Kowalski is with Inria, Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numériques, Gif-sur-Yvette, France (e-mail: matthieu.kowalski@universite-paris-saclay.fr).

The authors want to gratefully thank Julian Tachella for fruitful discussions.

This work was supported by grants from Digiteo France.

setting, dictionary learning can only learn a dictionary in $\ker(A)^\perp$. We study in Section II-B the possibility of dictionary learning through multiple operators. We show that the previous result extends straightforwardly, and in this case, seeing the whole space is insufficient because the measurement operators make the problem more challenging.

In a second contribution, we investigate the effectiveness of convolution-based techniques in signal recovery tasks. We conduct an extensive experiment in Section III to assess the practical behavior of three popular methods: (i) Convolutional Dictionary Learning to demonstrate our previous findings on dictionary learning, (ii) Deep Image Prior where the prior is its deep convolutional architecture, and (iii) Plug-and-Play methods (PnP), which uses a convolutive neural network denoiser as a proximal operator in variational methods. Our results reveal that while these methods are effective for inpainting, they require additional prior information for deblurring. In other words, this weak convolution prior must be adapted for deblurring while remaining suitable for inpainting tasks.

II. The main bottleneck of prior learning in inverse problems

When dealing with inverse problems, it is common for the dimension of the signal to be larger than the dimension of the measurements. This means that some of the information about the signal is lost during the observation process, particularly in the null space of the operator A. To address this issue, we investigate the impact of this dimension reduction on dictionary learning, which involves learning a set of basis vectors that can be used to reconstruct the signal. In Proposition II.2, we present a theoretical analysis of the impact of this degradation on dictionary learning for a single operator, and we demonstrate that this necessary condition extends to multiple operators and what are the practical implications.

A. Dictionary learning with a single measurement operator

Dictionary learning assumes that the signal can be decomposed into a sparse representation in a redundant basis of patterns – also called atoms. In other words, the goal is to recover the signals $X \in \mathbb{R}^{n\times N}$ as DZ where $Z \in \mathbb{R}^{L\times N}$ are sparse codes and $D \in \mathbb{R}^{n\times L}$ is a dictionary with L atoms. Taking the example of Lasso-based dictionary learning, recovering X would require solving a problem of the form

$$\min_{Z\in\mathbb{R}^{L\times N}, D\in\mathcal{C}} \frac{1}{2}\left\|ADZ - Y\right\|_2^2 + \lambda\left\|Z\right\|_1 \quad , \tag{1}$$

where λ is a regularization hyperparameter and $\mathcal{C}$ is a set of constraints, typically set so that columns of D have a norm smaller than 1.

We first aim to see the impact of A on the algorithm's ability to recover a proper dictionary. In Proposition II.1, we focus on inpainting where the measurement operator is a binary mask or equivalently a diagonal matrix with m non-zeros elements.

Proposition II.1. *Consider a diagonal measurement matrix* $A = diag(a_1, \dots, a_m, 0, \dots, 0) \in \mathbb{R}^{n\times n}$ *where* $m < n$ *and* $a_1 \geq \cdots \geq a_m > 0$ *. Let* $D_0 \in \mathbb{R}^{n\times L}$ *and* D' *be such that*

$$D' = \begin{pmatrix} \frac{\|D_{0,j}\|}{\|D_{0,j,m}\|} D_{0,j,m} \\ 0_{n-m} \end{pmatrix}_{1\leq j\leq L}, \text{ where } D_0 = \begin{pmatrix} D_{0,m} \\ D_{0,n-m} \end{pmatrix}$$

Then

$$\min_Z \frac{1}{2}\left\|AD'Z - Y\right\|_2^2 + \lambda\left\|Z\right\|_1 \leq \min_Z \frac{1}{2}\left\|AD_0Z - Y\right\|_2^2 + \lambda\left\|Z\right\|_1 \quad .$$

All proofs are deferred to Appendix A. In this simple case, our proposition shows that the optimal dictionary must be 0 in the null space of A. The core idea behind the proof is that due to invariances, the optimal solution for dictionary learning is contained in an equivalence class $\{PSD' + V\}$ where P is a permutation matrix, S is a scaling matrix, D' is a matrix of rank at most m and V is a matrix of rank at most $n-m$ such that $PSD' \in \ker(A)^\perp$ and $V \in \ker(A)$. Given a dictionary $PSD' + V$ in this equivalence class, the dictionary PSD' is always a better minimizer after proper rescaling. Therefore, the solver puts to 0 all directions from which A loses the information and increases the norm of the other rows while reducing the corresponding value in Z and therefore reducing the ℓ_1 norm. Proposition II.2 generalizes Proposition II.1 to the case of rectangular matrices.

Proposition II.2. *Let* $A \in \mathbb{R}^{m\times n}$ *be a measurement matrix where* $m < n$*, and let* $Y \in \mathbb{R}^{m\times N}$ *be the observed data. If a dictionary* $D \in \mathbb{R}^{n\times L}$ *is a solution to*

$$\min_{Z\in\mathbb{R}^{L\times N}, D\in\mathcal{C}} \frac{1}{2}\left\|ADZ - Y\right\|_2^2 + \lambda\left\|Z\right\|_1 \quad .$$

then $D \in \ker(A)^\perp$.

In essence, if we use only one measurement matrix, we can only anticipate to acquire a dictionary of rank that does not exceed m. According to [21], when the signal space is assumed to have no further constraints, a single operator with $m < n$ cannot recover the signal. Here, we complement this result by showing that one can only learn a dictionary within the range of the operator. In [28], it has been shown that the dictionary can be uniquely identified under specific constraints on the operator. Proposition II.2 shows that it is impossible to identify anything outside the range space of the operator.

Dimension reduction makes dictionary learning harder in the range space. Even in the range space of the signal, a good dictionary cannot always be learned reliably. Guarantees of identifiability or local recovery are strongly based on the accurate estimation of the sparse code Z [29]–[31]. As the dimension of the measurement m becomes smaller than the dimension of the signal n, this estimation becomes less stable. As an example, if D is a Gaussian random dictionary, the theory of compressed sensing states that $n \geq 2s\ln(\frac{L}{s})$, s being the sparsity of Z, is a sufficient condition to be able to recover Z with high probability [32]. When the dictionary is degraded by a matrix A, this constraint becomes $m \geq 2s\ln(\frac{L}{s})$ and the sparse code that can be reliably recovered needs to be much sparser, with a ratio close to $\frac{m}{n}$, to compensate for the loss

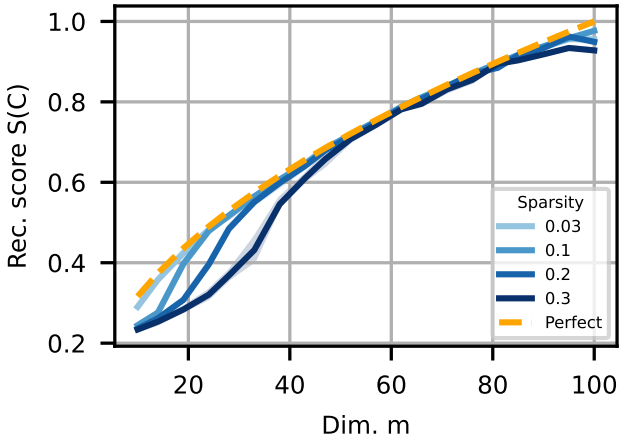


Fig. 1: Recovery score for Gaussian dictionaries 100×100, after degradation by a single compressed sensing operator $m \times 100$. When the dimension m decreases, the part of the dictionary not contained in the null space can be recovered only with sparse signals.

of information. This implies that recovering the part of the dictionary not contained in the null space of A becomes harder with the corruption of the data.

To demonstrate the challenges of retrieving a dictionary in inverse problems, we conducted an experiment as follows. We created a 100×100 random Gaussian dictionary and a sparse signal using the typical Bernoulli-Gaussian model. Then, we generated the measurement by applying a compressed sensing operator. After that, we estimated the dictionary and the signal by solving Eq. (1).

We evaluate the quality of the dictionary, based on the Pearson correlation of their columns. To make the metric sign and permutation invariant, we use a best linear sum assignment $S(C) = \max_{\sigma \in \mathfrak{S}_n} \frac{1}{n} \sum_{i=1}^{n} |C_{\sigma(i),i}|$, where $\mathfrak{S}_n$ is the group of permutations of $[1, n]$ and C is the cost matrix whose entry i, j compares the atom i of the first dictionary and j of the second. This metric can be computed using the Hungarian algorithm [33]. It is equal to 1 when the dictionary is perfectly recovered. We compare the score obtained for various sparsity levels and for varying measurement sizes m to the perfect score that can be achieved by taking the projection of the original dictionary in $\ker(A)^{\perp}$. Fig. 1 shows the recovery score depending on the size of the measurements and on the sparsity of the Bernoulli Gaussian signal. The recovery score drops when the dimension m decreases and small values of m require a high sparsity level to recover the dictionary in the range of A.

B. Seeing the data through multiple operators

Even though it is not possible to recover the whole dictionary from a single measurement operator, the situation changes when the measurement matrix is sample-dependent. Indeed, several operators may span different parts of the signal space and make it possible to recover a meaningful prior for the missing part of the signal. In this section, we focus on cases where the signals are observed through a set of N_m measurement matrices $(A_i)_{1 \leq i \leq N_m}$, and consider the task of learning a dictionary with the associated lasso-based optimization problem

$$\begin{aligned} &\min_{D \in \mathcal{C}} F(D) \\ &\triangleq \frac{1}{2} \sum_{i=1}^{N_m} \|A_i D Z_{A_i}(D) - Y_i\|_2^2 + \lambda_i \|Z_{A_i}(D)\|_1 \quad , \end{aligned} \tag{2}$$

with

$$Z_{A_i}(D) = \operatorname{argmin}_Z \frac{1}{2} \|A_i D Z - Y_i\|_2^2 + \lambda_i \|Z\|_1 \tag{3}$$

Here $Z_A(D) = (Z_{A_1}(D), \dots Z_{A_{N_m}}(D))$ denotes the sparse codes related to each operator.

This problem is non-convex and usually solved through gradient descent, to find a local minimum. In the following, we highlight cases when the local minima of Eq. (2) are also local minima for the problem without observation operators and provide an empirical analysis in different scenarios. With multiple measurement operators, the gradient of Eq. (2) is given by

$$\nabla_D F(D) = \sum_{i=1}^{N_m} A_i^T (A_i D Z_{A_i}(D) - Y_i) Z_{A_i}(D)^T \quad . \tag{4}$$

The main difficulty in studying this quantity is that the sparse codes estimate $Z_{A_i}(D)$ depends on D and A_i. Each operator provides measurements from a limited number of samples in the dataset, and the sparse codes are different with and without A. Thus, we consider the simplest case where $A_i = I$. This is an easier problem than the general formulation in Eq. (2), as if this is not feasible, then the original formulation is not feasible either. In this case, we have

$$\nabla_D F(D) = (\sum_i A_i^T A_i) \nabla_D F(D) \quad . \tag{5}$$

KKT conditions imply that the gradient $\nabla_D F(D)$ must vanish at local minima. Whenever $\sum_i A_i^T A_i$ is injective, $\nabla_D F(D)$ vanishes if and only if $\nabla_D F(D)$ vanishes. Thus, local minima of Eq. (2) are also local minima for the original problem where $A_i = I$. This means that when $\sum_i A_i^T A_i$ spans the entire space, the dictionary from the original problem can be recovered. This case boils down to the case previously studied in Section II, as $\sum_i A_i^T A_i$ is full rank whenever the rank of the matrix obtained by stacking the operators $(A_1^T, \cdots, A_{N_m}^T)$ is equal to n. Otherwise, local minima of $F(D)$ are not necessarily local minima of $F(D)$. This result stresses again that seeing the whole space through the measurement operators is necessary. It is however important to note that this is only a necessary condition to recover the dictionary, as sparse coding guarantees may not be met when the dimension m is too small.

We present in the following two examples of inverse problems with multiple operators to illustrate the challenge of unsupervised dictionary learning.

Example II.3 (Compressed sensing (CS))**.** *Consider the case where all A_i are random matrices with independent Gaussian entries. In this case, $(A_1^T, \cdots, A_{N_m}^T)$ is also a random*

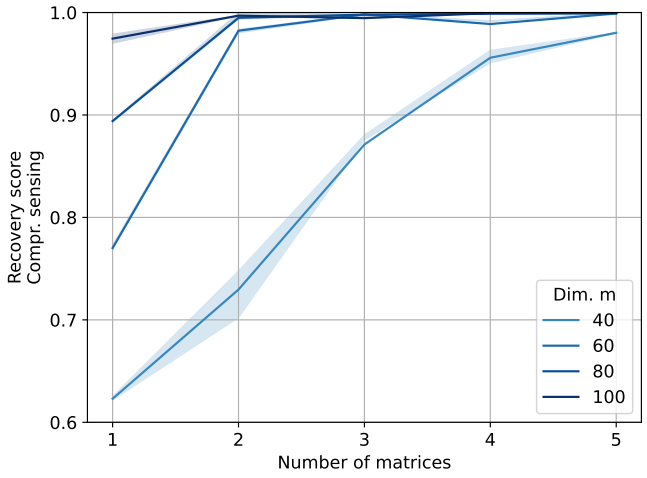


Fig. 2: Recovery score for Gaussian dictionaries 100×100, after degradation by N_m compressed sensing operators $m \times 100$ of Bernoulli Gaussian signals with sparsity $s = 0.1$. A necessary condition to recover the dictionary is $N_m \geq \lfloor \frac{n}{m} \rfloor + 1$ but it is not sufficient when m is too small.

Gaussian matrix of dimension $n \times N_m m$. Therefore, it is of rank n with probability 1 if $N_m \geq \lfloor \frac{n}{m} \rfloor + 1$. Fig. 2 illustrates that it is a necessary condition to recover D, but it is not sufficient when m is too small, possibly because the sparse coding becomes inefficient. See [34]–[36] for more theoretical insight on multiview compressive dictionary learning.

Example II.4 (Inpainting)**.** *Now consider the case where all A_i are binary masks with coefficients following Bernoulli distributions of parameters $p_1, \cdots, p_n$, i.e. $A_i = diag(a_i^1, \cdots, a_i^n)$ where each a_i^j is equal to 1 with probability p_j. The rank of $(A_1^T, \cdots, A_{N_m}^T)$ is equal to n if for each coordinate j there exists an index i such that $a_i^j = 1$. This happens with probability $\prod_j (1 - (1-p_j)^{N_m})$. Fig. 3 shows that similar to CS, this is a necessary but insufficient condition to recover a proper dictionary. Even when the number of samples compensates for missing values, the sparsity of the data plays a great role in the ability of the algorithm to recover the proper dictionary after heavy dimension reduction.*

To illustrate what happens with real data, we consider the example of image inpainting. Let $A \in \{0,1\}^{h \times w}$ be a binary mask used to observe an image $X \in [0,1]^{h \times w}$ and $Y = A \odot X$ be the observed image. While the operator is unique when we consider the whole image at once, learning a dictionary from patches of size n from the image is equivalent to learning a dictionary with multiple operators in Eq. (2). Denoting $A_{ij} = diag(A_{ni:n(i+1),nj:n(j+1)})$ and $Y_{ij} = vect(Y_{ni:n(i+1),nj:n(j+1)})$ the i,j-patch, patch-based dictionary learning solves

$$\min_{Z_{ij}, D \in \mathcal{C}} \sum_{i,j} \frac{1}{2} \|A_{ij} D Z_{ij} - Y_{ij}\|_2^2 + \lambda \|Z_{ij}\|_1 \qquad (6)$$

The dictionary should be recovered if the image is large enough and if there are not too many masked pixels. In Fig. 4, we show the PSNR (Peak Signal to Noise ratio) and the recovery score depending on the proportion of missing values in a grey-level 256×256 image with a binary mask containing a given level of missing pixels. Then we split this image into 10×10 patches and we learn a dictionary of size 100 on these patches, with $\lambda = 0.1$. We compute the recovery score by comparing dictionaries obtained with and without the mask, and we weight the costs by the average of sparse activations Z in absolute value for a given atom

$$W = \left(\textstyle\sum_{i=1}^N |Z_{1,i}|, \cdots, \sum_{i=1}^N |Z_{L,i}|\right)^T / \sum_{i,j} |Z_{i,j}| \qquad (7)$$

$$C = D_0^T (D^T \odot W)^T \ . \qquad (8)$$

This score better reflects the usefulness of the atoms, and allows to take into account the fact that dictionaries learned on natural signals may contain irrelevant atoms that are rarely used. We repeat this experiment 10 times. The recovery score drops when the proportion of missing values is larger than 50%. Otherwise, the image is successfully recovered even when the dictionary is learned from the degraded observation. This is why dictionary learning led to good results in unsupervised inpainting in the literature [23]–[25].

Proposition II.2 shows that dictionary learning won't operate in the null space of the measurement matrix. Using multiple operators can mitigate this issue, as the whole signal space is seen through different matrices A_i, reducing the effective null space. However, our experiments with synthetic and real data also show that this is only a necessary condition to learn a good dictionary. In some cases, the sparse codes cannot be recovered as the information is too degraded. Reducing the dimension of the observations could then be a hard limit to dictionary learning. In the following, we show that well-chosen weak prior knowledge can lift the problem and allow the recovery of the information from the kernel space of a single operator through the example of convolutions in imaging.

III. Weak prior knowledge through convolutions

The usage of convolutions in Deep Learning [37] has encountered tremendous success in a broad range of tasks from image classification to reconstruction. Convolutional neural networks efficiently analyze translation invariant data while reducing the number of parameters.

This section aims to explain the effectiveness and limitations of convolutions as weak prior knowledge for unsupervised image reconstruction. After briefly discussing the methods involved, we conduct a comprehensive experimental study of prior learning based on convolutions for deblurring and inpainting problems.

All computations have been performed on a GPU NVIDIA Tesla V100-DGXS 32GB using `PyTorch` [38].[1]

A. Prior learning methods based on convolutions

We study three methods based on prior learning: Convolutional Dictionary Learning [39], Plug and Play [11] and Deep Image Prior [22]. Convolutional dictionary learning (CDL) consists in learning convolutional kernels of relatively small

[1]Code is available at https://github.com/bmalezieux/dl_inv_prob.

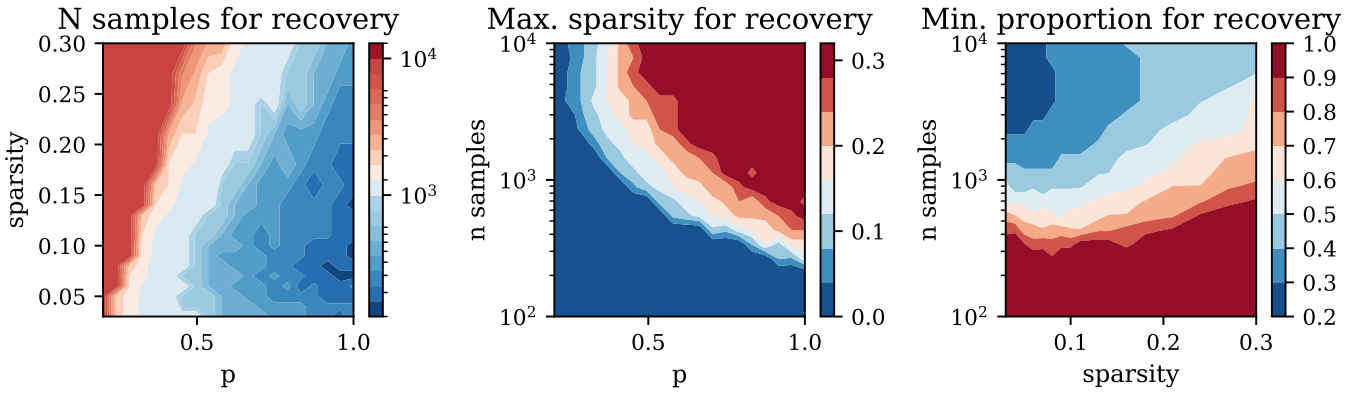


Fig. 3: We generate the data from a Gaussian dictionary of size 100×100 and Bernoulli Gaussian sparse codes of sparsity s (average rate of non-zero coordinates). Then, we degrade the data with a binary mask of variable rates of available coordinates p. We learn a dictionary of size 100×100 over several values of λ. We show the ability of the algorithm to recover the dictionary, depending on p, the number of training samples, and the level of sparsity in the data. Dictionary recovery is defined as obtaining a recovery score of at least 0.95. We display the results from three different perspectives: (**left**) Minimal number of samples necessary to recover the dictionary depending on sparsity and rate of available coordinates in the data. A number of samples larger than 10^4 means no recovery possible. (**center**) Maximal level of sparsity s (maximal proportion of non zero coordinates) to recover the dictionary depending on the number of samples and the rate of available coordinates. A level equal to 0 means no recovery possible. (**right**) Minimal rate of available coordinates for recovery depending on the number of samples and the level of sparsity. A level equal to 1 means no recovery possible. These figures show that there is a hard limit to what can be learned depending on the proportion of missing values and sparsity, regardless the number of training samples. Having access to the whole signal space is not a sufficient condition to recover the dictionary.

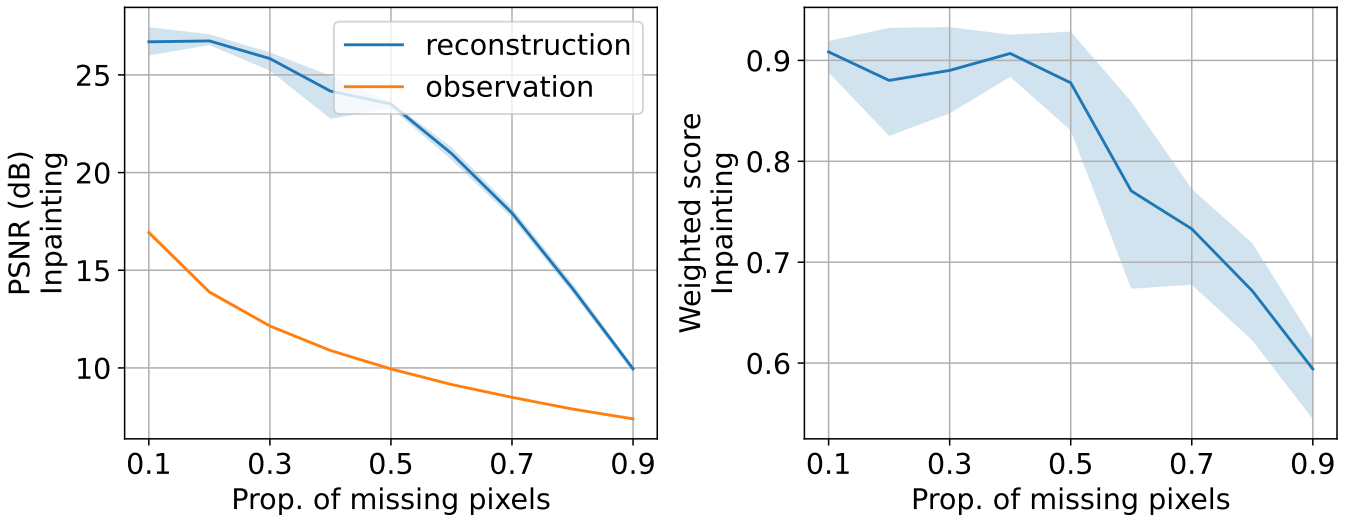


Fig. 4: For inpainting, PSNR (**left**) and weighted recovery score (**right**) depend on the proportion of missing values in dictionary learning on patches from a natural image. When the dimension of the measurement space is large enough, the algorithm successfully recovers the image and the supervised dictionary.

dimensions from a signal Y to sparsely reconstruct the signal. The ℓ_1-based formulation reads

$$\min_{z_k, d_k \in \mathcal{C}} \frac{1}{2} \left\| A \sum_k d_k * z_k - Y \right\|_2^2 + \lambda \sum_k \|z_k\|_1 \quad . \tag{9}$$

Deep Image Prior (DIP) takes advantage of CNN architectures to project the observed image into a well-suited range space by drawing a random code vector z in the latent space and optimizing the parameters of the network f as follows

$$\min_\theta \|Y - A f_\theta(z)\|_2^2 \quad . \tag{10}$$

Plug and Play (PnP) is an iterative algorithm inspired by proximal gradient descent, which recovers images from an observation Y with steps of the form

$$X_{n+1} = f_\theta \left(X_n - \tau A^*(AX_n - Y) \right) \quad \forall n \geq 1 \ , \tag{11}$$

where $X_0 = 0$, τ is a step size and f_θ is an image denoiser.

CDL and DIP are unsupervised methods that can be applied to a single observation, as the prior is learned directly from the degraded data or fixed without requiring a training set. In contrast, PnP usually resorts to a deep denoiser for f_θ, which is generally trained on a database of clean images. As we focus on the unsupervised setting, we adapt PnP by training the denoiser on degraded data instead. Note that here, we depart from the classical PnP literature, as f_θ is amenable to a reconstructor and not a denoiser, in line with the recent results extending PnP with more complex restoration operation [40]. In this case, we consider that we have access to a dataset $(Y_i)_{1\leq i\leq N}$ where each $Y_i = AX_i + \epsilon_i$ is an observation of an original image X_i degraded by the same operator A and a Gaussian noise ϵ_i. We generate noisy images $(Y_i')_{1\leq i\leq N}$ from our dataset of observations $Y_i' = Y_i + \epsilon_i'$, and we train a DnCNN [41] to recover Y_i from Y_i' in the range space of A by minimizing

$$\min_\theta \frac{1}{N}\sum_i \|A(f_\theta(Y_i') - Y_i)\|_2^2 \quad . \tag{12}$$

where X and Y have the same dimension (up to well-chosen boundary conditions). The idea is to check in which case the architecture can compensate for the lack of information in the kernel of A by learning from the information in the range space of A. To point out the limits of these prior learning algorithms, we will compare them to two reconstruction methods based on Total Variation (TV) [42] and sparse wavelets [4].

The purpose is to highlight the hard limits of unsupervised methods in various contexts. Therefore, we evaluate the performance reached by each algorithm over oracle hyper-parameters, namely hyper-parameters leading to the best performances. While evaluating hyper-parameter sensitivity is necessary when comparing different methods, it is orthogonal to our study, which considers the difference between supervised and unsupervised training of similar methods.

B. Why convolutions are likely to work on tasks like inpainting

Works on prior learning in unsupervised inverse problems often evaluate the performance of the methods they propose on an inpainting task [20], [22] and achieve very good performance compared to supervised learning techniques. Here, we provide elements to understand why this task is feasible when using convolutional dictionaries or neural networks without access to ground truth data.

a) Learning convolutional dictionaries from incomplete data: To understand what happens in inpainting, let's consider a simple one-dimensional signal example. Let X_t be a wide sense stationary (WSS) random process, and let A_t be an i.i.d Bernoulli process of mean ρ. The observed signal $Y_t = A_t X_t$ is also a WSS random process and its auto-correlation function $R_Y(\tau)$ is

$$R_Y(\tau) = \mathbb{E}[A_t X_t A_{t+\tau} X_{t+\tau}] = R_X(\tau)\,\mathbb{E}[A_t A_{t+\tau}] \tag{13}$$
$$= \rho^2 R_X(\tau)\mathbb{1}_{\tau\neq 0} + \rho R_X(\tau)\mathbb{1}_{\tau=0} \quad . \tag{14}$$

Then, the Wiener-Khintchine theorem assures that the power spectral density of X and Y are proportional. This shows that with sufficient samples in the signal, the masking process won't affect the spectrum of the original signal X, and translation invariant priors can take advantage of the information from all frequencies. This means that the original signal is completely observed, as all frequencies can be recovered, and the task can thus be solved. Intuitively, when learning parameters of small convolution kernels on translation-invariant signals, sub-parts of the signal act as different observations, with potentially different measurement operators. For sufficiently large signals, as the missing pixels are distributed randomly, the sub-parts of the signal are fully observed, making the inverse problem feasible.

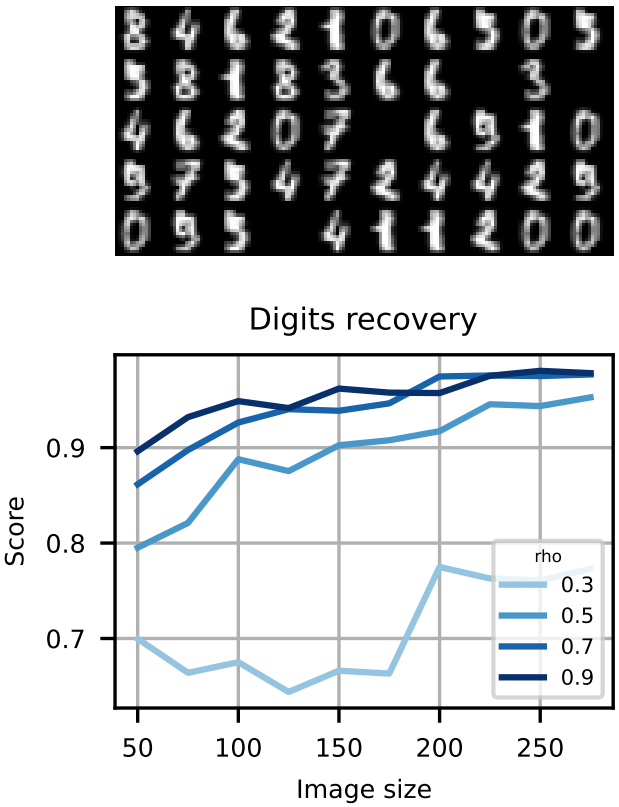
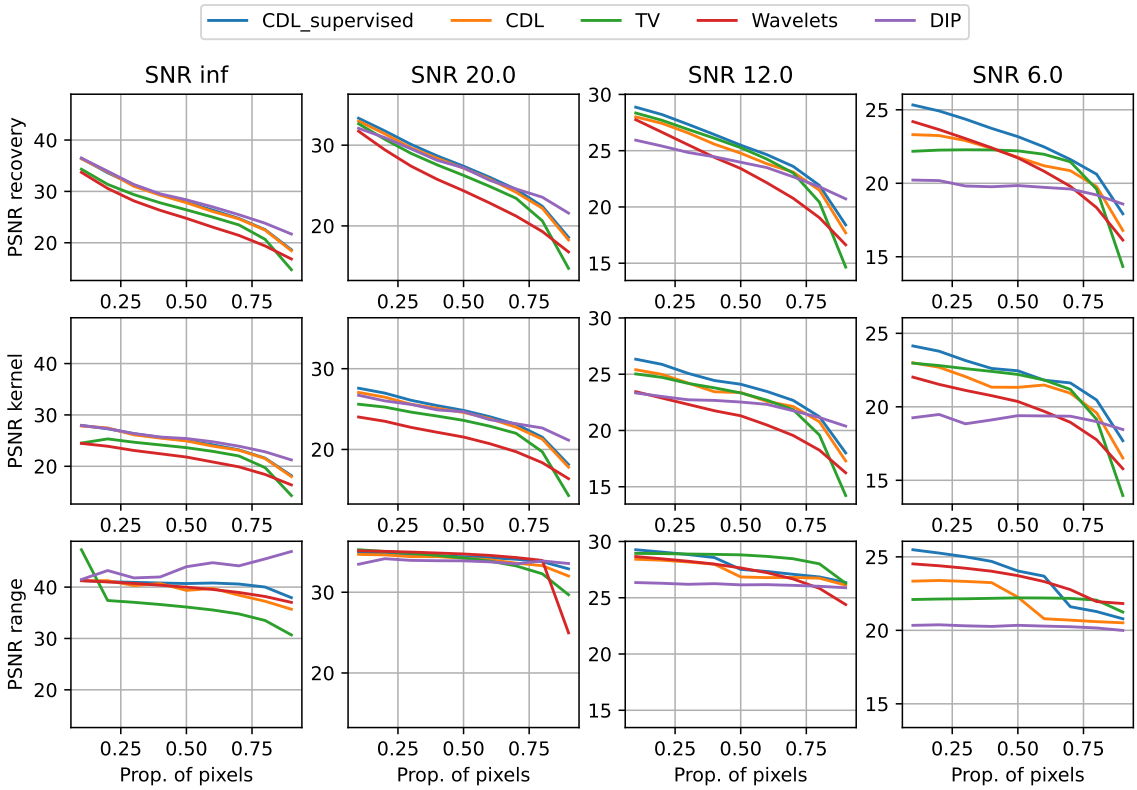
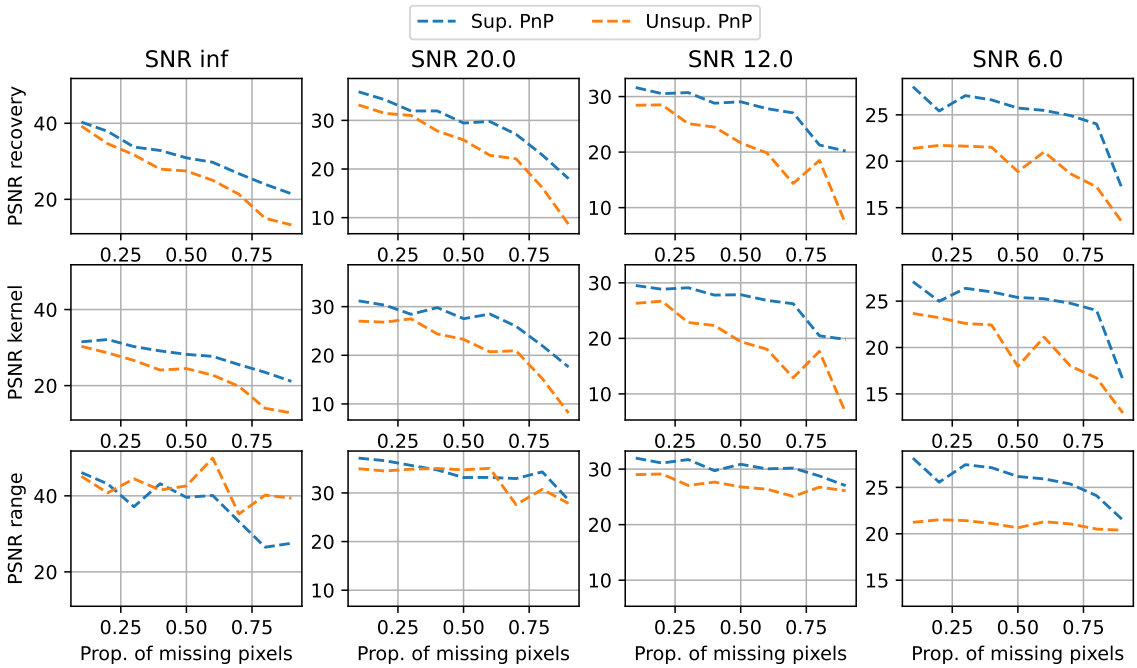


Fig. 5: The recovery score of convolutional dictionaries depends on the image size and the rate of available pixels ρ. Increasing the size improves the quality at high enough rates.

b) Dictionary recovery: We illustrate the practical implication of this observation on the ability of CDL to recover a dictionary composed of 10 digits from an image, depending on the size of the image and the rate of available pixels ρ in Fig. 5. As expected, the recovery of the dictionary used to generate the image increases with the size of the image when ρ is not too low (> 0.5). This shows that when the signal is large enough, learning the dictionary from the degraded image is possible when using convolution to leverage the translation invariance of the image. It is essential to note that having access to all frequencies is only a necessary condition to learn a good dictionary, as sparse coding assumptions are not met when there are too many missing values. Note that these



(a) PSNR depending on the proportion of missing pixels and noise for CDL, DIP, TV, and wavelets based reconstruction on a 256 $\times$ 256 grey-level image. Unsupervised prior learning works only when the noise is not too high.



(b) PSNR depending on the proportion of available pixels and noise for supervised and unsupervised Plug and Play. When the noise is too high, unsupervised PnP fails to recover the image both in the kernel and the range space of A.

Fig. 6: PSNR of unsupervised, self-supervised, and supervised methods for inpainting

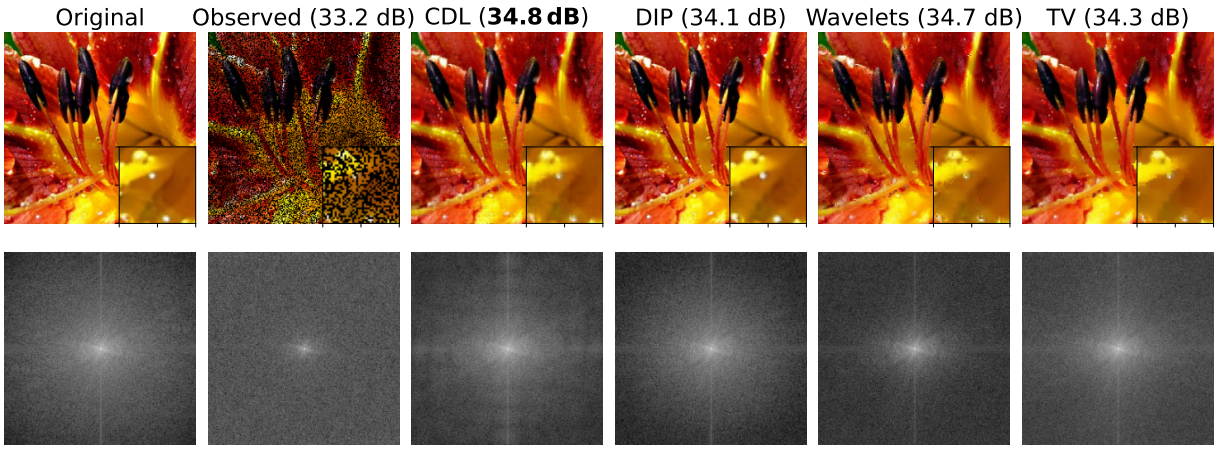


Fig. 7: Reconstruction, PSNR, and PSD of a 256 × 256 RGB image with 50% missing pixels in a noiseless scenario. The PSD reveals the presence of ringing artifacts in the reconstruction by CDL. Unsupervised algorithms recover the whole spectrum of the original image and do as well as hand-crafted methods.

observations would not stand for non-stationary signals, which would not be translation-invariant.

c) Unsupervised reconstruction: Similar effects can be observed for image reconstruction. Natural images are stable enough to allow convolution-based algorithms to learn from all frequencies that are present in the signal. Fig. 6a presents an example where a single natural image is degraded by a random binary mask and Gaussian noise and reports the PSNR of the reconstruction in the mask kernel space and range space for CDL, DIP, and methods based on TV and sparse wavelets for different rates of missing pixels with various levels of SNR: noiseless, 20dB, 12dB, 6dB. The reconstruction error is decomposed between the range space and the kernel space which have different behaviors. For low noise levels, we see that unsupervised approaches match the performance of the supervised approach, showing that the algorithms behave as if they were seeing the complete data. Note that the global loss in PSNR is also mostly driven by the PSNR decrease in the kernel space, as this space gets larger when the number of missing pixels increases. When the noise is too high, unsupervised prior learning methods fail to learn a proper prior even in the range space of A, leading to poor overall results.

Similar observations can be made with PnP methods. Here, supervised PnP refers to a PnP reconstruction algorithm with a denoiser learned on the clean signal, whereas unsupervised PnP means that the prior is learned directly from the observation with (12). In both cases, we train a DnCNN-based denoiser with images from the dataset Imagenette[2] and plug it into an iterative reconstruction algorithm. The results are shown in Fig. 6b for several values of SNR and the proportion of missing pixels. As for the single image example, unsupervised PnP can recover information in the kernel space of A from what is learned in the range space of A and performs closely to its supervised counterpart as long as the rate of masked pixels and the noise are not too large, i.e., when SNR $\geq$ 20 and the rate of missing pixels stays below 50%. With higher noise levels, the performance of unsupervised PnP degrades in the range space, hence in the kernel space.

The experiments highlight that unsupervised methods work as well as supervised CDL and are better than hand-crafted priors in the kernel of A when the noise level is not too high (SNR $\geq$ 20). They succeed in learning in the range space of A and generalizing in the kernel space. TV and wavelets tends to be more robust when the noise increases, as it becomes more challenging to learn the signal's structure. Fig. 7 provides a visual example in the noiseless case. Unsupervised algorithms successfully recover the original image after degradation by a binary mask with 50% pixels missing. The PSD shows that low and high frequencies are retrieved, despite ringing artifacts in the case of CDL. However, CDL and DIP are sensitive to noise and fail to recover relevant frequencies from the observations in noisy scenarios.

C. The pitfall of convolutions in deblurring

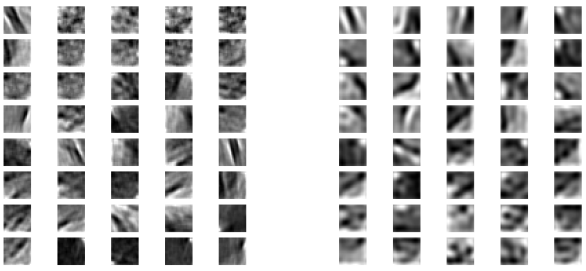


Fig. 8: 40 atoms of a dictionary learned with CDL ($\lambda = 0.1$, atoms of size 20×20) for inpainting (left) and deblurring (right) on the image of Figure 7. High frequencies are put to 0 in the case of deblurring, leading to blurry atoms.

[2]The data are available at https://github.com/fastai/imagenette

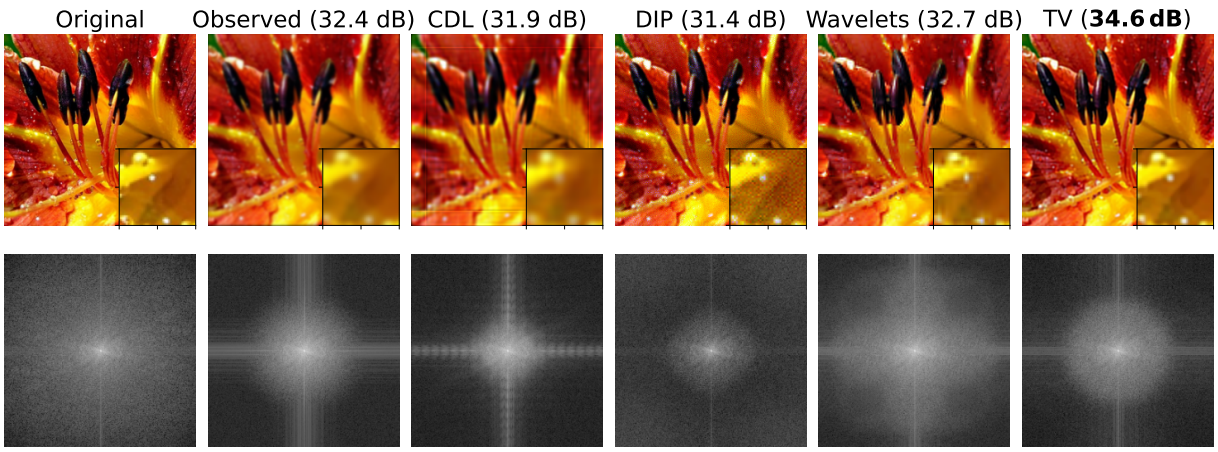


Fig. 9: In a noiseless scenario, reconstruction, PSNR and PSD of a 256 $\times$ 256 RGB image blurred by a normalized 10 $\times$ 10 Gaussian kernel with standard deviation 0.3. The PSD clearly shows that nothing is learned in high frequencies compared to what was obtained in inpainting.

Convolutions work well when all frequencies are preserved, as shown for inpainting. However, several inverse problems involve recovering a signal where part of the spectrum is missing. In super-resolution, all odd frequencies in the signal are lost. In deblurring, the signal is observed after degradation by a low-pass filter. As mentioned in Section II-B, learning from the degraded observation leads to failure in this case. And unlike with inpainting, the range space of such Fourier-based operators stays the same after composition with a translation operator. Thus, these problems do not benefit from convolutive priors since they do not allow to complete the spectrum and a part of the signal space is left unseen. We will focus on the example of deblurring in the following.

Using the Parseval equality, the CDL problem can be re-written in the Fourier domain

$$\min_{z_k, d_k} \frac{1}{2} \left\| \hat{A} \sum_k \hat{d}_k \hat{z}_k - \hat{Y} \right\|_2^2 + \lambda \sum_k \|z_k\|_1 \quad , \qquad (15)$$

with $\hat{x}$ denoting the Fourier transform of a signal x. For deblurring, as the spectrum $\hat{A}$ is low-pass, all information about the is high frequencies are lost. Thus, optimal dictionaries contain atoms $(d_k)_k$ with high frequencies set to 0, for the same reason as pointed out in Proposition II.1. This is illustrated in Fig. 8 where we can see that the atoms learned on blurred images do not contain high frequencies (they appear blurry) whereas atoms learned on inpainted images contain high frequencies.

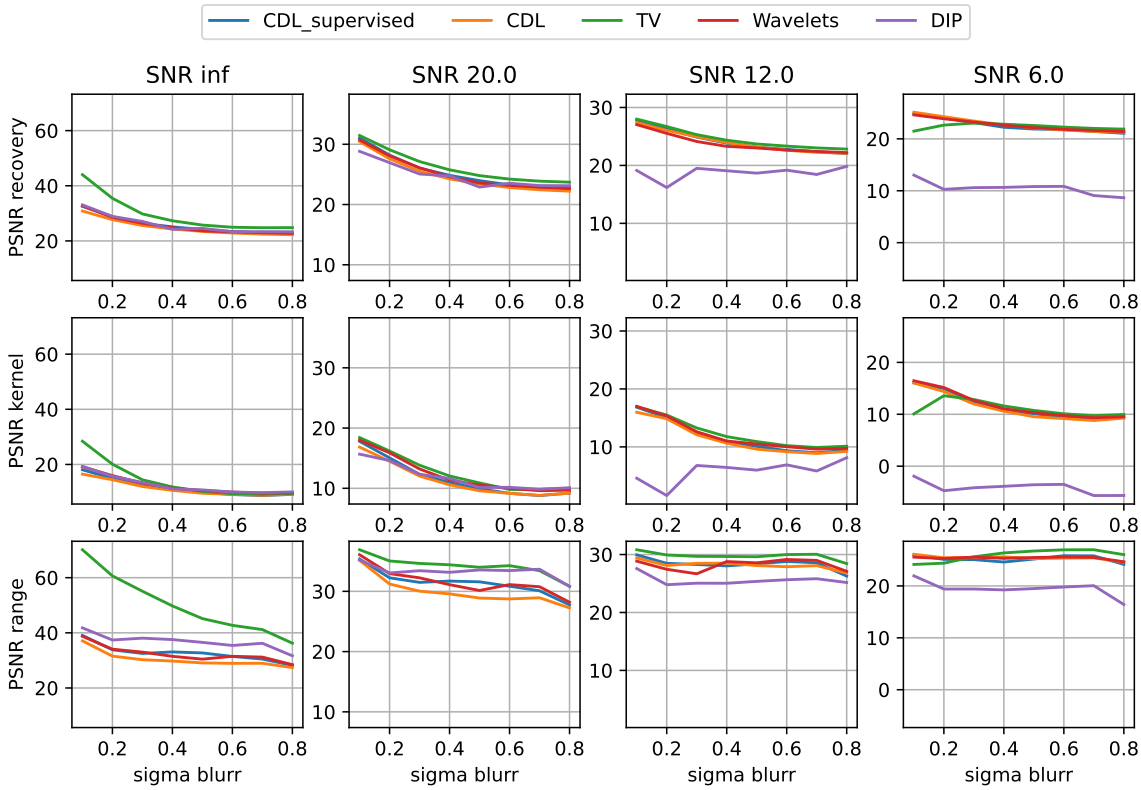
For a blurred image, Fig. 9 displays its reconstructions and their PSD for various methods. These results indicate that neither CDL nor DIP recover information outside the span of the blur, i.e., in high frequencies. While CDL puts all high frequencies to 0, DIP adds noise. We also see that Wavelets are able to better recover high frequencies due to their particular structure that links the low and high frequencies. Fig. 10a shows the performances of these methods for various blur sizes, decomposed between the kernel and range spaces. All methods show a rapid decrease in performance, due to bad reconstruction in the kernel space. We also notice that supervised CDL does not improve over unsupervised CDL. We conjecture that this is because the learned dictionary does not allow to as it does not link the information from the kernel space and the range space.

The same phenomenon appears with PnP in Fig. 10b: there is a performance gap between supervised and unsupervised learning in the kernel space. This shows that the performance loss from unsupervised learning are again mostly driven by the loss in the kernel space. However, unlike with inpainting, even for low blur sizes, the loss between unsupervised and supervised learning is significant at all noise level, showing that this problem is intrinsically harder than inpainting, in the sense that we need a prior with a structure adapted to recover high frequencies, such as scale invariance proposed in [43].

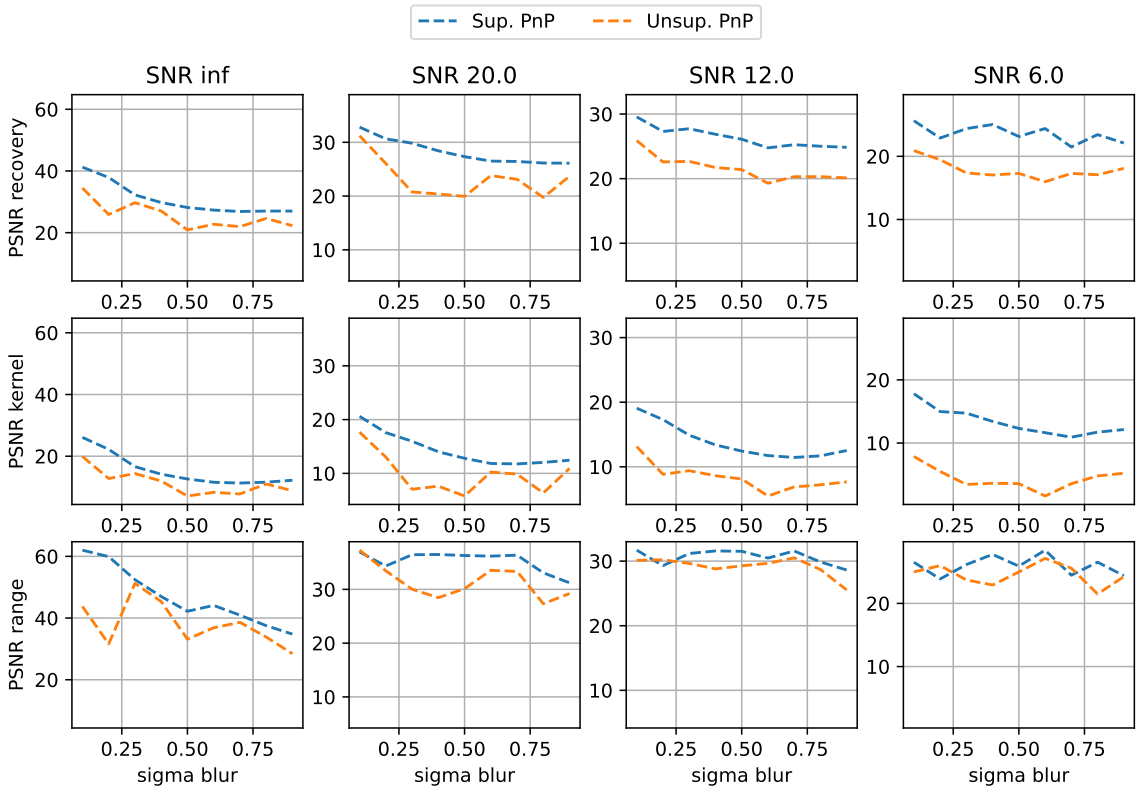
IV. Conclusion

To solve unsupervised inverse problems, it is important to have multiple operators that cover the entire space or appropriate constraints in the model to take advantage of the data's structure and invariance. However, when the operator is too ill-conditioned, such as in deblurring, the prior knowledge must compensate for the lack of information. Relying solely on convolutions is insufficient, even in supervised settings with access to clean data.

Current approaches have limitations, highlighting the need for innovative strategies capable of learning dictionaries that are robust to operator degradation and versatile enough to adapt to a wide range of inverse problems. Such methods would be a significant step forward in our ability to effectively reconstruct signals in the presence of complex, ill-conditioned operators. Developing robust, operator-resistant dictionaries requires a multifaceted approach that integrates advanced machine-learning techniques with deep insights into the problems' structure and physics.



(a) PSNR depending on the size of the blur in the full space, kernel space, and range space of the blur operator for reconstruction methods based on CDL, DIP, TV, and wavelets for a 256 $\times$ 256 grey-level image with several values of SNR. This time, unsupervised prior learning methods fail to recover information in the kernel space. More surprisingly, supervised CDL also struggles in the kernel space.



(b) PSNR depending on the size of the blur in the full space, kernel space, and range space for PnP-based reconstruction on 160 $\times$ 160 grey-level images with several values of SNR. This time, unsupervised prior learning methods fail to recover information in the kernel space. More surprisingly, supervised PnP also struggles in the kernel space.

Fig. 10: PSNR of unsupervised, self-supervised, and supervised methods for deblurring

APPENDIX

A. Proof of Proposition II.1

Let $Z_0 \in \operatorname{argmin}_Z \frac{1}{2}\|AD_0Z - Y\|_2^2 + \lambda\|Z\|_1$. Let $Z'_j = \frac{\|D_{0,j,m}\|}{\|D_{0,j}\|} Z_{0j}$. Then

$$\|AD'Z' - Y\|_2 = \|AD'_0Z'_0 - Y\|_2 \tag{16}$$

$$\|Z'\|_1 \leq \|Z_0\|_1 \tag{17}$$

The result follows.

B. Proof of Proposition II.2

Let $A \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{R}^{m\times T}$. We aim to solve

$$\min_{D\in\mathcal{C},Z} \frac{1}{2}\|ADZ - Y\|_2^2 + \lambda\|Z\|_1 \tag{18}$$

Performing a SVD on A leads to

$$A = U\Lambda V^* \quad \text{such that } U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$$
$$\text{and } UU^* = I_m, \quad VV^* = I_n$$
$$\Lambda = \begin{bmatrix} a_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_m & \cdots & 0 \end{bmatrix}$$

Then,

$$\min_{D\in\mathcal{C},Z} \frac{1}{2}\|ADZ - Y\|_2^2 + \lambda\|Z\|_1 \tag{19}$$

$$= \min_{D\in\mathcal{C},Z} \frac{1}{2}\|U\Lambda V^*DZ - Y\|_2^2 + \lambda\|Z\|_1 \tag{20}$$

$$= \min_{D\in\mathcal{C},Z} \frac{1}{2}\|\Lambda V^*DZ - U^*Y\|_2^2 + \gamma\|Z\|_1 \tag{21}$$

$$= \min_{\tilde{D}\in\mathcal{C},Z,D=V\tilde{D},\tilde{Y}=U^*Y} \frac{1}{2}\left\|\Lambda\tilde{D}Z - \tilde{Y}\right\|_2^2 + \gamma\|Z\|_1 \tag{22}$$

Adding zeros to Λ to make it square, and adding zeros at the end of the measurement vector U^*Y to respect dimensions, the problem reduces to

$$\begin{aligned} &\min_{D\in\mathcal{C},Z} \frac{1}{2}\left\|\Lambda\tilde{D}Z - \tilde{Y}\right\|_2^2 + \lambda\|Z\|_1 \\ &\text{s.t. } \Lambda = \operatorname{diag}(a_1,\cdots,a_m,0,\cdots,0), \tilde{Y} = \begin{pmatrix} U^*Y \\ 0_{n-m} \end{pmatrix} \end{aligned} . \tag{23}$$

Then, Proposition II.1 applies and an optimal dictionary is contained in $\ker(A)^\perp$.

REFERENCES

[1] A. Ribes and F. Schmitt, "Linear inverse problems in imaging," IEEE Signal Processing Magazine, vol. 25, no. 4, pp. 84–99, 2008.

[2] A. Gramfort, M. Kowalski, and M. Hämäläinen, "Mixed-norm estimates for the m/eeg inverse problem using accelerated gradient methods," Physics in medicine and biology, vol. 57, pp. 1937–61, 2012.

[3] J.-L. Starck, "Sparsity and inverse problems in astrophysics," Journal of Physics: Conference Series, vol. 699, 2016.

[4] S. Mallat, A Wavelet Tour of Signal Processing. Academic press, 2008.

[5] M. Elad, Sparse and redundant representations: from theory to applications in signal and image processing. Springer Science & Business Media, 2010.

[6] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" Vision research, vol. 37, no. 23, pp. 3311–3325, 1997.

[7] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Transactions on Signal Processing, vol. 54, pp. 4311 – 4322, 2006.

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," Journal of Machine Learning Research, vol. 11, 2009.

[9] F. Yellin, B. D. Haeffele, and R. Vidal, "Blood cell detection and counting in holographic lens-free imaging by convolutional sparse dictionary learning and coding," in International Symposium on Biomedical Imaging. IEEE, 2017, pp. 650–653.

[10] T. Dupré la Tour, T. Moreau, M. Jas, and A. Gramfort, "Multivariate convolutional sparse coding for electromagnetic brain signals," Advances in Neural Information Processing Systems, vol. 31, pp. 3292–3302, 2018.

[11] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play admm for image restoration: Fixed-point convergence and applications," IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 84–98, 2016.

[12] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (red)," SIAM Journal on Imaging Sciences, vol. 10, no. 4, pp. 1804–1844, 2017.

[13] J. Rick Chang, C.-L. Li, B. Poczos, B. Vijaya Kumar, and A. C. Sankaranarayanan, "One network to solve them all–solving linear inverse problems using deep projection models," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5888–5897.

[14] A. Brifman, Y. Romano, and M. Elad, "Turning a denoiser into a super-resolver using plug and play priors," in 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016, pp. 1404–1408.

[15] B. Malézieux, T. Moreau, and M. Kowalski, "Understanding approximate and unrolled dictionary learning for pattern recovery," in International Conference on Learning Representations, 2022.

[16] B. Tolooshams and D. E. Ba, "Stable and interpretable unrolled dictionary learning," Transactions on Machine Learning Research, 2022.

[17] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," in International Conference on Machine Learning. PMLR, 2018, pp. 2965–2974.

[18] A. Bora, E. Price, and A. G. Dimakis, "Ambientgan: Generative models from lossy measurements," in International conference on learning representations, 2018.

[19] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, "Rare: Image reconstruction using deep priors learned without groundtruth," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 6, pp. 1088–1099, 2020.

[20] D. Chen, J. Tachella, and M. E. Davies, "Equivariant imaging: Learning beyond the range space," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4379–4388.

[21] J. Tachella, D. Chen, and M. Davies, "Sensing theorems for unsupervised learning in linear inverse problems," Journal of Machine Learning Research, vol. 24, no. 39, pp. 1–45, 2023.

[22] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9446–9454.

[23] Z. Szabó, B. Póczos, and A. Lőrincz, "Online group-structured dictionary learning," in CVPR 2011. IEEE, 2011, pp. 2865–2872.

[24] C. Studer and R. G. Baraniuk, "Dictionary learning from sparsely corrupted or compressed signals," in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012, pp. 3341–3344.

[25] V. Naumova and K. Schnass, "Dictionary learning from incomplete data for efficient image restoration," in 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017, pp. 1425–1429.

[26] A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux, "Dictionary learning for massive matrix factorization," in International Conference on Machine Learning. PMLR, 2016, pp. 1737–1746.

[27] ——, "Stochastic subsampling for factorizing huge matrices," IEEE Transactions on Signal Processing, vol. 66, no. 1, pp. 113–128, 2017.

[28] S. Gleichman and Y. C. Eldar, "Blind compressed sensing," IEEE Transactions on Information Theory, vol. 57, no. 10, pp. 6958–6975, 2011.

[29] S. Arora, R. Ge, T. Ma, and A. Moitra, "Simple, efficient, and neural algorithms for sparse coding," in Conference on learning theory. PMLR, 2015, pp. 113–149.

[30] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: dictionary learning with noise and outliers," IEEE Transactions on Information Theory, vol. 61, no. 11, pp. 6298–6319, 2015.

[31] N. Chatterji and P. L. Bartlett, "Alternating minimization for dictionary learning with random initialization," Advances in Neural Information Processing Systems, vol. 30, 2017.

[32] S. Foucart and H. Rauhut, "An invitation to compressive sensing," in A mathematical introduction to compressive sensing. Springer, 2013, pp. 1–39.

[33] D. F. Crouse, "On implementing 2d rectangular assignment algorithms," IEEE Transactions on Aerospace and Electronic Systems, vol. 52, no. 4, pp. 1679–1696, 2016.

[34] F. P. Anaraki and S. M. Hughes, "Compressive k-svd," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 5469–5473.

[35] F. Pourkamali-Anaraki, S. Becker, and S. M. Hughes, "Efficient dictionary learning via very sparse random projections," in 2015 International Conference on Sampling Theory and Applications (SampTA). IEEE, 2015, pp. 478–482.

[36] T. Chang, B. Tolooshams, and D. Ba, "Randnet: Deep learning with compressed measurements of images," in 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2019, pp. 1–6.

[37] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., "Pytorch: An imperative style, high-performance deep learning library," in Advances in neural information processing systems, 2019, pp. 8026–8037.

[39] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-Invariant Sparse Coding for Audio Classification," Cortex, vol. 8, p. 9, 2007.

[40] Y. Hu, M. Delbracio, P. Milanfar, and U. Kamilov, "A restoration network as an implicit prior," in The Twelfth International Conference on Learning Representations, 2024. [Online]. Available: https://openreview.net/forum?id=x7d1qXEn1e

[41] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3142–3155, 2017.

[42] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, "An introduction to total variation for image analysis," Theoretical foundations and numerical methods for sparse recovery, vol. 9, no. 263-340, p. 227, 2010.

[43] J. Scanvic, M. Davies, P. Abry, and J. Tachella, "Self-supervised learning for image super-resolution and deblurring," arXiv preprint arXiv:2312.11232, 2023.