Sparse denoising: aggregation versus global optimization

Diego Carrera, Giacomo Boracchi
Politecnico di Milano, Italy
{diego.carrera, giacomo.boracchi}@polimi.it

Alessandro Foi
Tampere University of Technology, Finland
alessandro.foi@tut.fi

Brendt Wohlberg
Los Alamos National Laboratory, USA
brendt@lanl.gov

Introduction: Denoising is often addressed via sparse coding with respect to an overcomplete translation-invariant dictionary. There are two main approaches for dictionaries composed of translates of an orthonormal basis. The classical approach is *cycle spinning* [1], which aggregates partial estimates, each of which is sparse with respect to a different shift of the orthonormal basis. An alternative is offered by *convolutional sparse representations* [2] [3, Sec. II], which perform a global optimization over the entire dictionary. It is tempting to view the former approach as providing a suboptimal solution of the latter. Here we compare the two approaches and show that, while the global optimization produces estimates with lower bias than the corresponding aggregation procedure, these are also characterized by a higher variance. In practice, the computationally demanding global optimization outperforms the simpler aggregation of partial estimates only when images admit an extremely sparse representation w.r.t. the dictionary, while they perform similarly on natural images.

Denoising: The input signal $\mathbf{s} \in \mathbb{R}^N$ is modeled as $\mathbf{s} = \mathbf{y} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{y} \in \mathbb{R}^N$ denotes the unknown noise-free signal. We consider denoising methods that approximate $\mathbf{y}$ as a sparse linear combination of atoms from an overcomplete translation-invariant dictionary $D \in \mathbb{R}^{N \times N^2}$, formed as the union of all shifted copies D_i, $i \in \{1, \ldots, N\}$, of an orthonormal basis $D_1 \in \mathbb{R}^{N \times N}$, i.e. $\widehat{\mathbf{y}} = D\widehat{\mathbf{x}}$ where $D = (\, D_1 \; \cdots \; D_N \,) \in \mathbb{R}^{N \times N^2}$ and $\widehat{\mathbf{x}} \in \mathbb{R}^{N^2}$.

Cycle spinning [1] separately seeks sparsity separately with respect to each orthonormal basis D_i, by solving the optimization problems

$$\widehat{\mathbf{x}}_i = \underset{\mathbf{u} \in \mathbb{R}^N}{\arg\min} \frac{1}{2} \left\| D_i \mathbf{u} - \mathbf{s} \right\|_2^2 + \lambda \mathcal{R}(\mathbf{u}), \quad i \in \{1, \ldots, N\}, \tag{1}$$

where $\mathcal{R}(\cdot)$ is a regularization term, which is typically $\|\cdot\|_0$ or $\|\cdot\|_1$. The final estimate $\widehat{\mathbf{y}}_{\mathsf{aggr}}$ is obtained by aggregating the N estimates $D_i \widehat{\mathbf{x}}_i$:

$$\widehat{\mathbf{y}}_{\mathsf{aggr}} = \frac{1}{N} \sum_{i=1}^{N} D_i \widehat{\mathbf{x}}_i = D \frac{\left(\widehat{\mathbf{x}}_0^T \; \cdots \; \widehat{\mathbf{x}}_N^T \right)^T}{N} = D \widehat{\mathbf{x}}_{\mathsf{aggr}} \,. \tag{2}$$

An obvious but more computationally expensive alternative defines a single estimate via a global optimization over the entire dictionary:

$$\widehat{\mathbf{x}}_{\mathsf{glob}} = \underset{\mathbf{x} \in \mathbb{R}^{N^2}}{\arg\min} \frac{1}{2} \left\| D\mathbf{x} - \mathbf{s} \right\|_2^2 + \lambda \mathcal{R}(\mathbf{x}) \,. \tag{3}$$

This problem can also formulated in a convolutional form, replacing $D\mathbf{x}$ by convolutions against $M \leq N$ filters:[1]

$$\widehat{\mathbf{x}}_{\mathsf{glob}} = \underset{\mathbf{x} \in \mathbb{R}^{N^2}}{\arg\min} \frac{1}{2} \left\| \sum_{m=1}^{M} \mathbf{d}_m * \mathbf{x}_{[m]} - \mathbf{s} \right\|_2^2 + \lambda \mathcal{R}(\mathbf{x}) \,, \tag{4}$$

where $*$ denotes the convolution operator, $\mathbf{d}_m$ denotes the m^{th} column of D_1, $\mathbf{x}_{[m]} \in \mathbb{R}^N$ is a subvector of $\mathbf{x}$ with $\mathbf{x}_{[m]}(j) = \mathbf{x}(m + (j-1)N)$, $j \in \{1, \ldots, N\}$, and $\mathbf{x}_{[m]} \equiv \mathbf{0}$ for $m > M$. The final estimate is then given by

$$\widehat{\mathbf{y}}_{\mathsf{glob}} = D \widehat{\mathbf{x}}_{\mathsf{glob}} = \sum_{m=1}^{M} \mathbf{d}_m * \widehat{\mathbf{x}}_{[m]} \,, \tag{5}$$

[1] If D_1 contains shifted versions of the same column, e.g. when D_1^T is a wavelet basis, then the number M of filters in (4) can be smaller than N. We rearrange the columns of D_1 so that the first $M \leq N$ columns are all distinct modulo shifts and discard the remaining columns from (4).

where $\widehat{\mathbf{x}}_{[m]}$, typically referred to as coefficient map, is the subvector with the representation coefficients associated with $\mathbf{d}_m$, i.e., $\widehat{\mathbf{x}}_{[m]}(j) = \widehat{\mathbf{x}}_{\mathsf{glob}}(m + (j-1)N)$, $j \in \{1, \ldots, N\}$.

Experiments and Discussion: We consider both $\|\mathbf{x}\|_0$ and $\|\mathbf{x}\|_1$ as choices for the regularization term $\mathcal{R}(\mathbf{x})$. The solution of problem (1) is given by the proximal operator of $\lambda\mathcal{R}(\mathbf{x})$, corresponding to hard- and soft-thresholding [4], [5] for $\|\mathbf{x}\|_0$ and $\|\mathbf{x}\|_1$, respectively. Problem (3) can be approached via a variety of optimization methods. When $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_1$, the problem is convex, and the convolutional form (4) can be efficiently solved in the Fourier domain via an ADMM algorithm [3]. When $\mathcal{R}(\mathbf{x}) = \|x\|_0$, problem (4) can be addressed via the Iterative Hard Thresholding Algorithm [6], which converges to a local minimum since the problem is non-convex.

We take D_1 as the Daubechies db3 wavelet basis with 4 decomposition levels. Since wavelet coefficients from the coarsest level are not sparse [7], one typically shrinks only the detail coefficients [1], [8]. This corresponds to not regularizing the approximation coefficients in (1). This is not a viable solution for the convolutional case, since if we remove the coefficient map $\mathbf{x}_{[1]}$ from $\mathcal{R}(\mathbf{x})$ in (4), then $\widehat{\mathbf{x}}_{[1]}$ is the deconvolution of the noisy $\mathbf{s}$ w.r.t. to $\mathbf{d}_1$. Since this solution leads to a poor estimate $\widehat{\mathbf{y}}_{\mathsf{glob}}$, we perform convolutional sparse coding (4) not on $\mathbf{s}$ but on a high-pass filtered $\mathbf{s}_h$ computed by setting to 0 the approximation coefficients of the overcomplete wavelet transform D^T of $\mathbf{s}$. Hence, we exclude $\mathbf{d}_1$ and $\mathbf{x}_{[1]}$ from the data-fidelity and regularization terms in (4), and add $\mathbf{s} - \mathbf{s}_h$ back to $\widehat{\mathbf{y}}_{\mathsf{glob}}$ in (5).

We compute global and aggregated estimates from natural images corrupted by different amount of noise σ using the ℓ_1 regularization in (1) and (4) and show the PSNR in Figure 1(a). For each σ we separately tune λ in (1) and (4) to achieve the best results for each method and regularization. The two estimates achieve similar performance, with the aggregated estimate slightly outperforming the global one only for small σ. To analyze this result we decompose the mean squared error into squared bias and variance. Figure 1(b) shows that the global optimization outperforms aggregation in terms of bias, but exhibits a larger variance; we speculate that one of the causes is the very high overcompleteness factor of D. When the ℓ_0 regularization is used, the global estimate also achieves a lower bias, but suffers from an even larger variance, see Figure 1(c). Figure 1(a) shows that in this case aggregation outperforms global optimization. It is not clear whether such larger performance gap is due to the inherent superiority of aggregation for functionals involving ℓ_0 regularization, or whether it is due to computational issues: while hard-thresholding provides the closed-form solution of (1), problem (4) is not convex for $\mathcal{R}(\mathbf{x}) = \|\mathbf{x}\|_0$, making it very difficult to find the global minimum.

This effect is not observed when $\mathbf{y}$ is extremely sparse w.r.t. D, as in our synthetic experiments reported in Figure 2. There, when the number of nonzero coefficients of $\mathbf{x}_{\mathsf{glob}}$ is small, the global estimate outperforms the aggregated one in terms of SNR. However, as the number of non-zero coefficients or the noise increases, the variance of the global optimization increases and the two approaches become comparable, due to the larger variance of the global estimate. Further results and discussion can be found in [9].

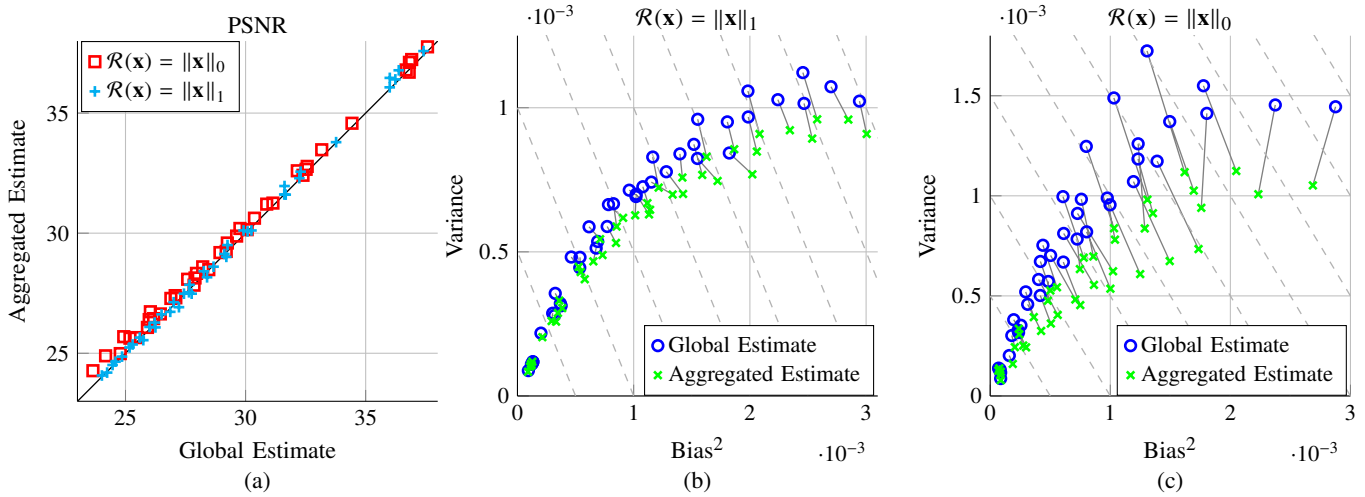

Figure 1. Comparison between the aggregation of partial estimates (1) and global optimization (4) on 5 test images (Lena, Barbara, Man, Peppers, Cameraman), corrupted by different noise levels $\sigma \in \{5, 10, \ldots, 40\}$, according to ℓ_0 and ℓ_1 regularization. The penalty parameter λ is separately tuned for each method and regularization to achieve the best result. Each point in (a) represents the PSNR achieved by $\widehat{\mathbf{y}}_{\text{aggr}}$ (vertical coordinate) and $\widehat{\mathbf{y}}_{\text{glob}}$ (horizontal coordinate) for each image and σ pair. With ℓ_1 regularization (cyan pluses +), the two methods attain similar PSNR values under strong noise, while at low noise levels the aggregation of partial estimates slightly outperforms global optimization. When considering the ℓ_0 regularization (red squares □) the aggregation of partial estimates outperforms the global minimization. In (b) and (c) we decompose the mean squared error obtained by each estimate in its squared bias (horizontal coordinate) and variance (vertical coordinate) components. Thus, anti-diagonals (dashed lines) are contour lines of the mean squared error. Blue circles ○ represent global optimization and green ×-marks the aggregation of partial estimates; markers corresponding to the same noisy image are linked by a segment. The relative position of linked circles and crosses reveals that estimates from global optimization feature a lower bias and higher variance than those from aggregation. This effect can be better appreciated with ℓ_0 regularization (c) than with ℓ_1 regularization (b).

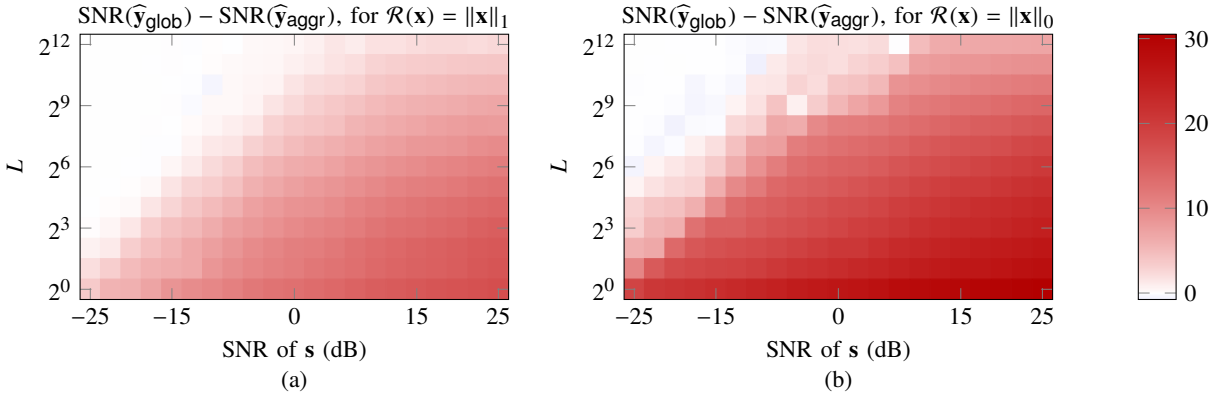

Figure 2. Comparison between the aggregation of partial estimates (1) and global optimization (4) on very sparse synthetic images, according to ℓ_0 and ℓ_1 regularization. Noise-free images of $N = 128 \times 128$ pixels are synthesized as $\mathbf{y} = D\mathbf{x}_{\text{init}}$ where $\mathbf{x}_{\text{init}}$ contains $L \in \{1, 2, 4, 8, \ldots, 4096\}$ nonzero coefficients, which have been randomly selected and set to 1. The variance σ^2 of the additive noise η is set so that the noisy image $\mathbf{s} = \mathbf{y} + \eta$ has SNR in $\{-25, \ldots, 25\}$. In the case of ℓ_0 regularization, we initialize algorithm [6] with the extremely sparse vector $\mathbf{x}_{\text{init}}$ used to generate the image $\mathbf{y}$, since at least when L and σ are small, we expect that the solution of (4) is close to $\mathbf{x}_{\text{init}}$ and that [6] practically approaches global minimum. In (a) and (b) we show the SNR difference between the solutions $\widehat{\mathbf{y}}_{\text{glob}}$ (4) and $\widehat{\mathbf{y}}_{\text{aggr}}$ (1) when using the ℓ_0 and ℓ_1 regularization, respectively. Results are organized according to the SNR of the input noisy image $\mathbf{s}$ (horizontal axis) and the number of nonzero coefficients L (vertical axis). Both plots indicate that the global optimization yields better estimates when $\mathbf{y}$ is very sparse (bottom rows). As L increases, the advantage of the global estimate wanes, particularly at low input SNR values (leftmost region of each plot).

References

[1] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, Eds. Springer, 1995, pp. 125–150.

[2] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Conf. Comp. Vis. Pat. Rec.*, Jun. 2010, pp. 2528–2535.

[3] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.

[4] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B (Met.)*, vol. 58, pp. 267–288, 1996.

[6] M. Kowalski, "Thresholding rules and iterative shrinkage/thresholding algorithm: A convergence study," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 4151–4155.

[7] I. W. Selesnick and M. A. Figueiredo, "Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors," in *SPIE Opt. Eng.+Appl.*, 2009, pp. 74 460D–74 460D.

[8] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Am. Stat. Assoc.*, vol. 90, no. 432, pp. 1200–1224, 1995.

[9] D. Carrera, G. Boracchi, A. Foi, and B. Wohlberg, "Sparse overcomplete denoising: Aggregation versus global optimization," 2017, submitted.