Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
www.gpeyre.com
www.numerical-tours.com

September 24, 2017

Chapter 11

Sparse Regularization

TODO.
Ref [21, 33, 28]

11.1 Sparsity Priors

11.1.1 Ideal sparsity prior.

As detailed in Chapter **??**, it is possible to use an orthogonal basis $\mathcal{B} = \{\psi_m\}_m$ to efficiently approximate an image f in a given class $f \in \Theta$ with a few atoms from $\mathcal{B}$.

To measure the complexity of an approximation with $\mathcal{B}$, we consider the ℓ^0 prior, which counts the number of non-zero coefficients in $\mathcal{B}$

$$J_0(f) = \#\{m\ ;\ \langle f,\, \psi_m\rangle \neq 0\} = \|a\|_0 \quad \text{where} \quad a[m] = \langle f,\, \psi_m\rangle.$$

We have introduced the ℓ^0 pseudo-norm $\|a\|_0$, which we treat here as an ideal sparsity measure for the coefficients a of f in $\mathcal{B}$.

Natural images are not exactly composed of a few atoms, but they can be well approximated by a function f_M with a small ideal sparsity $M = J_0(f)$. In particular, the best M-term approximation defined in (6.3) is defined by

$$f_M = \sum_{|\langle f,\, \psi_m\rangle| > T} \langle f,\, \psi_m\rangle \psi_m \quad \text{where} \quad M = \#\{m\ ;\ |\langle f,\, \psi_m\rangle| > T\}.$$

As detailed in Section 6.2, discontinuous images with bounded variation have a fast decay of the approximation error $\|f - f_M\|$. Natural images f are well approximated by images with a small value of the ideal sparsity prior J_0.

Figure 11.1 shows an examples of decomposition of a natural image in a wavelet basis, $\psi_m = \psi^{\omega}_{j,n}$ $m = (j, n, \omega)$. This shows that most $\langle f,\, \psi_m\rangle$ are small, and hence the decomposition is quite sparse.

11.1.2 Convex relaxation

Unfortunately, the ideal sparsity prior J_0 is difficult to handle numerically because $J_0(f)$ is not a convex function of f. For instance, if f and g have non-intersecting supports of there coefficients in $\mathcal{B}$, then $J_0((f+g)/2) = J_0(f) + J_0(g)$, which shows the highly non-convex behavior of J_0.

This ideal sparsity J_0 is thus not amenable to minimization, which is an issue to solve general inverse problems considered in Section **??**.

We consider a family of ℓ^q priors for $q > 0$, intended to approximate the ideal prior J_0

$$J_q(f) = \sum_m |\langle f,\, \psi_m\rangle|^q.$$

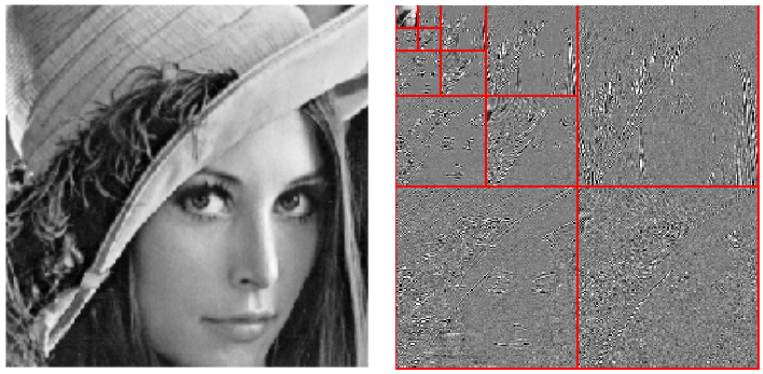

Image f Coefficients $\langle f, \psi_m \rangle$

Figure 11.1: Wavelet coefficients of natural images are relatively sparse.

As shown in Figure 11.2, the unit balls in $\mathbb{R}^2$ associated to these priors are shrinking toward the axes, which corresponds to the unit ball for the ℓ^0 pseudo norm. In some sense, the J_q priors are becoming closer to J_0 as q tends to zero, and thus J_q favors sparsity for small q.

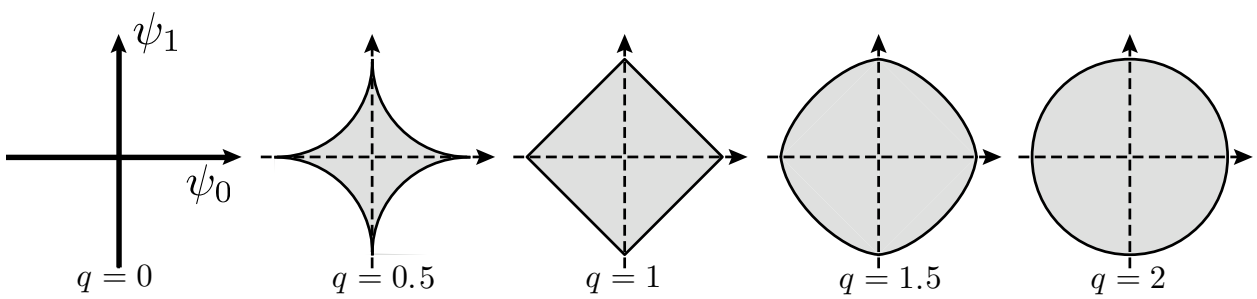

Figure 11.2: ℓ^q balls $\{x \,;\, J_q(x) \leqslant 1\}$ for varying q.

The prior J_q is convex if and only if $q \geqslant 1$. To reach the highest degree of sparsity while using a convex prior, we consider the ℓ^1 sparsity prior J_1, which is thus defined as

$$J_1(f) = \|(\langle f, \psi_m \rangle)\|_1 = \sum_m |\langle f, \psi_m \rangle|. \tag{11.1}$$

In the following, we consider discrete orthogonal bases $\mathcal{B} = \{\psi_m\}_{m=0}^{N-1}$ of $\mathbb{R}^N$.

11.1.3 Sparse Regularization and Thresholding

Given some orthogonal basis $\{\psi_m\}_m$ of $\mathbb{R}^N$, the denoising by regularization (9.15) is written using the sparsity J_0 and J_1 as

$$f^\star_{\lambda,q} = \operatorname*{argmin}_{g \in \mathbb{R}^N} \frac{1}{2} \|f - g\|^2 + \lambda J_q(f)$$

for $q = 0$ or $q = 1$. It can be re-written in the orthogonal basis as

$$f^\star_{\lambda,q} = \sum_m a^\star_{\lambda,q}[m] \psi_m$$

$$a^\star_{\lambda,q} = \operatorname*{argmin}_{b\in\mathbb{R}^N} \sum_m \frac{1}{2}|a[m]-b[m]|^2 + \lambda\varphi_q(b[m])$$

where $a[m] = \langle f,\, \psi_m\rangle$, and with

$$\varphi_1(x) = |x| \quad \text{and} \quad \varphi_0(x) = \begin{cases} 0 & \text{if} \quad x=0, \\ 1 & \text{otherwise.} \end{cases}$$

Each coefficients of the denoised image is the solution of

$$a^\star_{\lambda,q}[m] = \operatorname*{argmin}_{\alpha\in\mathbb{R}} \frac{1}{2}|a[m]-\beta|^2 + \lambda\varphi_q(\beta)$$

and one can shows that this optimization is solved exactly in closed form using thresholding

$$a^\star_{\lambda,q}[m] = S^q_T(a[m]) \quad \text{where} \quad \begin{cases} T=\sqrt{2\lambda} & \text{for} \quad q=0, \\ T=\lambda & \text{for} \quad q=1, \end{cases} \tag{11.2}$$

where S^0_T is the hard thresholding introduced in (8.3), and S^1_T is the soft thresholding introduced in (8.4).

One thus has

$$f_{\lambda,q} = \sum_m S^q_T(\langle f,\, \psi_m\rangle)\psi_m.$$

As detailed in Section 8.3, these denoising methods has the advantage that the threshold is simple to set for Gaussian white noise w of variance σ^2. Theoritical values indicated that $T=\sqrt{2\log(N)}\sigma$ is asymptotically optimal, see Section 8.3.3. In practice, one should choose $T\approx 3\sigma$ for hard thresholding (ℓ^0 regularization), and $T\approx 3\sigma/2$ for soft thresholding (ℓ^1 regularization), see Figure 8.14.

11.2 Sparse Regularization of Inverse Problems

Sparse ℓ^1 regularization in an orthogonal basis $\{\psi_m\}_m$ of $\mathbb{R}^N$ makes use of the J_1 prior defined in (11.1), so that the inversion is obtained by solving the following convex program

$$f^\star \in \operatorname*{argmin}_{f\in\mathbb{R}^N} \frac{1}{2}\|y-\Phi f\|^2 + \lambda\sum_m |\langle f,\, \psi_m\rangle|. \tag{11.3}$$

This corresponds to the basis pursuit denoising for sparse approximation introduced by Chen, Donoho and Saunders in [6]. The resolution of (11.3) can be perform using an iterative thresholding algorithm as detailed in Section 11.3.

For noiseless measurements $y=\Phi f_0$, one solves a constraint basis pursuit problem

$$f^\star \in \operatorname*{argmin}_{\Phi f=y} \sum_m |\langle f,\, \psi_m\rangle|.$$

This can be recasted as a convex linear program, which can in turn by solved by various solver such as simplex, interior points, or Douglas-Rachford iterations.

11.3 Proximal Gradient Algorithm

This section details an iterative algorithm that compute a solution of (10.6) for either the TV prior $J=J_{\mathrm{TV}}$ or the sparse ℓ^1 prior, which corresponds respectively to the minimizations (10.11) and (11.3).

This algorithm was derived by several authors, among which [14, 10], and belongs to the general family of forward-backward splitting in proximal iterations [8]. We note that faster algorithms can be used, such as Nesterov scheme [23].

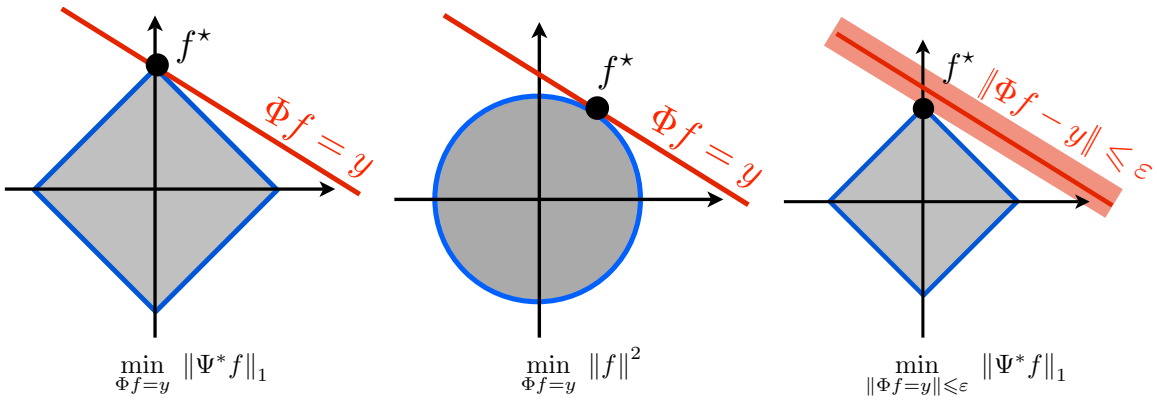

Figure 11.3: Geometry of convex optimizations.

Surrogate functionals. The energy to minimize in (10.11) and (11.3) is written as

$$E(f) = \frac{1}{2} \|\Phi f - y\|^2 + \lambda J(f).$$

The difficulty is the presence of the operator Φ in the ℓ^2 norm, which makes this problem significantly more difficult than the simple denoising by regularization (9.15).

To derive an iterative algorithm, we modify the energy $E(f)$ to obtain a surrogate functional $E(f, f^{(k)})$ whose minimization corresponds to a simpler denoising problem.

Given some guess $f^{(k)} \in \mathbb{R}^N$ of the solution $f^\star$, the surrogate functional is defined as

$$E(f, f^{(k)}) = E(f) - \frac{1}{2} \left\| \Phi f - \Phi f^{(k)} \right\|^2 + \frac{1}{2\tau} \left\| f - f^{(k)} \right\|^2.$$

One has

$$E(f, f^{(k)}) \geqslant E(f) \quad \text{and} \quad E(f^{(k)}, f^{(k)}) = E(f^{(k)}) \tag{11.4}$$

so that $E(f, f^{(k)})$ is a proxy for the minimization of $E(f)$.

Proximal iterations. A proximal iterative algorithm computes

$$f^{(k+1)} = \operatorname*{argmin}_{f \in \mathbb{R}^N} E(f, f^{(k)}).$$

Property (11.4) guarantees that E is decaying

$$E(f^{(k+1)}) \leqslant E(f^{(k)}).$$

Furtheremore, one has

$$E(f, f^{(k)}) = C + \frac{1}{2\tau} \left\| f - f^{(k)} + \tau \Phi^* (\Phi f^{(k)} - y) \right\|^2 + \lambda \sum_m |\langle f, \psi_m \rangle|$$

where C is independent of f. Defining a proximal operator

$$\operatorname{prox}_{\lambda J}(\tilde{f}) = \operatorname*{argmin}_{f \in \mathbb{R}^N} \frac{1}{2} \left\| \tilde{f} - f \right\|^2 + \lambda J(f), \tag{11.5}$$

that corresponds to the variational denoiser introduced in (9.15), one thus has

$$f^{(k+1)} = \operatorname{prox}_{\lambda \tau J}\left(\tilde{f}^{(k)} \right) \quad \text{where} \quad \tilde{f}^{(k)} = f^{(k)} - \tau \Phi^* (\Phi f^{(k)} - y).$$

One can prove, see [], that if $\tau < 2 / \|\Phi^* \Phi\|_S$, then

$$f^{(k)} \to f^\star.$$

Noiseless case. If $\sigma = 0$, so that one observe noiseless measures $y = \Phi f_0$, an heuristic to compute approximately the solution $f^\star$ of (10.7) is to use a decaying value of $\lambda = \lambda^{(k)}$ during the iterations. One can for instance use $\lambda_k = \lambda_{\max}/k$, although there is no proof of convergence to $f^\star$.

constrained problem. The constrained problem

$$f_\varepsilon^\star \in \operatorname*{argmin}_{\|\Phi f - y\| \leqslant \varepsilon} \sum_m |\langle f, \psi_m \rangle|.$$

is equivalent to the problem (10.6), in the sense that $f^\star$ is a solution of (10.6) for a suitable value of λ that depends on ε. Unfortunately, the correspondence between λ and ε is unknown and depends on y.

An heuristic to automatically find this correspondence is to iteratively update the value of $\lambda = \lambda^{(k)}$

$$\lambda^{(k+1)} = \lambda^{(k)} \frac{\varepsilon}{\left\|\Phi f^{(k)} - y\right\|}.$$

Sparse regularization. For the case of $J = J_1$, the proximal denoising operator (11.5) is computed in closed form using a soft thresholding, as already noticed in (11.2).

The resulting proximal iterative algorithm corresponds to the iterative soft thresholding algorithm, that alternates a gradient descent step

$$\tilde{f}^{(k)} = f^{(k)} - \tau \Phi^* (\Phi f^{(k)} - y). \tag{11.6}$$

and soft thresholding

$$f^{(k+1)} = \sum_m S_{\lambda\tau}^1 (\langle \tilde{f}^{(k)}, \psi_m \rangle) \psi_m. \tag{11.7}$$

Table **??** details the implementation of this method, when the data is assumed to be sparse in the trivial identity basis.

TV regularization. For the case of $J = J_{\mathrm{TV}}$, the proximal operator (11.5) does not have a closed form solution. One thus has to use inner iteration of Chambolle's scheme (9.21) to compute the proximal map.

11.4 Example: Sparse Deconvolution

11.4.1 Sparse Spikes Deconvolution

Sparse spikes deconvolution makes use of sparsity in the spacial domain, which corresponds to the orthogonal basis of Diracs $\psi_m[n] = \delta[n - m]$. This sparsity was first introduced in the seismic imaging community [], where the signal f_0 represent the change of density in the underground and is assumed to be composed of a few Diracs impulse.

In a simplified linearized 1D set-up, ignoring multiple reflexions, the acquisition of underground data f_0 is modeled as a convolution $y = h \star f_0 + w$, where h is a so-called "wavelet" signal sent in the ground. This should not be confounded with the construction of orthogonal wavelet bases detailed in Chapter **??**, although the term "wavelet" originally comes from seismic imaging.

The wavelet filter h is typically a band pass signal that perform a tradeoff between space and frequency concentration especially tailored for seismic exploration. Figure (11.4) shows a typical wavelet that is a second derivative of a Gaussian, together with its Fourier transform. This shows the large amount of information removed from f during the imaging process.

The sparse ℓ^1 regularization in the Dirac basis reads

$$f^\star = \operatorname*{argmin}_{f \in \mathbb{R}^N} \frac{1}{2} \|f \star h - y\|^2 + \lambda \sum_m |f[m]|.$$

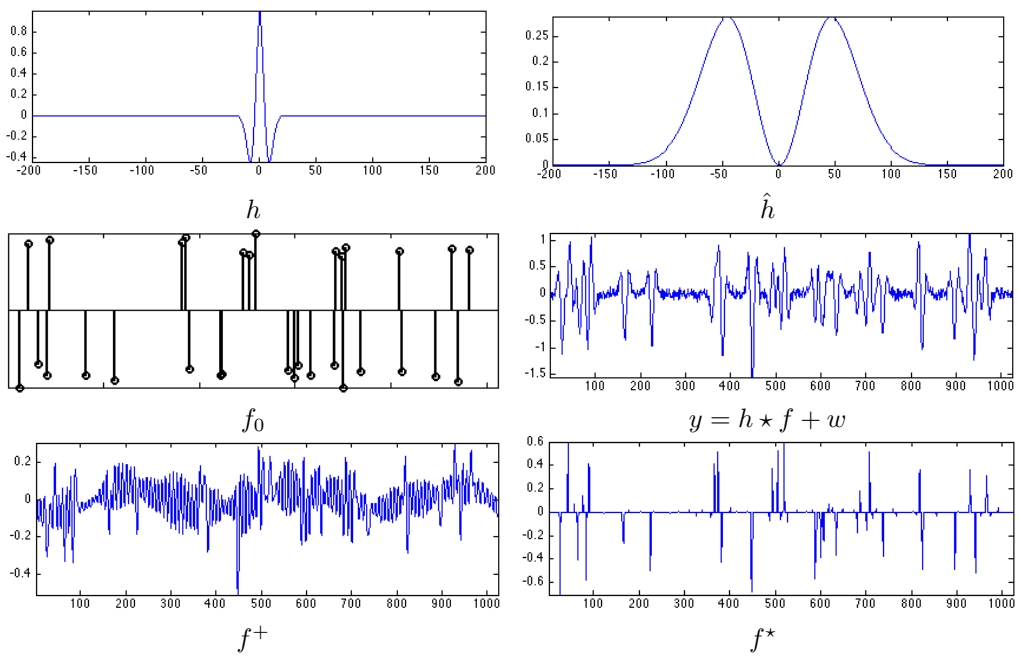

Figure 11.4: Pseudo-inverse and ℓ^1 sparse spikes deconvolution.

Figure 11.4 shows the result of ℓ^1 minimization for a well chosen λ parameter, that was optimized in an oracle manner to minimize the error $\|f^\star - f_0\|$.

The iterative soft thresholding for sparse spikes inversion iterates

$$\tilde{f}^{(k)} = f^{(k)} - \tau h \star (h \star f^{(k)} - y)$$

and

$$f^{(k+1)}[m] = S^1_{\lambda\tau}(\tilde{f}^{(k)}[m])$$

where the step size should obeys

$$\tau < 2/\|\Phi^*\Phi\| = 2/\max_\omega |\hat{h}(\omega)|^2$$

to guarantee convergence. Figure 11.5 shows the progressive convergence of the algorithm, both in term of energy minimization and iterates. Since the energy is not strictly convex, we note that convergence in energy is not enough to guarantee convergence of the algorithm.

11.4.2 Sparse Wavelets Deconvolution

Signal and image acquired by camera always contain some amount of blur because of objects being out of focus, movements in the scene during exposure, and diffraction. A simplifying assumption assumes a spatially invariant blur, so that Φ is a convolution

$$y = f_0 \star h + w.$$

In the following, we consider h to be a Gaussian filter of width $\mu > 0$. The number of effective measurements can thus be considered to be $P \sim 1/\mu$, since Φ nearly set to 0 large enough Fourier frequencies. Table **??** details the implementation of the sparse deconvolution algorithm.

Figures 11.6 and 11.7 shows examples of signal and image acquisition with Gaussian blur.

Sobolev regularization (10.13) improves over ℓ^2 regularization (10.12) because it introduces an uniform smoothing that reduces noise artifact. It however fail to recover sharp edge and thus does a poor job in

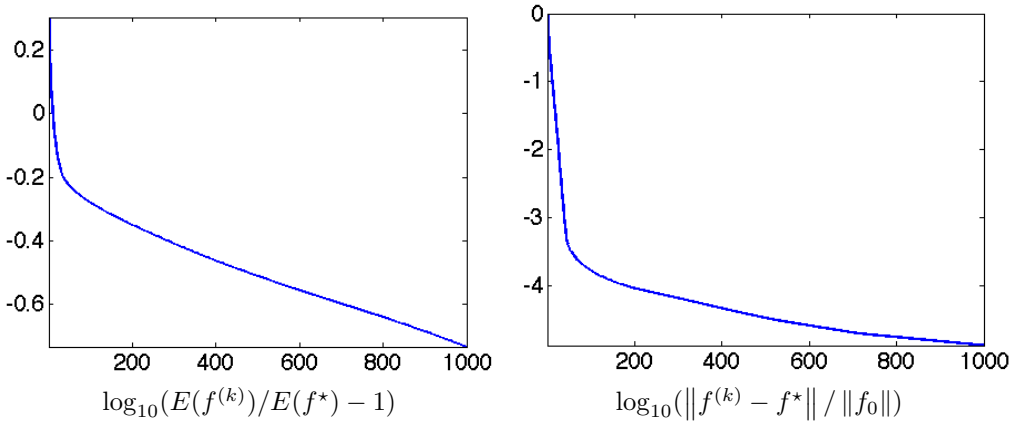

Figure 11.5: Decay of the energy and convergence through the iterative thresholding iterations.

inverting the operator. To recover sharper transition and edges, one can use either a TV regularization or a sparsity in an orthogonal wavelet basis.

Figure 11.6 shows the improvement obtained in 1D with wavelets with respect to Sobolev. Figure 11.7 shows that this improvement is also visible for image deblurring. To obtain a better result with fewer artifact, one can replace the soft thresholding in orthogonal wavelets in during the iteration (11.7) by a thresholding in a translation invariant tight frame as defined in (8.6).

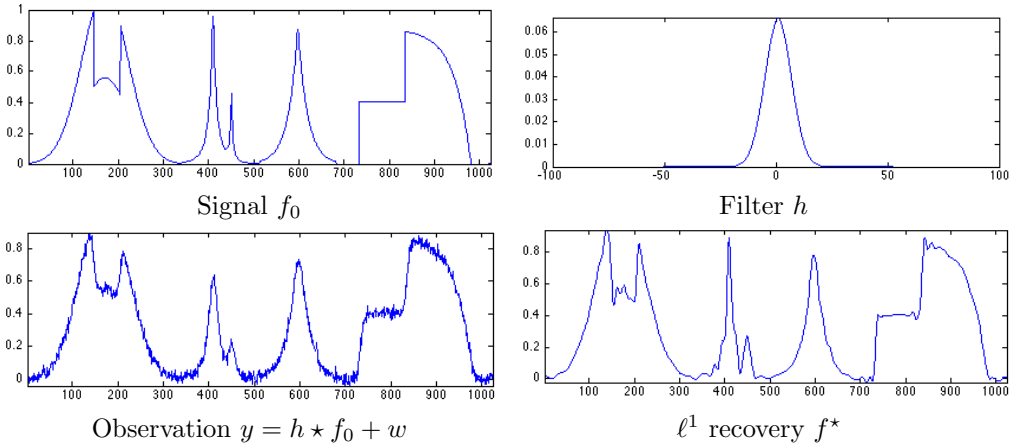

Figure 11.6: Sparse 1D deconvolution using orthogonal wavelets.

Figure 11.8 shows the decay of the SNR as a function of the regularization parameter λ. This SNR is computed in an oracle manner since it requires the knowledge of f_0. The optimal value of λ was used in the reported experiments.

11.4.3 Sparse Inpainting

This section is a follow-up of Section 10.4.2.

To inpaint using a sparsity prior without noise, we use a small value for λ. The iterative thresholding

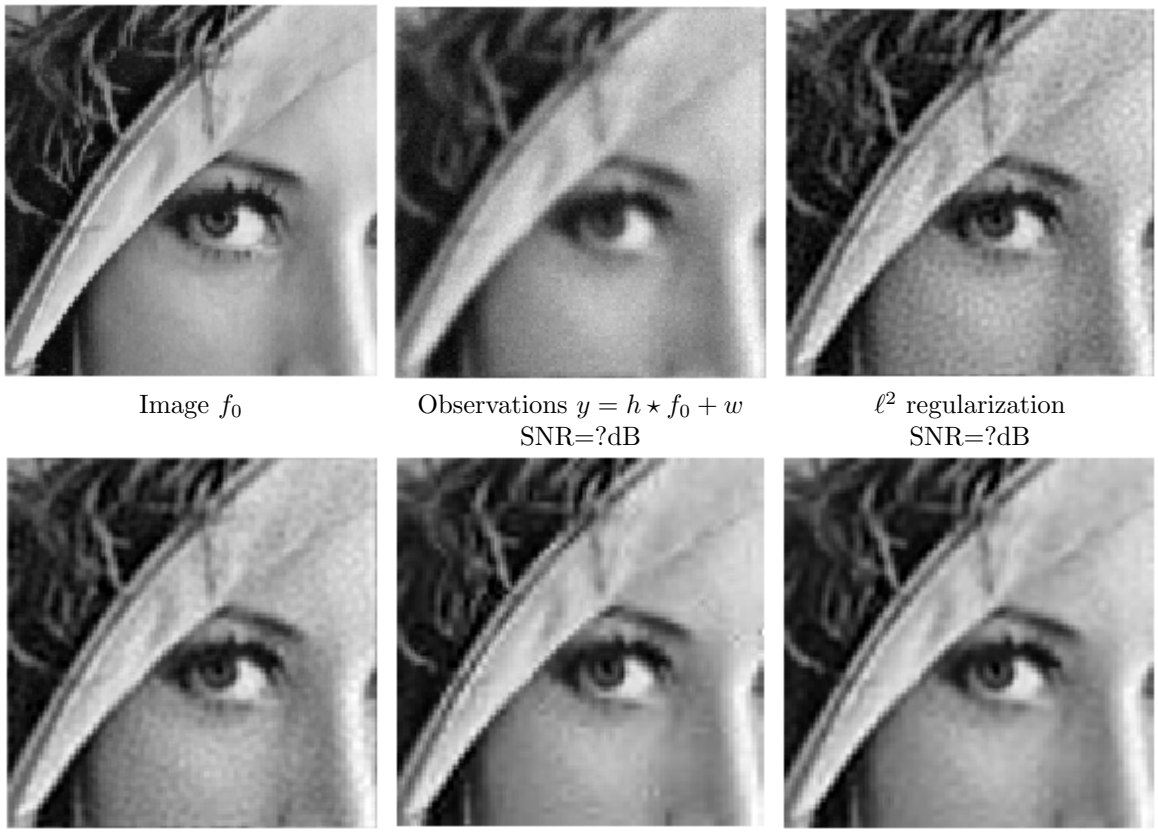

Image f_0 | Observations $y = h \star f_0 + w$ SNR=?dB | ℓ^2 regularization SNR=?dB

Sobolev regularization SNR=?dB | ℓ^1 regularization SNR=?dB | ℓ^1 invariant SNR=?dB

Figure 11.7: Image deconvolution.

algorithm (11.7) is written as follow for $\tau = 1$,

$$f^{(k+1)} = \sum_m S_\lambda^1(\langle P_y(f^{(k)}),\, \psi_m\rangle)\psi_m$$

Figure 11.9 shows the improvevement obtained by the sparse prior over the Sobolev prior if one uses soft thresholding in a translation invariant wavelet frame.

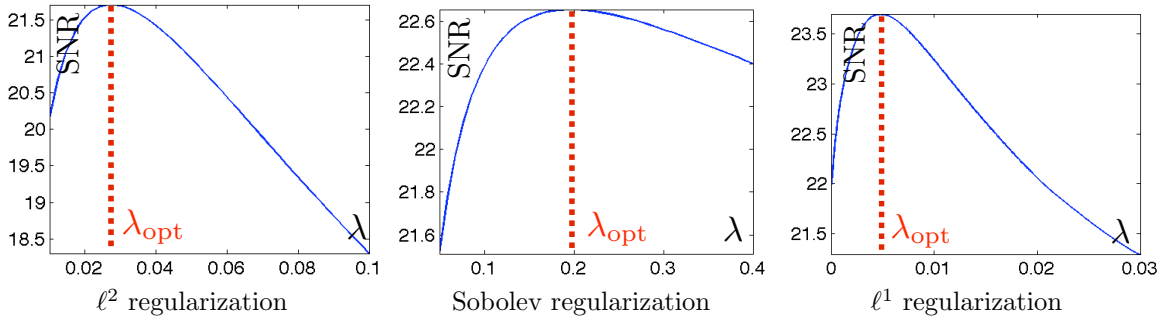

Figure 11.8: SNR as a function of λ.

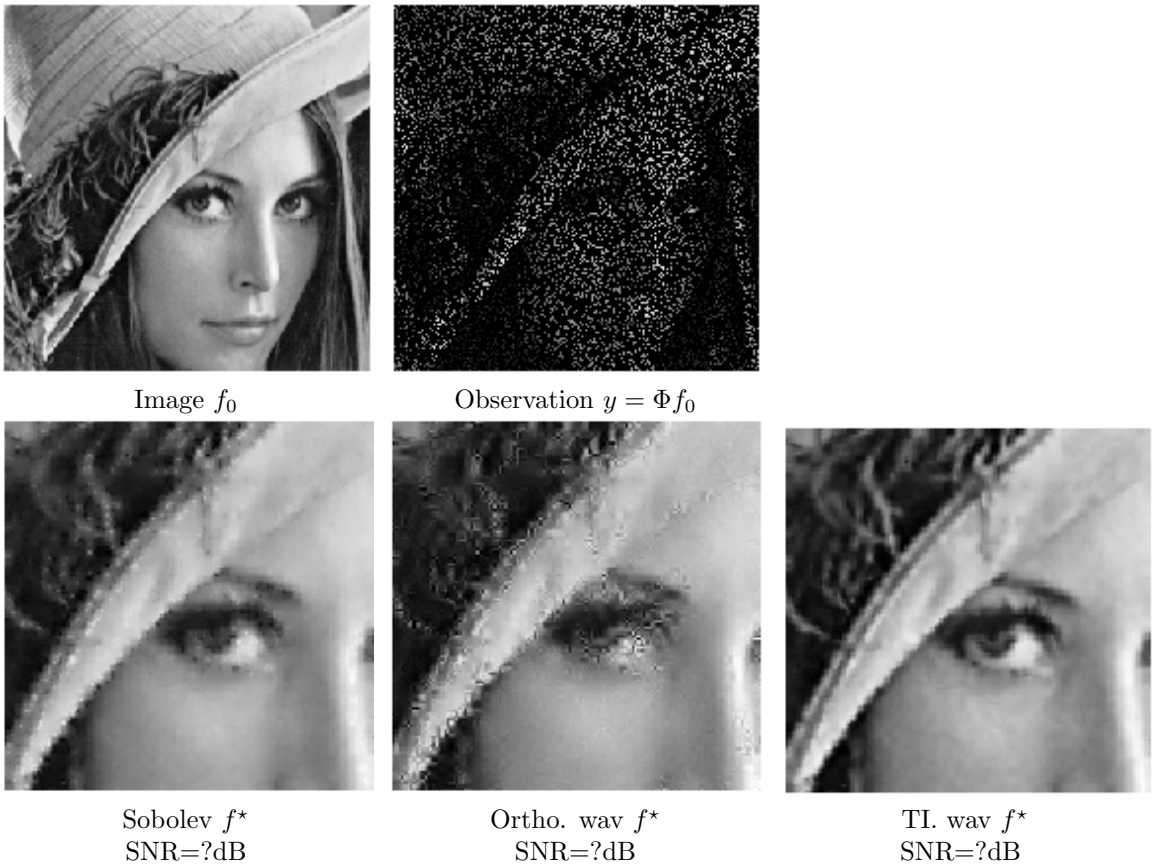

Figure 11.9: Inpainting with Sobolev and sparsity.

Bibliography

[1] P. Alliez and C. Gotsman. Recent advances in compression of 3d meshes. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 3–26. Springer Verlag, 2005.

[2] P. Alliez, G. Ucelli, C. Gotsman, and M. Attene. Recent advances in remeshing of surfaces. In *AIM@SHAPE repport*. 2005.

[3] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.

[4] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.

[5] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.

[6] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[7] F. R. K. Chung. Spectral graph theory. *Regional Conference Series in Mathematics, American Mathematical Society*, 92:1–212, 1997.

[8] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.

[9] P. Schroeder et al. D. Zorin. Subdivision surfaces in character animation. In *Course notes at SIGGRAPH 2000*, July 2000.

[10] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.

[11] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267, 1998.

[12] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.

[13] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[14] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.

[15] M. S. Floater and K. Hormann. Surface parameterization: a tutorial and survey. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 157–186. Springer Verlag, 2005.

[16] I. Guskov, W. Sweldens, and P. Schröder. Multiresolution signal processing for meshes. In Alyn Rockwood, editor, *Proceedings of the Conference on Computer Graphics (Siggraph99)*, pages 325–334. ACM Press, August8–13 1999.

[17] A. Khodakovsky, P. Schröder, and W. Sweldens. Progressive geometry compression. In *Proceedings of the Computer Graphics Conference 2000 (SIGGRAPH-00)*, pages 271–278, New York, July 23–28 2000. ACMPress.

[18] L. Kobbelt. $\sqrt{3}$ subdivision. In Sheila Hoffmeyer, editor, *Proc. of SIGGRAPH'00*, pages 103–112, New York, July 23–28 2000. ACMPress.

[19] M. Lounsbery, T. D. DeRose, and J. Warren. Multiresolution analysis for surfaces of arbitrary topological type. *ACM Trans. Graph.*, 16(1):34–73, 1997.

[20] S. Mallat. *A Wavelet Tour of Signal Processing, 3rd edition.* Academic Press, San Diego, 2009.

[21] Stephane Mallat. *A wavelet tour of signal processing: the sparse way.* Academic press, 2008.

[22] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.

[23] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152, 2005.

[24] Gabriel Peyré. *L'algèbre discrète de la transformée de Fourier.* Ellipses, 2004.

[25] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.

[26] E. Praun and H. Hoppe. Spherical parametrization and remeshing. *ACM Transactions on Graphics*, 22(3):340–349, July 2003.

[27] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[28] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging.* Springer, 2009.

[29] P. Schröder and W. Sweldens. Spherical Wavelets: Efficiently Representing Functions on the Sphere. In *Proc. of SIGGRAPH 95*, pages 161–172, 1995.

[30] P. Schröder and W. Sweldens. Spherical wavelets: Texture processing. In P. Hanrahan and W. Purgathofer, editors, *Rendering Techniques '95.* Springer Verlag, Wien, New York, August 1995.

[31] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[32] A. Sheffer, E. Praun, and K. Rose. Mesh parameterization methods and their applications. *Found. Trends. Comput. Graph. Vis.*, 2(2):105–171, 2006.

[33] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis.* Cambridge university press, 2015.

[34] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Applied and Computation Harmonic Analysis*, 3(2):186–200, 1996.

[35] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.