Reconciling ”priors” and ”priors” without prejudice?

Rémi Gribonval

Reconciling "priors" and "priors" without prejudice ?

Remi Gribonval*
* Inria Rennes – Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France.

***Abstract*—We discuss a long-lasting *qui pro quo* between regularization-based and Bayesian-based approaches to inverse problems, and review some recent results that try to reconcile both viewpoints. This sheds light on some tradeoff between computational efficiency and estimation accuracy in sparse regularization.**

I. Introduction

A central problem in modern signal processing is to solve inverse problems of the type $y = \mathcal{A}(x) + n$ where $\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^m$ $(m \leq N)$ models a linear or nonlinear observation process, and n accounts for additive noise in this process. Addressing such problems amounts to designing estimators $\Delta : \mathbb{R}^m \to \mathbb{R}^N$, also called decoders.

The last decade has witnessed a particularly impressive amount of work dedicated to linear dimensionality reducing observations processes ($m \ll N$), where $\mathcal{A}(x) = \mathbf{A}x$, with $\mathbf{A} \in \mathbb{R}^{m \times N}$. Many *sparse* decoders (greedy algorithms, iterative reweighed or thresholding schemes) have been carefully designed and their performance guarantees have been scrutinized on various types of signals.

Regularization: A particular family of decoders is associated to *regularization* through global optimization of a cost function

$$\Delta_\phi(y) := \arg\min_x \frac{1}{2}\|y - \mathcal{A}(x)\|_2^2 + \phi(x) \qquad (1)$$

where $\phi : \mathbb{R}^N \to \mathbb{R}$ is a penalty function. The ℓ^1 decoder associated to $\phi(x) = \lambda\|x\|_1$ has attracted a particular attention, for the associated optimization problem is convex, and generalizations to other "mixed" norms are being intensively studied. Several facts explain the popularity of such approaches: a) these penalties have well-understood geometric interpretations; b) they are known to be sparsity promoting (the minimizer has many zeroes); c) this can be exploited in active set methods for computational efficiency; d) convexity offers a comfortable framework to ensure both a unique minimum and a rich toolbox of efficient and provably convergent optimization algorithms.

Bayesian modeling: While the convex and deterministic viewpoint on inverse problems has gained a strong momentum, there is another major route: the Bayesian one. Assuming prior distributions $x \sim P_X$, $n \sim P_N$ on the unknown and the noise, and measuring the risk with the squared loss $\|\Delta(\mathcal{A}(x)) - x\|_2^2$, the optimum decoder, in the sense of the minimum expected risk is the conditional mean, also known as posterior mean or minimum mean squared error (MMSE),

$$\Delta^\star(y) := \mathbb{E}(x|y) = \int_{\mathbb{R}^N} x\ p(x|y)dx. \qquad (2)$$

Its computation involves high-dimensional integration, which raises substantial issues typically addressed through sampling (MCMC, etc.).

II. Reconciling two worlds?

Regularization and Bayesian estimation seemingly yield radically different viewpoints on inverse problems. In fact, they are underpinned by distinct ways of defining signal models or "priors". The "regularization prior" is embodied by the penalty function $\phi(x)$ which promotes certain solutions, somehow carving an implicit signal model. In the Bayesian framework, the "Bayesian prior" is embodied by where the mass of the signal distribution P_X lies.

The MAP qui pro quo*:* A *qui pro quo* been these distinct notions of priors has crystalized around the notion of *maximum a posteriori* (MAP) estimation, leading to a long lasting incomprehension between two worlds. In fact, a simple application of Bayes rule shows that under a Gaussian noise model $n \sim \mathcal{N}(0, \mathbf{I}_m)$ and *Bayesian prior* $P_X(x \in E) = \int_E p_X(x)dx$, $E \subset \mathbb{R}^N$, MAP estimation yields the optimization problem (1) with *regularization prior* $\phi_X(x) := -\log p_X(x)$. As an unfortunate consequence of an erroneous "reverse reading" of this fact, the optimization problem (1) with regularization prior $\phi(x)$ is now routinely called "MAP with prior $\exp(-\phi(x))$". With the ℓ^1 penalty, it is often called "MAP with a Laplacian prior". This *qui pro quo* has given rise to the erroneous but common myth that the optimization approach is particularly well adapted when the unknown is distributed as $\exp(-\phi(x))$.

A myth disproved: As a striking counter-example to this last myth, it has recently been proved [1] that when x is drawn i.i.d. Laplacian and $\mathbf{A} \in \mathbb{R}^{m \times N}$ is drawn from the Gaussian ensemble, the ℓ^1 decoder – and indeed *any* sparse decoder – will be outperformed by the least squares decoder $\Delta(y) = \mathbf{A}^+ y$, unless $m \gtrsim 0.15N$.

Reconciliation?: Can these routes be reconciled ? In the context of additive white Gaussian noise denoising ($m = N$, $\mathbf{A} = \mathbf{I}_m$, $n \sim \mathcal{N}(0, \mathbf{I}_m)$) it has been shown [2] that the truly Bayesian estimator $\Delta^\star(y)$ is in fact the solution of an optimization problem (1), for some *regularization prior* $\phi^\star$ fully determined by the *Bayesian prior* P_X:

$$\Delta^\star(y) = \arg\min_x \frac{1}{2}\|y - x\|_2^2 + \phi^\star(x). \qquad (3)$$

Moreover, for any $y \in \mathbb{R}^m$, the global minimum $\Delta^\star(y)$ is indeed the *unique stationary point* of the resulting optimization problem (3). In other words, for AWGN denoising, Bayesian estimation with any postulated Bayesian prior P_X can be expressed as a regularization with a certain regularization prior $\phi^\star$.

Is the reverse true ? The results in [2] show that the resulting regularization prior $\phi^\star$ is necessarily smooth everywhere. Hence, many popular sparsity-promoting regularization priors cannot correspond to any Bayesian prior. In particular, the ℓ^1 penalty cannot be the MMSE estimator for any Bayesian prior P_X. In other words, the performance of any sparse-regularization scheme is necessarily sub-optimal. The talk will discuss consequences of these results in terms of tradeoffs between computational complexity and estimation performance, as well as possible extensions to under-determined linear or nonlinear problems.

References

[1] R. Gribonval, V. Cevher, and M. Davies, E., "Compressible Distributions for High-dimensional Statistics," *IEEE Transactions on Information Theory*, 2012, was previously entitled "Compressible priors for high-dimensional statistics". [Online]. Available: http://hal.inria.fr/inria-00563207

[2] R. Gribonval, "Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?" *Signal Processing, IEEE Transactions on*, vol. 59, no. 5, pp. 2405–2410, 2011. [Online]. Available: http://hal.archives-ouvertes.fr/inria-00486840/PDF/MAPvsOptim.pdf