Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?

Rémi Gribonval

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?

Rémi Gribonval

N° 7484

Mai 2010

Audio, Speech, and Language Processing

ISSN 0249-6399 ISRN INRIA/RR--7484--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
RENNES - BRETAGNE ATLANTIQUE

Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?

Rémi Gribonval *

Theme : Audio, Speech, and Language Processing
Équipes-Projets METISS

Rapport de recherche n° 7484 — Mai 2010 — 14 pages

Abstract: Penalized least squares regression is often used for signal denoising and inverse problems, and is commonly interpreted in a Bayesian framework as a Maximum A Posteriori (MAP) estimator, the penalty function being the negative logarithm of the prior. For example, the widely used quadratic program (with an ℓ^1 penalty) associated to the LASSO / Basis Pursuit Denoising is very often considered as MAP estimation under a Laplacian prior in the context of additive white Gaussian noise (AWGN) reduction. This paper highlights the fact that, while this is *one* possible Bayesian interpretation, there can be other equally acceptable Bayesian interpretations. Therefore, solving a penalized least squares regression problem with penalty $\phi(x)$ need not be interpreted as assuming a prior $C \cdot \exp(-\phi(x))$ and using the MAP estimator. In particular, it is shown that for *any* prior P_X, the minimum mean square error (MMSE) estimator is the solution of a penalized least square problem with some penalty $\phi(x)$, which can be interpreted as the MAP estimator with the prior $C \cdot \exp(-\phi(x))$. Vice-versa, for *certain* penalties $\phi(x)$, the solution of the penalized least squares problem is indeed the MMSE estimator, with a certain prior P_X. In general $dP_X(x) \neq C \cdot \exp(-\phi(x))dx$.

Key-words: Bayesian estimation; Maximum A Posteriori; Minimum Mean Square Error; MAP; MMSE; penalized least squares regression; LASSO; Basis Pursuit; nonconvex optimization; proximity operator

* Rémi Gribonval is with INRIA, Centre Inria Rennes - Bretagne Atlantique, Campus de Beaulieu, F-35042 Rennes Cedex, Rennes, France. Phone: +33 2 99 84 25 06. Fax: +33 2 99 84 71 71. Email: remi.gribonval@inria.fr.
Rémi Gribonval is a member of the METISS project-team at IRISA, Rennes, France. This work was supported in part by the European Union through the project SMALL (Sparse Models, Algorithms and Learning for Large-Scale data). The project SMALL acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 225913.

Faut-il interpréter la régression aux moindres carrés pénalisée comme de l'estimation au Maximum A Posteriori ?

Résumé : La régression aux moindres carrés pénalisée est souvent utilisée dans le cadre du débruitage de signaux et des problèmes inverses, où elle est communément interprétée comme un estimateur au Maximum A Posteriori (MAP), la fonction de pénalité étant alors liée au logarithme de la probabilité *a priori*. Par exemple, le programme quadratique (avec pénalité ℓ^1) associé au *Basis Pursuit Denoising* / LASSO, qui est très largement utilisé en traitement du signal et de l'image, est souvent considéré comme un estimateur MAP avec a priori Laplacien dans le contexte d'un bruit additif blanc et Gaussien. Cet article met en lumière le fait que cette interprétation Bayesienne, bien que possible, n'est pas unique ni nécessairement fondée. Ainsi, la résolution d'un problème de régression aux moindres carrés avec un pénalité $\phi(x)$ n'a pas vocation à être systématiquement interprétée comme un estimateur MAP avec probabilité a priori $C \cdot \exp(-\phi(x))$. En particulier, on montre que pour *toute* loi a priori P_X, l'estimateur d' erreur quadratique moyenne minimale (EQMM) est la solution d'un problème de régression aux moindres carrés pénalisé avec une fonction de pénalité bien choisie $\phi(x)$, et peut dont être également interprété comme l'estimateur MAP avec a priori $C \cdot \exp(-\phi(x))$. Vice-versa, pour *certaines* pénalités $\phi(x)$, la solution du problème aux moindres carrés pénalisé est de fait également l'estimateur EQMM, avec une loi a priori P_X bien choisie. En général, $dP_X(x) \neq C \cdot \exp(-\phi(x))dx$.

Mots-clés : Estimation Bayesienne; Maximum A Posteriori; Erreur Quadratique Moyenne Minimala; MAP; EQMM; régression aux moindres pénalisée; LASSO; Basis Pursuit; optimisation non-convexe; opérateur proximal

1 Introduction

Consider the problem of estimating an unknown signal $x \in \mathbb{R}^n$ from a noisy observation $y = x + b$, also known as *denoising*. Given an arbitrary noisy observation y the goal is to estimate the noiseless signal x: in practice, designing a denoising scheme amounts to choosing a function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ which provides estimates of the form $\hat{x} = \psi(y)$. However, unless one specifies further what is meant by "noise" and "signal", denoising is a completely ill-posed problem since any pair x, b such that $y = x+b$ can be replaced by a pair x', b' where $x' = x+z$, $b' = b - z$. Practical denoising schemes hence have to rely on various types of prior information on x and b to design an appropriate denoising function ψ.

1.1 Bayesian estimation

A standard statistical approach to the denoising problem consists in assuming that x and b are drawn independently at random from known *prior* probability distributions P_X and P_B. Under this *model*, given a cost function $\mathcal{C}(\hat{x}, x)$ that measures the quality of an estimator $\hat{x}$ in comparison to the true quantity to estimate x, the Bayes estimator is defined as an estimator ψ with minimum expected cost:

$$\arg\min_{\psi} \mathbb{E}\left\{\mathcal{C}(\psi(X + B), X)\right\}.$$

For a quadratic cost function $\mathcal{C}(\hat{x}, x) := \|\hat{x} - x\|_2^2$ the Bayes estimator is the minimum mean square error (MMSE) estimator [5], also called conditional mean, posterior mean, or conditional expectation:

$$\psi_{\texttt{MMSE}}(y) := \mathbb{E}(X|Y = y). \tag{1.1}$$

Even though this estimator is "optimal" in the above defined sense, its computation involves a high-dimensional integral and cannot generally be done explicitly. In practice, Monte-Carlo simulations can be used to approximate the integral.

Often more amenable to efficient numerical optimization is the popular Maximum A Posteriori (MAP) criterion, which is the Bayes estimator associated to the $0-1$ cost function ($\mathcal{C}(\hat{x}, x) = 1$, when $\hat{x} \neq x$; $\mathcal{C}(\hat{x}, x) = 0$, when $\hat{x} = x$). Exploiting Bayes rule and assuming that both the noise and the unknown noiseless signal have probability density functions (pdf), p_X and $p_B(b)$, the MAP estimator reads:

$$\begin{aligned} \psi_{\texttt{MAP}}(y) &:= \arg\max_{x\in\mathbb{R}^n} p(x|y) = \arg\max_{x\in\mathbb{R}^n} p(y|x)p(x) \\ &= \arg\min_{x\in\mathbb{R}^n} \left\{-\log p_B(y - x) - \log p_X(x)\right\}. \end{aligned}$$

For white Gaussian noise b we have $p_B(b) \propto \exp(-\|b\|_2^2/2)$, where $\|b\|_2^2 = \sum_{i=1}^n b_i^2$ and the notation $f(x) \propto g(x)$ means $f(x) = C \cdot g(x)$ for all x, with $C \neq 0$ some constant independent of x. Hence the MAP estimator under the prior $p_X(x)$ can be expressed as

$$\psi_{\texttt{MAP}}(y) = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + [-\log p_X(x)]. \tag{1.2}$$

1.2 Regularization

Optimization problems of the type (1.2) have also been often considered in signal processing without explicit reference to probabilities or priors, under the generic form

$$\arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|y-x\|_2^2 + \phi(x). \tag{1.3}$$

The deterministic objective is to achieve a tradeoff between the data-fidelity term $\|y-x\|_2^2$ and the penalty term $\phi(x)$, which promotes solutions with certain properties. In particular, when the function ϕ is non-smooth at the origin, such as $\phi(x) = \|x\|_p^p := \sum_{i=1}^n |x_i|^p, 0 < p \leq 1$, the optimum of the criterion (1.3) is known to have few nonzero entries. Regularization with such penalty functions is at the basis of *shrinkage* techniques for signal denoising (see e.g. [3] with $p = 1$, or [6] with $0 < p \leq 1$). More recently, these approaches have become a very popular means of promoting *sparse* solutions to under-determined or ill-conditioned linear inverse problems $y = \mathbf{A}x + b$, and are now a key tool for compressed sensing [4].

1.3 Plurality of Bayesian interpretations of regularization

Given the identity of the optimization problems (1.2) and (1.3) when $\phi(x) = \phi_{\texttt{MAP}}(x) := -\log p_X(x)$, the regularization problem (1.3) is often interpreted [1] as "solving the MAP under the prior $p_X(x) = \exp(-\phi(x))/C_\phi$", where

$$C_\phi := \int_{\mathbb{R}^n} \exp(-\phi(x))dx. \tag{1.4}$$

In particular, when $\phi(x) = \|x\|_1$, a possible interpretation of (1.3) is MAP denoising under a Laplacian prior on x and white Gaussian noise.

The main objective of this paper is to highlight the fact that while *one* Bayesian interpretation of the penalized least-squares estimator (1.3) with penalty function $\phi(x)$ is the MAP estimator $\psi_{\texttt{MAP}}(y)$ with prior $p_X(x) = \exp(-\phi(x))/C_\phi$, *there can be other admissible Bayesian interpretations.*

We focus on white Gaussian denoising and show that *for any prior* P_X and any noisy observation $y \in \mathbb{R}^n$, the MMSE estimate $\psi_{\texttt{MMSE}}(y)$ under the prior P_X is the solution of a penalized least-squares problem (1.3) with an appropriate penalty function $\phi_{\texttt{MMSE}}(x)$. Thus, the problem (1.3) with penalty $\phi_{\texttt{MMSE}}(x)$ can equally be interpreted as: a) the MAP estimator $\tilde{\psi}_{MAP}(y)$ with a prior associated to the pdf $\tilde{p}_X(x) = \exp(-\phi_{\texttt{MMSE}}(x))/C_{\phi_{\texttt{MMSE}}}$; or b) the MMSE estimator with prior P_X. In general $dP_X(x) \neq \tilde{p}_X(x)dx$.

2 Main results

From now on we focus on Gaussian denoising: $B \in \mathbb{R}^n$ is a centered normal Gaussian variable with law $P_B = \mathcal{N}(0, \mathbf{I}_n)$ and pdf $p_B(b) \propto \exp(-\|b\|_2^2/2)$. We let $X \in \mathbb{R}^n$ be a random variable independent of B, with law P_X. The probability distribution of the noisy observation $Y = X + B$ has a pdf

$$p_Y(y) := p_B \star P_X(y) = \int_{\mathbb{R}^n} p_B(y-x)dP_X(x) \tag{2.1}$$

[1] This interpretation only makes sense if $C_\phi < \infty$ is integrable. Otherwise some authors refer to a "non-informative prior".

which is sometimes refered to as the *evidence* of the observation y. When P_X is associated to a pdf $p_X(x)$, the evidence is given by a standard convolution between pdfs $p_Y = p_B \star p_X$. Even when P_X is not associated to a pdf, p_Y infinitely differentiable, i.e., $p_Y \in C^\infty(\mathbb{R}^n)$.

In this setting, using techniques going back to Stein's unbiased risk estimator [9, 1], one can express the MMSE estimator as [8]

$$\psi_{\texttt{MMSE}}(y) = y + \frac{1}{p_Y(y)} \left[\frac{\partial}{\partial y_i} p_Y(y) \right]_{i=1}^{n} = y + \nabla \log p_Y(y). \tag{2.2}$$

All vectors $u \in \mathbb{R}^n$, such as the gradient $\nabla \log p_Y(y) \in \mathbb{R}^n$, are in column form. Their transpose u^T is in row form.

Next we study whether $\psi_{\texttt{MMSE}}$ can also be written as the optimum of an optimization problem of the MAP type (1.3), with an appropriate choice of ϕ. Namely, we investigate when $\psi_{\texttt{MMSE}}$ can be identified with the *proximity operator* [2] of a function ϕ, where we recall the definition

$$\mathrm{prox}_\phi(y) := \arg\min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y - z\|_2^2 + \phi(z) \right\}. \tag{2.3}$$

Rereading Equation (1.2) the MAP estimator (with prior $p_X(x)$) can be written as $\psi_{\texttt{MAP}} = \mathrm{prox}_{\phi_{\texttt{MAP}}}$ where

$$\phi_{\texttt{MAP}}(x) := -\log p_X(x). \tag{2.4}$$

For smooth ϕ we have the implicit characterization [2]

$$\mathrm{prox}_\phi(y) := y - \nabla\phi[\mathrm{prox}_\phi(y)], \quad \forall y \in \mathbb{R}^n. \tag{2.5}$$

Comparing (2.2) with (2.5), we see that if $\psi_{\texttt{MMSE}} = \mathrm{prox}_\phi$ then

$$\nabla\phi[\psi_{\texttt{MMSE}}(y)] = -\nabla \log p_Y(y), \quad \forall y \in \mathbb{R}^n. \tag{2.6}$$

Indeed, the relation (2.6) characterizes all functions ϕ such that $\psi_{\texttt{MMSE}} = \mathrm{prox}_\phi$, thanks to the following lemma.

Lemma 2.1. *Let X $P_X, B \sim P_B\mathcal{N}(0, \mathbf{I})$ be independent random variables in $\mathbb{R}^n$. Assume that there is no pair $v \in \mathbb{R}^n$, $c \in \mathbb{R}$ such that $\langle X, v\rangle = c$ with probability one. Then the MMSE estimator $y \mapsto \psi_{MMSE}(y)$ has the following properties:*

1. *it is* one-to-one *from $\mathbb{R}^n$ onto $Im\psi_{MMSE} \subset \mathbb{R}^n$: for any pair $y, y' \in \mathbb{R}^n$, if $\psi_{MMSE}(y) = \psi_{MMSE}(y')$ then $y = y'$.*
2. *it is $C^\infty(\mathbb{R}^n)$; so is its inverse $\psi_{MMSE}^{-1} : Im\psi_{MMSE} \to \mathbb{R}^n$.*
3. *when $n = 1$ we further have that ψ_{MMSE} is increasing.*

The proof is in Appendix .1. Note that the probability distribution P_X in Lemma 2.1 can be almost arbitrary, provided that there is no lower-dimensional affine space of $\mathbb{R}^n$ to which X belongs almost surely. In particular, P_X *need not be separable.* In light of this lemma, (2.6) is equivalent to

$$\nabla\phi(z) = -\nabla \log p_Y[\psi_{\texttt{MMSE}}^{-1}(z)], \quad \forall z \in \mathrm{Im}\psi_{\texttt{MMSE}}.$$

As shown by our main theorem (the proof is in Appendix .2), this equation is satisfied by the function $\phi_{\texttt{MMSE}} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined as:

$$\phi_{\texttt{MMSE}}(x) := \begin{cases} -\frac{1}{2}\|\psi_{\texttt{MMSE}}^{-1}(x) - x\|_2^2 & -\log p_Y[\psi_{\texttt{MMSE}}^{-1}(x)]; \\ & \text{for } x \in \mathrm{Im}\psi_{\texttt{MMSE}}; \\ +\infty, & \text{for } x \notin \mathrm{Im}\psi_{\texttt{MMSE}}. \end{cases} \tag{2.7}$$

Theorem 2.2. *Let X $P_X, B \sim P_B = \mathcal{N}(0, \mathbf{I})$ be independent random variables in $\mathbb{R}^n$. Assume that there is no lower-dimensional affine space of $\mathbb{R}^n$ to which X belongs almost surely. Then $prox_{\phi_{MMSE}} = \psi_{MMSE}$ and:*

1. *the function ϕ_{MMSE} is C^∞ on its domain $Im\psi_{MMSE}$;*
2. *for every $y \in \mathbb{R}^n$, the vector $\psi_{MMSE}(y) = prox_{\phi_{MMSE}}(y)$ is the* unique global minimum*, as well as the* unique stationary point *of the function $x \mapsto \frac{1}{2}\|y - x\|^2 + \phi_{MMSE}(x)$;*
3. *for every $y \in \mathbb{R}^n$, we have $\phi_{MMSE}(y) \geq -\log p_Y(y)$;*
4. *we have $C_{\phi_{MMSE}} = \int_{\mathbb{R}^n} \exp(-\phi_{MMSE}(x))dx < \infty$.*

Therefore, the MMSE estimator with prior P_X and white Gaussian noise is also the MAP estimator with the prior which pdf is $\tilde{p}_X(x) = \exp(-\phi_{MMSE}(x))/C_{\phi_{MMSE}}$.

Remark 2.1. Note that $\psi(y)$ is not only the unique global minimum of $x \mapsto \frac{1}{2}\|y-x\|^2+\phi_{\texttt{MMSE}}(x)$: it is also its unique stationary point. This is much stronger: this means that descent algorithms used to solve the optimization problem (1.3) with $\phi = \phi_{\texttt{MMSE}}$ cannot be trapped in a spurious local minimum.

Remark 2.2. When X belongs with probability one to a lower-dimensional affine space $V \subset \mathbb{R}^n$, we have $\mathrm{Im}\psi_{\texttt{MMSE}} \subset V$. Letting V be the smallest such affine space, the restriction of $\psi_{\texttt{MMSE}}$ to V still has a well defined C^∞ inverse $\psi_{\texttt{MMSE}}^{-1} : \mathrm{Im}\psi_{\texttt{MMSE}} \to V$ which can be used to define $\phi_{\texttt{MMSE}}$ as in (2.7) and to generalize Theorem 2.2 to an arbitrary prior P_X.

3 Worked example

Let us illustrate Theorem 2.2 with a simple example: we consider the one-dimensional ($n = 1$) mixture of two Gaussians prior on the unknown noiseless data x,

$$p_X(x) := p \cdot \frac{e^{-\frac{x^2}{2\sigma_0^2}}}{\sqrt{2\pi\sigma_0^2}} + (1-p) \cdot \frac{e^{-\frac{x^2}{2\sigma_1^2}}}{\sqrt{2\pi\sigma_1^2}}, \tag{3.1}$$

where $p \in (0,1)$ and $0 < \sigma_0 < \sigma_1$. The evidence of the observed noisy data $Y = X + B$ with $B \sim \mathcal{N}(0,1)$ is then

$$p_Y(y) = p \cdot \frac{e^{-\frac{y^2}{2(\sigma_0^2+1)}}}{\sqrt{2\pi(\sigma_0^2+1)}} + (1-p) \cdot \frac{e^{-\frac{y^2}{2(\sigma_1^2+1)}}}{\sqrt{2\pi(\sigma_1^2+1)}},$$

hence

$$p_Y'(y) = -y \cdot \left\{ p \cdot \frac{e^{-\frac{y^2}{2(\sigma_0^2+1)}}}{\sqrt{2\pi(\sigma_0^2+1)^3}} + (1-p) \cdot \frac{e^{-\frac{y^2}{2(\sigma_1^2+1)}}}{\sqrt{2\pi(\sigma_1^2+1)^3}}. \right\}$$

By straightforward computations, we obtain

$$\psi_{\texttt{MMSE}}(y) = y \cdot \frac{\frac{\sigma_0^2}{\sigma_0^2+1} + \frac{\sigma_1^2}{\sigma_1^2+1} \cdot ae^{by^2}}{1+ae^{by^2}} \text{ with}$$
$$a := \frac{1-p}{p}\sqrt{\frac{\sigma_0^2+1}{\sigma_1^2+1}},\ b = \frac{1}{\sigma_0^2+1} - \frac{1}{\sigma_1^2+1} \in (0,1).$$

The limiting case $\sigma_0^2 \to 0$ corresponds to the so called Bernoulli-Gaussian prior (see, e.g., [11]): the value $x = 0$ is drawn with probability $p > 0$, hence vectors with i.i.d. entries distributed according to p_X are typically sparse. The MMSE estimator takes a simplified form [10] when $\sigma_0^2 \to 0$

$$\psi_{\texttt{MMSE}}(y) = y \cdot \frac{\sigma_1^2}{\sigma_1^2+1} \cdot \frac{ae^{by^2}}{1+ae^{by^2}}.$$

We illustrate in Figure 1 the case $p = 0.9$, $\sigma_0^2 \to 0$, $\sigma_1^2 = 10$. Figure 1(a) shows $\psi_{\texttt{MMSE}}$ (solid line) and its inverse $\psi_{\texttt{MMSE}}^{-1}$ (dashed line). The latter does not seem to have an analytic expression. Figure 1(b) shows $\phi_{\texttt{MAP}}(x) = -\log p_X(x)$ (dotted line), $-\log p_Y(x)$ (dashed line) and the penalty function $\phi_{\texttt{MMSE}}(x)$ (solid line). While the penalty function $\phi_{\texttt{MMSE}}(x)$ does not seem to admit an analytic expression, one can obtain an analytic expression for $\phi_{\texttt{MMSE}}[\psi_{\texttt{MMSE}}(y)] = -\frac{1}{2}\|y - \psi_{\texttt{MMSE}}(y)\|_2^2 - \log p_Y(y)$. The explicit analytic expression –which is long and rather uninteresting– was used to plot $\phi_{\texttt{MMSE}}(x)$ on Figure 1(a) using the parameterized curve $y \mapsto (\psi_{\texttt{MMSE}}(y), \phi_{\texttt{MMSE}}[\psi_{\texttt{MMSE}}(y)])$. Observing on Figure 1(b) the plot of

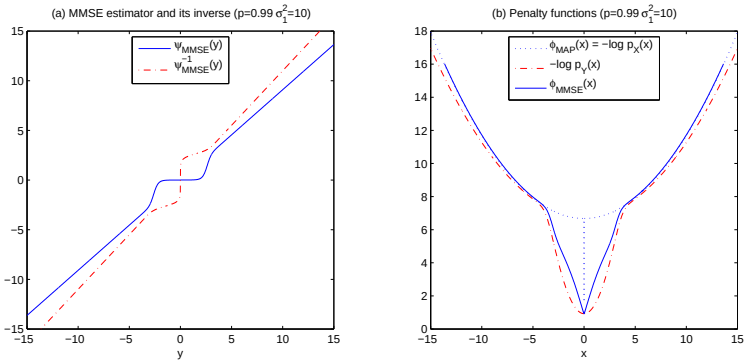

Figure 1: Left: MMSE estimator $\psi_{\texttt{MMSE}}(y)$ (solid line) and its inverse $\psi_{\texttt{MMSE}}^{-1}(y)$ (dashed line),, in the Bernoulli-Gaussian case, $p = 0.9$, $\sigma_0^2 \to 0$, $\sigma_1^2 = 10$. Right: MAP penalty $\phi_{\texttt{MAP}}(x) = -\log p_X(x)$ (dotted line), negative log-evidence ($-\log p_Y(x)$) (dashed line) and MMSE penalty $\phi_{\texttt{MMSE}}(x)$ (solid line).

$\phi_{\texttt{MMSE}}(x)$ for the above Bernoulli-Gaussian prior yields a number of observations.

1. For small x, the penalty $\phi_{\texttt{MMSE}}(x)$ is approximately shaped as the absolute value: $\phi_{\texttt{MMSE}}(x) \approx c|x|$ for some constant c. This is tempered by the fact that $\phi_{\texttt{MMSE}}(x)$ is C^∞, thus, unlike $|x|$, it must be smooth at zero.

2. The penalty $\phi_{\texttt{MMSE}}(x)$ is unimodal (it is decreasing until its global minimum, then increasing) but *it is not convex*.

The second observation could seem surprising given that Theorem 2.2 guarantees the uniqueness of the global minimizer / stationary point of $x \mapsto \frac{1}{2}\|y - x\|^2 + \phi_{\texttt{MMSE}}(x)$. However, this property is not a characteristic of convex penalties. As a matter of fact, a function $f : \mathbb{R} \to \mathbb{R}$ (i.e., in the case $n = 1$) can be written $f = \text{prox}_g$ with g a proper lower semi-continuous *convex* function from $\mathbb{R}$ to $\mathbb{R} \cup \{+\infty\}$ if, and only if, the function f is increasing and *non-expansive* [2, Proposition 2.4]:

Definition 3.1. A function $f : \mathbb{R} \to \mathbb{R}$ is non-expansive if $|f(y') - f(y)| \leq |y' - y|$ for all y, y'. When f is differentiable, it is non-expansive if and only if $|f'(y)| \leq 1$ for all y.

By Lemma 2.1, in dimension $n = 1$, the MMSE estimator $\psi_{\texttt{MMSE}}$ is increasing for any prior P_X. However, for certain priors P_X, it can indeed be proved to be expansive (see the proof in Appendix .3):

Proposition 3.2. *Assume that X has a symmetric pdf [$\forall x \in \mathbb{R},\ p_X(-x) = p_X(x)$] and that there exists $\varepsilon > 0$ such that $p_X(x) = 0$ for all x with $|x| < 1 + \varepsilon$. Then the penalty $\phi_{\texttt{MMSE}}$ cannot be convex.*

4 Discussion

Theorem 2.2 shows that for general priors P_X we have $\psi_{\texttt{MMSE}} = \text{prox}_{\phi_{\texttt{MMSE}}}$. Similarly, when X has a pdf, we have $\psi_{\texttt{MAP}} = \text{prox}_{\phi_{\texttt{MAP}}}$, where for a given prior the MAP penalty $\phi_{\texttt{MAP}}(x)$ has the simple expression (2.4) while the MMSE penalty $\phi_{\texttt{MMSE}}(x)$ has the much more intricate definition (2.7).

For Gaussian priors $P_X = \mathcal{N}(0, \Sigma)$, the MMSE estimator is the Wiener filter, which is also the MAP and the minimum mean square linear estimator [5], so $\phi_{\texttt{MMSE}} = \phi_{\texttt{MAP}}$ (up to a constant additive term).

However, for most priors with a pdf $p_X(x)$, the MMSE estimator does not coincide with the MAP estimator (i.e., $\psi_{\texttt{MMSE}} \neq \psi_{\texttt{MAP}}$), hence $\phi_{\texttt{MMSE}} \neq \phi_{\texttt{MAP}}$ (even up to a constant additive term). Indeed, by Theorem 2.2, the penalty $\phi_{\texttt{MMSE}}(x)$ defined in (2.7) has a number of specific properties. Therefore, if $\phi_{\texttt{MAP}}(x) = -\log p_X(x)$ fails to satisfy one of these properties, then the identity $\phi_{\texttt{MMSE}}(x) = \phi_{\texttt{MAP}}(x) + c$ (for some constant $c \in \mathbb{R}$ and all $x \in \mathbb{R}^n$) cannot be satisfied.

For example, generalized Gaussian priors $p_X(x) \propto \exp(-\alpha\|x\|_p^p)$ with $0 < p \leq 1$ are *not smooth* at $x = 0$, hence they are not C^∞: as a result for such priors there is not even any pair $a, b \in \mathbb{R}$ such that $\phi_{\texttt{MMSE}}(x) = a + b \cdot \phi_{\texttt{MAP}}(x)$ for all x.

One may also wonder whether a reciprocal to Theorem 2.2 is possible: given a penalty function $\phi(x)$, does there exist a prior P_X such that the MMSE estimator $\psi_{\texttt{MMSE}}$ with this prior is associated to the penalty $\phi_{\texttt{MMSE}}(x) = \phi(x)$ (up to a constant additive term) ? When this prior exists, can we characterize it in terms of the penalty function ϕ ? Even though one can always define the tentatively associated "MMSE estimator" $\psi(y) = \text{prox}_\phi(y)$, the main difficulty is to understand when there exists a probability measure P_X such that $\psi(y) - y = \nabla \log(p_B \star P_X)(y)$. This combined integration and Gaussian deconvolution problem often does not admit a solution, for example: when ψ is not one to one; when $\phi(x)$ is not sufficiently smooth.

5 Conclusion and perspectives

We proved that the MMSE estimator for Gaussian denoising with *any* prior can be written as the MAP estimator with a possibly different prior (and that the MAP estimator with *certain* priors can be interpreted as a MMSE estimator with a possibly different prior). These results, in conjunction with Nikolova's highlighting of model distortions brought by MAP estimation [7], indicate that one should be cautious when interpreting penalized least squares regression schemes in terms of priors:

- *If* the unknown noiseless data x follows a prior with pdf $p_X(x) \propto \exp(-\phi(x))$ and *if* we choose the MAP as a criterion for estimating it, *then* the resulting denoising scheme leads to penalized least squares regression with penalty $\phi(x)$. This MAP estimator may however have poor denoising performance[2] for this type of data [7].

- In practice, the choice of penalized least squares regression with penalty $\phi(x)$ is seldomly associated to the *belief* that the unknown noiseless data follows a prior with pdf $p_X(x) \propto \exp(-\phi(x))$. Instead, it rather stems from the *need* for numerical efficiency and the *empirical observation* that it achieves good denoising performance for the considered class of data.

  By definition, optimum denoising (as measured by the mean squares error) is achieved by the MMSE estimator. As shown in this paper, the latter is indeed always associated to a penalized least squares scheme[3]. This sheds a new light on the popularity of such schemes for Gaussian denoising.

Quite obviously, the denoising performance of penalized least squares regression with a given penalty $\phi(x)$ heavily depends on the prior P_X underlying the unknown noiseless data. We focused in this paper on the case where the penalized least squares regression estimator $\psi(y) = \mathrm{prox}_{\phi}(y)$ coincides with the MMSE estimator: its denoising performance $\mathbb{E}(\|\mathrm{prox}_{\phi}(Y) - X\|_2^2)$ is optimum. An interesting open problem related to the results of this paper would be to understand for which priors P_X we obtain "good" denoising performance with $\psi(y) = \mathrm{prox}_{\phi}(y)$, i.e., when the denoising performance is bounded by a constant $C > 1$ times the optimum performance.

One can imagine concrete applications of the results presented here for certain priors: in general the MMSE estimator $\psi_{\texttt{MMSE}}(y)$ is *a priori* expressed as an intractable high-dimensional integral; however, if the penalty function $\phi_{\texttt{MMSE}}(x)$ admits a simple expression amenable to efficient numerical optimization (e.g., convex optimization), then the MMSE estimator can be computed efficiently. Developping such approaches requires a more in-depth understanding of the properties of penalty functions $\phi_{\texttt{MMSE}}(x)$ obtained through Theorem 2.2. Of particular interest would be the construction of explicit examples where $\phi_{\texttt{MMSE}}(x)$ is "simple" while $p_Y(y)$ involves an intractable integral.

Another interesting perspective is to obtain alternate statistical interpretations of a larger class of penalized least squares regression estimators (e.g., with

[2] Even though, as shown in this paper, this MAP scheme can sometimes be interpreted as an MMSE estimator with a different prior, this re-interpretation does not alter the denoising scheme nor its denoising performance.

[3] Yet, the associated penalized least squares problem may not be more computationally tractable than the original MMSE.

non-smooth $\phi(x)$ such as those leading to sparse estimates). As remarked above, the lack of smoothness makes it impossible to interpret such estimators in terms of a MMSE estimator, however one may seek interpretations that leave the strict Bayesian framework: for example, one may wish to obtain an interpretation as the optimum of a hybrid Bayesian cost function $\min_\psi \{\mathbb{E}\mathcal{C}(\psi(X+B),X)+\mathbf{K}(\psi)\}$ where the term $\mathbf{K}(\cdot)$ forces the function ψ to be in some function class. Eventually, one may also wish to extend theses results to ill-posed linear inverse problems of the type $y=\mathbf{A}x+b$, and to deal with non-Gaussian noise.

6 Acknowledgements

This existence of this paper owes much to several discussions with Mike Davies about the Bayesian "interpretation" of sparse regularization, as well as intense discussions with Jérôme Idier on the same topic during the second French Spring School of Inverse Problems in Signal Processing, held in the beautiful mediterranean island of Porquerolles in the spring of 2010. The author is very thankful to Mike, Jérôme, and to the organizers of the Spring school for these passionate discussions, and would also like to thank Jean-Christophe Pesquet, who provided his insight on proximity operators, as well as Patrick Perez and Miki Elad, whose comments on a draft version of this paper were precious. Last, the comments of the three anonymous reviewers were very helpful to improve the final version of the paper.

.1 Proof of Lemma 2.1

Lemma .1. *Denote $\psi_{\mathtt{MMSE}}(y)=\left(\psi^i_{\mathtt{MMSE}}(y)\right)_{i=1}^n$ where $\psi^i_{\mathtt{MMSE}}:\mathbb{R}^n\to\mathbb{R}$ is scalar valued. Under the assumptions of Lemma 2.1, the $n\times n$ Jacobian matrix $J[\psi_{\mathtt{MMSE}}](y):=\left(\frac{\partial}{\partial y_j}\psi^i_{\mathtt{MMSE}}(y)\right)_{ij}$ satisfies the identity*

$$J[\psi_{\mathtt{MMSE}}](y)=\left(\delta_{ij}+\frac{\partial^2\log p_Y(y)}{\partial y_i\partial y_j}\right)_{ij}=\mathbf{I}+\nabla^2\log p_Y(y) \tag{.1}$$

and is symmetric positive definite:

$$\langle v, J[\psi_{\mathtt{MMSE}}](y)\cdot v\rangle>0,\quad \forall y\in\mathbb{R}^n, v\neq 0. \tag{.2}$$

Proof. Without loss of generality we consider a unit norm vector $\|v\|_2=1$. For brevity we omit the dependency in the variable y when possible. First, by (1.2) we have

$$\psi_{\mathtt{MMSE}}(y)=y+\nabla\log p_Y(y)=y+\nabla p_Y(y)/p_Y(y)$$

hence

$$J[\psi_{\mathtt{MMSE}}] \;=\; \mathbf{I}+\nabla^2\log p_Y=\mathbf{I}+\frac{\nabla^2 p_Y}{p_Y}-\frac{\nabla p_Y\cdot(\nabla p_Y)^T}{[p_Y]^2}$$

and

$$\langle J[\psi_{\mathtt{MMSE}}]\cdot v, v\rangle=\frac{p_Y^2+p_Y\langle\nabla^2 p_Y\cdot v, v\rangle-\langle\nabla p_Y, v\rangle^2}{p_Y^2}. \tag{.3}$$

We will now prove that the numerator in (.3) is positive for all y. Since $p_B(b) \propto e^{-\|b\|_2^2/2}$, we have

$$\begin{aligned} \nabla p_B(b) &= (-b) \cdot p_B(b), \\ \nabla^2 p_B(b) &= (bb^T - \mathbf{I}) \cdot p_B(b). \end{aligned}$$

Since $p_Y = p_B \star P_X$, $\nabla p_Y = \nabla p_B \star P_X$, $\nabla^2 p_Y = \nabla^2 p_B \star P_X$ this yields

$$\begin{aligned} p_Y &= \int p_B(y-x)\ dP_X(x) \\ \langle \nabla p_Y, v\rangle &= \int (-\langle y-x, v\rangle) \cdot p_B(y-x) dP_X(x) \\ \langle \nabla^2 p_Y \cdot v, v\rangle &= \int \left(\langle y-x, v\rangle^2 - 1\right) \cdot p_B(y-x) dP_X(x) \end{aligned}$$

hence

$$\begin{aligned} p_Y \langle \nabla^2 p_Y \cdot v, v\rangle = \iint & \left(\langle y-x, v\rangle^2 - 1\right) \\ & \cdot p_B(y-x) p_B(y-x') dP_X(x) dP_X(x') \end{aligned}$$

The above expression is also valid is we exchange the role of the integration variables b and b', hence by taking the average of these two equal expressions we obtain

$$\begin{aligned} p_Y \langle \nabla^2 p_Y \cdot v, v\rangle = \iint & \left[\tfrac{\langle y-x, v\rangle^2 + \langle y-x', v\rangle^2}{2} - 1\right] \\ & \cdot p_B(y-x) p_B(y-x') dP_X(x) dP_X(x') \end{aligned}$$

Similarly we can write

$$\begin{aligned} p_Y^2 &= \iint p_B(y-x) p_B(y-x') dP_X(x) dP_X(x') \\ \langle \nabla p_Y, v\rangle^2 &= \iint \langle y-x, v\rangle \langle y-x', v\rangle \\ & \qquad \cdot p_B(y-x) p_B(y-x') dP_X(x) dP_X(x') \end{aligned}$$

Overall, the numerator of the right hand side in (.3) becomes

$$\iint \frac{\langle x'-x, v\rangle^2}{2} p_B(y-x) p_B(y-x') dP_X(x) dP_X(x'). \tag{.4}$$

Now, since there is no c such that $\langle X, v\rangle = c$ with probability one, there exists $x_1, x_2 \in \mathbb{R}^n$, $d = \langle x_2 - x_1, v\rangle \neq 0$, such that the Euclidean balls $B_i = B(x_i, d/3) \subset \mathbb{R}^n$, have positive probability $P_X(B_i) > 0$. For $(x, x') \in B_1 \times B_2$ the function $g(x, x') := \frac{\langle x'-x, v\rangle^2}{2} p_B(y-x) p_B(y-x')$ is bounded from below by some constant $\eta > 0$, hence the integral in (.4) is bounded from below by

$$\iint_{B_1 \times B_2} g(x, x') dP_X(x) dP_X(x') \geq \eta \cdot P_X(B_1) P_X(B_2) > 0.$$

We conclude that $\langle J[\psi_{\texttt{MMSE}}] \cdot v, v\rangle > 0$.

□

We are now equipped to prove Lemma 2.1.

Proof of Lemma 2.1. We let the reader check that p_Y cannot vanish. Since it is C^∞, $\psi_{\texttt{MMSE}}$ is also C^∞. To prove that $\psi_{\texttt{MMSE}}$ is one-to-one, we proceed by contradiction, assuming that $\psi_{\texttt{MMSE}}(y) = \psi_{\texttt{MMSE}}(y')$ while $y' \neq y$. We define $v := (y'-y)/\|y'-y\|_2$ and the function $f : t \mapsto f(t) := \langle v, \psi_{\texttt{MMSE}}(y+tv)\rangle \in \mathbb{R}$. Since the function f is smooth and $f(0) = f(\|y'-y\|_2)$, by Rolle's theorem the derivative of f must vanish for some $0 < t < \|y'-y\|_2$. However by Lemma .1 we have $f'(t) = \langle v, J[\psi_{\texttt{MMSE}}](y+tv)\cdot v\rangle > 0$ which yields a contradiction. Therefore, the inverse function $\psi^{-1}_{\texttt{MMSE}}$ exists as claimed. The fact that it is also C^∞ follows from the positivity of the Jacobian of $\psi_{\texttt{MMSE}}$ and the inverse function theorem. □

.2 Proof of Theorem 2.2

The fact that $\phi_{\texttt{MMSE}}$ is C^∞ on $\mathrm{Im}\psi_{\texttt{MMSE}}$ is a straightforward consequence of its definition (2.7) and of the fact that p_Y as well as $\psi^{-1}_{\texttt{MMSE}}$ are C^∞ (Lemma 2.1). We wish to check that the proximity operator of $\phi_{\texttt{MMSE}}$ defined by (2.7) is indeed $\psi_{\texttt{MMSE}}$. The definition of $\phi_{\texttt{MMSE}}(x)$ for $x \notin \mathrm{Im}\psi_{\texttt{MMSE}}$ ensures that $\mathrm{prox}_{\phi_{\texttt{MMSE}}}$ takes its values in $\mathrm{Im}\psi_{\texttt{MMSE}}$. We let the reader check that a consequence of Lemma .1 is that the set $\mathrm{Im}\psi_{\texttt{MMSE}}$ is open. For brevity we denote $q(y) = \log p_Y(y)$ and

$$\begin{aligned} g(u) :=& \frac{1}{2}\|y - \psi_{\texttt{MMSE}}(u)\|_2^2 + \phi_{\texttt{MMSE}}[\psi_{\texttt{MMSE}}(u)] \\ =& \frac{1}{2}\|\psi_{\texttt{MMSE}}(u) - y\|_2^2 - \frac{1}{2}\|\nabla q(u)\|_2^2 - q(u). \end{aligned}$$

Since $J[\psi_{\texttt{MMSE}}](u) = \mathbf{I} + \nabla^2 q(u)$ (Lemma .1) and $\psi_{\texttt{MMSE}}(u) = u + \nabla q(u)$ (Equation (2.2)), we obtain

$$\begin{aligned} \nabla g(u) =& J[\psi_{\texttt{MMSE}}](u)\cdot\big[\psi_{\texttt{MMSE}}(u) - y\big] \\ & - \nabla^2 q(u)\cdot\nabla q(u) - \nabla q(u) \\ =& J[\psi_{\texttt{MMSE}}](u)\cdot\big[\psi_{\texttt{MMSE}}(u) - y - \nabla q(u)\big] \\ =& J[\psi_{\texttt{MMSE}}](u)\cdot\big[u - y\big] \end{aligned}$$

Now consider $f_v(t) := g(y+tv)$ with $v \neq 0$ an arbitrary vector. Its derivative is

$$\begin{aligned} f_v'(t) &= \langle \nabla g(y+tv), v\rangle = \langle J[\psi_{\texttt{MMSE}}](y+tv)\cdot tv, v\rangle \\ &= t\cdot\langle J[\psi_{\texttt{MMSE}}](y+tv).v, v\rangle \end{aligned}$$

which, by Lemma .1, has the sign of t, showing that f_v admits its strict global minimum at $t = 0$. Since this is true for any choice of v it follows that g has no stationary point other that $u = y$, and that $g(u) > g(y)$ whenever $u \neq v$, that is to say $x \mapsto \frac{1}{2}\|y-x\|_2^2 + \phi_{\texttt{MMSE}}(x)$ admits a unique global minimum at $x = \psi_{\texttt{MMSE}}(y)$. To conclude, since $\psi_{\texttt{MMSE}}(y) = \mathrm{prox}_{\phi_{\texttt{MMSE}}}(y)$, we have for any y

$$\begin{aligned} \phi_{\texttt{MMSE}}(y) &= \frac{1}{2}\|y - y\|_2^2 + \phi_{\texttt{MMSE}}(y) \\ &\geq \frac{1}{2}\|y - \psi_{\texttt{MMSE}}(y)\|_2^2 + \phi_{\texttt{MMSE}}[\psi_{\texttt{MMSE}}(y)] \\ &= -\log p_Y(y). \end{aligned}$$

As a result $0 \leq \exp(-\phi_{\texttt{MMSE}}(y)) \leq p_Y(y)$, and since $p_Y(y)$ is integrable so is $\exp(-\phi_{\texttt{MMSE}}(y))$.

.3 Proof of Lemma 3.2

Thanks to (.4), and since both p_B and p_X are symmetric, the numerator of (.3) for $y = 0$ reads

$$\begin{aligned}
&\iint \frac{(x'-x)^2}{2} \cdot p_B(-x)p_B(-x')p_X(x)p_X(x')dxdx' \\
&= \int \frac{x^2}{2} \cdot p_B(x)p_X(x)dx \cdot \int p_B(x')p_X(x')dx' \\
&\quad + \int \frac{(x')^2}{2} \cdot p_B(x')p_X(x')dx' \cdot \int p_B(x)p_X(x)dx \\
&\quad - \int x' \cdot p_B(x')p_X(x')dx' \cdot \int x \cdot p_B(x)p_X(x)dx \\
&= \int x^2 \cdot p_B(x)p_X(x)dx \cdot \int p_B(x')p_X(x')dx'
\end{aligned}$$

Since $p_Y(y) = \int p_B(x)p_X(x)dx$, inserting the above expression in (.3) for $y = 0$ and using that $p_X(x) = 0$ for $|x| < 1 + \varepsilon$ we obtain

$$\begin{aligned}
\psi'_{\texttt{MMSE}}(0) &= \frac{\int x^2 \cdot p_B(x)p_X(x)dx}{\int p_B(x)p_X(x)dx} \\
&= \frac{\int_{|x|\geq 1+\varepsilon} x^2 \cdot p_B(x)p_X(x)dx}{\int_{|x|\geq 1+\varepsilon} p_B(x)p_X(x)dx} \geq (1+\varepsilon)^2 > 1.
\end{aligned}$$

Therefore, $\psi_{\texttt{MMSE}}$ is expansive. Since it is also increasing, the associated $\phi_{\texttt{MMSE}}$ is C^∞ (Theorem 2.2) hence it is proper and continuous. As a result of [2, Proposition 2.4], since $\psi_{\texttt{MMSE}} = \text{prox}_{\phi_{\texttt{MMSE}}}$, the penalty $\phi_{\texttt{MMSE}}$ cannot be convex. Similar examples can be built in higher dimensions.

References

[1] T. Blu and F. Luisier. The SURE-LET approach to image denoising. *IEEE Trans. Image Proc.*, 16(11):2778–2786, 2007.

[2] P. L. Combettes and J.-C. Pesquet. Proximal thresholding algorithm for minimization over orthonormal bases. *SIAM J. Optim.*, 18:1351–1376, 2007.

[3] D. Donoho. Denoising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3):613–627, May 1995.

[4] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[5] S. Kay. *Fundamentals of Statistical Signal Processing : Estimation Theory.* Signal Processing. Prentice Hall, 1993.

[6] P. Moulin and J. Liu. Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. *IEEE Trans. on Information Theory*, 45(3):909–919, April 1999.

[7] M. Nikolova. Model distortions in Bayesian MAP reconstruction. *Inverse Problems and Imaging*, 1(2):399–422, 2007.

[8] M. Raphan and E. P. Simoncelli. Learning to be Bayesian without supervision. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Adv. Neural Information Processing Systems (NIPS*06)*, volume 19, pages 1170–1177, Cambridge, MA, may 2007. MIT Press.

[9] C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

[10] J. S. Turek, I. Yavneh, M. Protter, and M. Elad. On MMSE and MAP denoising under sparse representation modeling over a unitary dictionary. *submitted to Applied and Computational Harmonic Analysis*, Technion University, 2010.

[11] H. Zayyani, M. Babaie-Zadeh, and C. Jutten. Bayesian pursuit algorithm for sparse representation. In *Proceedings of* ICASSP2009, pages 1549–1552, Taipei, Taiwan, 19–24 April 2009.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399