A Study of Methods for Choosing the Unknown Parameters in the Bayesian and Regularization Approaches to Image Restoration.

J. Abad, J. Mateos, R. Molina and F.J. Cortijo
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada. 18071 Granada, Spain.

Abstract

The application of regularization to ill-conditioned problems necesitates the choice of a regularization parameter which trades fidelity to the data with smoothness of the solution. The value of the regularization parameter depends on the variance of the noise in the data and characteristics of the image. The Bayesian approach to image restoration uses image and noise models which depend on unknown parameters that have to be estimated to perform the restoration. In this work we compare these approaches to the estimation of the unknown parameters in the image restoration problem.

1 Introduction

The purpose of image restoration can be formulated as the estimation of the original image $\mathbf{f}$, when a noisy and blurred version $\mathbf{g}$, given by

$$\mathbf{g} = \mathbf{D}\mathbf{f} + \mathbf{w} \tag{1}$$

is observed (see, for instance, [?]). We shall assume that $\mathbf{g}$ and $\mathbf{f}$ are $m \times n$ matrices, $\mathbf{w}$ is a random vector of independent errors with variance $\sigma_n^2 = \beta^{-1}$ and $\mathbf{D}$ is a known blurring matrix.

The simplest way to approach the restoration problem is to use least squares estimation and then select $\hat{\mathbf{f}}$, the estimation of $\mathbf{f}$, as

$$\hat{\mathbf{f}} = arg\{ \max_{\mathbf{f}} \frac{1}{Z_{noise}(\beta)} \exp\left[-\frac{1}{2}\beta \parallel \mathbf{g} - \mathbf{D}\mathbf{f} \parallel^2 \right] \} \tag{2}$$

where $Z_{noise}(\beta) = (2\pi/\beta)^{p/2}$ and $p = m \times n$. However, as it is well known, this approach does not lead to useful restorations (see [?]).

Regularization is an effective method for obtaining satisfactory solutions to problems that involve inversion of ill-conditioned operators. According to the regularization approach the solution of (2) is replaced by the minimization of

$$J_\lambda(\mathbf{f}) = \parallel \mathbf{g} - \mathbf{D}\mathbf{f} \parallel^2 + \lambda \parallel \mathbf{Q}\mathbf{f} \parallel^2 \tag{3}$$

where $\mathbf{Q}$ is a real valued matrix of size $mn \times mn$, known as the regularizing operator. Minimization of (3) produces the estimate

$$\hat{\mathbf{f}}(\lambda) = (\mathbf{D}^t\mathbf{D} + \lambda\mathbf{Q}^t\mathbf{Q})^{-1}\mathbf{D}^t\mathbf{g} = A(\lambda)\mathbf{g} \tag{4}$$

Expression (3) can be rewritten as

$$\Delta(\mathbf{f},\mathbf{g}) + \lambda\Phi(\mathbf{f}) \tag{5}$$

funtional, measuring roughness in a well defined sense and λ is a regularization or smoothing parameter. Typical choices of Φ are the quadratic norm and $2-D$ Laplacian of the image. Expository articles of the regularization methods are [?],[?],[?],[?].

Let us now introduce the Bayesian paradigm. Bayesian methods start with a prior distribution, a probability distribution over images $\mathbf{f}$. It is here where we incorporate information on the expected structure within an image. It is also necessary to specify $p(\mathbf{g}|\mathbf{f})$, the probability distribution of observed images $\mathbf{g}$ if $\mathbf{f}$ were the 'true' image. The Bayesian paradigm dictates that inference about the true $\mathbf{f}$ should be based on $p(\mathbf{f}|\mathbf{g})$ given by

$$p(\mathbf{f}|\mathbf{g}) = p(\mathbf{g}|\mathbf{f})p(\mathbf{f})/p(\mathbf{g}) \propto p(\mathbf{g}|\mathbf{f})p(\mathbf{f}) \tag{6}$$

To show just one restoration it is common (but not obligatory) to choose the mode of $p(\mathbf{f}|\mathbf{g})$, that is to display the image $\hat{\mathbf{f}}$ which satisfies

$$\hat{\mathbf{f}} \text{ maximizes } p(\mathbf{f}|\mathbf{g}) \tag{7}$$

This is known as the MAP (maximum a posteriori) estimate of $\mathbf{f}$.

Typical knowledge about the original image has the form

$$p_\alpha(\mathbf{f}) \propto \alpha^{q/2} exp\{-\frac{1}{2}\alpha S(\mathbf{f})\} \tag{8}$$

where $S(\mathbf{f})$ is a non negative quadratic form, q is the number of positive eigenvalues of S and α is a constant. A commonly used degradation model is the normal distribution associated to (1), which defines $p_\beta(\mathbf{g}|\mathbf{f})$.

If α and β are known, then, following the Bayesian paradigm, it is customary to select as the restoration of $\mathbf{f}$ the image $\mathbf{f}_{mp(\alpha,\beta)}$ defined by

$$\mathbf{f}_{mp(\alpha,\beta)} = arg\{\min_{\mathbf{f}} [\, \alpha S(\mathbf{f}) + \beta \| \, \mathbf{g} - \mathbf{D}\mathbf{f} \, \|^2 \,]\} \tag{9}$$

The problem comes when α and/or β are unknown. The empirical Bayesian approach does not assume any distribution on these parameters and try to estimate them by principles like maximum likelihood from the marginal distribution of the observations. The hierarchical Bayesian paradigm introduces a second stage. In that stage an hyperprior $p(\alpha,\beta)$ is defined, giving the distribution $p(\alpha,\beta,\mathbf{f},\mathbf{g})$. Since of interest is in the posterior distribution of $\mathbf{f}$ given $\mathbf{g}$, the hierarchical approach marginalizes out α and β and draws conclusions from $p(\mathbf{f}|\mathbf{g})$.

In this work we shall compare the regularization paradigm and the empirical and hierarchical Bayesian approaches to the estimation of the unknown parameters in an image restoration problem. The paper is divided as follows. In section 2 we shall describe the regularization approach to the estimation of the smoothness parameter. The empirical Bayesian approach will be studied in section 3. In section 4 we shall describe the hierarchical approach by defining precisely the hyperpriors on the hyperparameters. Finally, in section 5 we test and compare the models on real world images.

2 Regularization Approach

In this framework considerable use of the residual sum of squares function, $RSS(\lambda)$, defined by

$$RSS(\lambda) = \| \, \mathbf{g} - \mathbf{D}\hat{\mathbf{f}}(\lambda) \, \|^2 \tag{10}$$

is made.

Hunt ([?]) was the first to address the problem of selecting the value of the regularization parameter in restoration problems. He used a deterministic constrained least squares framework. According to it, the parameter λ was selected such that the following equation holds

$$RSS(\lambda)/p = 1/\beta \tag{11}$$

let us denote by λ_{CLS} such a parameter. Apart from the fact that the above choice of λ requires the prior knowledge of the noise variance, it has been observed and reported by a number of researchers that this choice of λ produces an oversmoothed solution $\hat{\mathbf{f}}(\lambda)$.

2.2 Equivalent Degrees of Freedom (EDF)

Using the CLS approach is equivalent to assuming that the error $(\mathbf{g} - \mathbf{D}\hat{\mathbf{f}}(\lambda))$ is IID with variance $1/\beta$. In other words $RSS(\lambda)$ is Chi-square distributed with variance $1/\beta$ and p degrees of freedom. In this context, in [?], the authors propose to take into account the linear dependency between the observed image and the restoration, and so $RSS(\lambda)$ is Chi-square distributed with variance $1/\beta$ and $trace[I - \mathbf{D}A(\lambda)]$ degrees of freedom. Thus, in this case the constraint equation to be used for computing λ_{EDF} is given by the solution of

$$RSS(\lambda)/EDF(\lambda) = 1/\beta \tag{12}$$

where $EDF(\lambda) = trace[I - \mathbf{D}A(\lambda)]$.

2.3 Generalized Cross-Validation (GCV)

In many practical situations the noise variance is not known and so the methods just described cannot be used. Cross validation is a method that allows the selection of the regularization parameter when the noise variance is not known. It has been shown experimentally that CV yields sharper estimations of the original image than the CLS method.

The ordinary cross validation and the generalized cross validation can be derived from the leave one out principle (see [?] for details). GCV and CV do not differ greatly, and the former has more pleasing mathematical properties. The value of λ_{GCV} is obtained by minimizing

$$GCV(\lambda) = \frac{RSS(\lambda)}{\{trace[I - \mathbf{D}A(\lambda)]\}^2} \tag{13}$$

3 Empirical Bayesian Approach

In this approach we have knowledge about the structural form of the noise (noise model) and about the structural behaviour of the restoration (image model). Here we shall only discuss the image models and assume as noise model the one given by (1).

Although the theory we shall develop here can be applied to any quadratic prior on $\mathbf{f}$ we shall study in full two particular priors, the models we have been applying to image restoration in astronomy (see [?], [?], [?]).

Our prior knowledge about the smoothness of the object luminosity distribution makes it possible to model the distribution of $\mathbf{f}$ by a conditional autoregression (CAR). Thus,

$$p_\alpha(\mathbf{f}) \propto \exp\{-\frac{1}{2}\alpha\, \mathbf{f}^t(I - \phi N)\mathbf{f}\}$$

than 0.25. The term $\mathbf{f}^t(I-\phi N)\mathbf{f}$ represents in matrix notation the sum of squares of the values $\mathbf{f}_i$ minus ϕ times the sum of $\mathbf{f}_i\mathbf{f}_j$ for neighbouring pixels i and j. We shall regard $\mathbf{g}$ as a column vector of values of $\mathbf{g}_i$, and similarly for $\mathbf{f}$.

The parameters can be interpreted by the following expressions describing the conditional distribution

$$\begin{aligned} E(\mathbf{f}_i \mid \mathbf{f}_j,\ j \neq i) &= \phi \sum_{j \text{ nhbr } i} \mathbf{f}_j \\ \operatorname{var}(\mathbf{f}_i \mid \mathbf{f}_j,\ j \neq i) &= \alpha^{-1} \end{aligned}$$

where the suffix 'j nhbr i' denotes the four neighbour pixels at distance one from pixel i. The parameter α measures the smoothness of the 'true' image.

This model on log scale has been used in the galaxy deconvolution problem (see [?], [?]) a problem in which no more knowledge than the exponential decay of the luminosity can be incorporated, and on lineal scale in planet deconvolution problems (see [?]).

Assuming a toroidal edge correction, the eigenvalues of the matrix $I-\phi N$ are $\lambda_{ij} = 1 - 2\phi(\cos(2\pi i/m) + \cos(2\pi j/n)),\ i = 1,2,\ldots,m,\ j = 1,2,\ldots,n$. So for ϕ just less than 0.25, it can be shown that the density of $\mathbf{f}$ has the form

$$p_\alpha(\mathbf{f}) = \frac{1}{Z_{prior}(\alpha)} \exp\{-\frac{1}{2}\alpha\,\mathbf{f}^t(I-C)\mathbf{f}\}$$

where $Z_{prior}(\alpha) = (\prod_{i,j\neq 0,0} \lambda_{ij})^{-1/2}(2\pi/\alpha)^{p/2}$, $C = \phi N$.

We shall also use the simultaneous autorregresive model (SAR). This model is characterized by

$$p_\alpha(\mathbf{f}) = \frac{1}{Z_{prior}(\alpha)} \exp\{-\frac{1}{2}\alpha\,\mathbf{f}^t(I-C)^2\,\mathbf{f}\}$$

where, for this prior, we have $Z_{prior}(\alpha) = (\prod_{i,j\neq 0,0} \lambda_{ij}^2)^{-1/2}(2\pi/\alpha)^{p/2}$.

It is important to note that these two choices of the image models correspond to setting contraints on the first or second derivatives in the regularization framework. Furthermore, α/β correspond to the regularization parameter λ.

A typical procedure to estimate the unknown parameters is to select $\hat{\alpha}$ and $\hat{\beta}$ as the maximum likelihood estimates, *mle*, of α and β from $p_{\alpha,\beta}(\mathbf{g})$. This is the estimation method we shall use in our examples.

The mle estimation of α and β satisfies (see [?] for details)

$$\begin{aligned} S(\mathbf{f}_{mp(\alpha,\beta)}) + trace[Q(\alpha,\beta)^{-1}A] &= p/\alpha, \quad (14) \\ \| \mathbf{g} - \mathbf{D}\mathbf{f}_{mp(\alpha,\beta)} \|^2 + trace[Q(\alpha,\beta)^{-1}\mathbf{D}^t\mathbf{D}] &= p/\beta. \quad (15) \end{aligned}$$

where

$$Q(\alpha,\beta) = \alpha A + \beta \mathbf{D}^t\mathbf{D}$$

and $A = (I-C)$ or $A = (I-C)^2$ depending on whether we use the CAR or the SAR model for $\mathbf{f}$.

If we use the EM-algorithm to find a stationary point of $p(\mathbf{g}|\alpha,\beta)$, with $\mathcal{X} = (\mathbf{f},\mathbf{g})^t$ and $\mathcal{Y} = \mathbf{g} = \mathcal{T}\mathcal{X}$, as the complete and incomplete data, respectively, we obtain the following iterative procedure for α and β (see [?] for details)

$$\begin{aligned} \alpha_{i+1}^{-1} &= \{S(\mathbf{f}_{mp(\alpha_i,\beta_i)}) + trace[Q(\alpha_i,\beta_i)^{-1}A]\}/p, \quad (16) \\ \beta_{i+1}^{-1} &= \{\| \mathbf{g} - \mathbf{D}\mathbf{f}_{mp(\alpha_i,\beta_i)} \|^2 + trace[Q(\alpha_i,\beta_i)^{-1}\mathbf{D}^t\mathbf{D}]\}/p. \quad (17) \end{aligned}$$

scheme in our examples.

4 The Hierarchical Bayesian Approach

Together with the ingredients of the empirical approach, the hierarchical paradigm defines a hyperprior on the hyperparameters where information about those hyperparameters is included.

Although in some cases it would be possible to know, from previous experience, relations between the hyperparameters, we shall study here the model where the global probability is defined as

$$p(\alpha, \beta, \mathbf{f}, \mathbf{g}) = p(\alpha)p(\beta)p(\mathbf{f}|\alpha)p(\mathbf{g}|\mathbf{f}, \beta), \tag{18}$$

where $p_\alpha(\mathbf{f})$ now becomes $p(\mathbf{f}|\alpha)$ and $p_\beta(\mathbf{g}|\mathbf{f})$ is written as $p(\mathbf{g}|\mathbf{f}, \beta)$.

We shall use the improper noninformative priors $p(\alpha) \propto 1/\alpha$ and $p(\beta) \propto 1/\beta$ both over $[0, \infty)$. It is also possible to use, instead, an improper noninformative hyperprior on linear scale but since we have so many data the results are very similar. Other posibilities are described in [?].

After having defined $p(\alpha, \beta, \mathbf{f}, \mathbf{g})$, the hierarchical Bayesian analysis is performed by marginalizing out α and β to draw conclusions from $p(\mathbf{f}|\mathbf{g}) \propto p(\mathbf{f}, \mathbf{g})$.

For our problem we have

$$p(\mathbf{f}, \mathbf{g}) \propto \int_\alpha \int_\beta \frac{1}{Z_{prior}(\alpha)} \frac{1}{Z_{noise}(\beta)} \exp\{-\frac{1}{2}[\alpha S(\mathbf{f}) + \beta \parallel \mathbf{g} - \mathbf{D}\mathbf{f} \parallel^2]\}\, d\alpha\, d\beta$$

where $S(\mathbf{f}) = \mathbf{f}^t(I - C)\mathbf{f}$ or $S(\mathbf{f}) = \mathbf{f}^t(I - C)^2\,\mathbf{f}$ depending on whether we use the CAR or SAR model. Using the well known result from the gamma distribution,

$$\int_0^\infty y^{u-1} e^{-ay} dy = \Gamma(u) a^{-u}$$

we have

$$p(\mathbf{f}, \mathbf{g}) \propto S(\mathbf{f})^{-p/2}\, [\parallel \mathbf{g} - D\mathbf{f} \parallel^2]^{-p/2}$$

So to obtain the restoration we would just minimize

$$p \log S(\mathbf{f}) + p \log \parallel \mathbf{g} - \mathbf{D}\mathbf{f} \parallel^2 \tag{19}$$

This equation suggests the following iterative scheme to find the restoration. Starting at step 0 with $\mathbf{f}^0 = \mathbf{g}$, we use at step i of our iterative procedure $\alpha_{i-1}^{-1} = S(\mathbf{f}^{i-1})/p$ and $\beta_{i-1}^{-1} = \parallel \mathbf{g} - \mathbf{D}\mathbf{f}^{i-1} \parallel^2 /p$ and then define $\mathbf{f}^i$ as

$$\mathbf{f}^i = arg\, \{\min_{\mathbf{f}}\, \alpha_{i-1} S(\mathbf{f}) + \beta_{i-1} \parallel \mathbf{g} - \mathbf{D}\mathbf{f} \parallel^2 \}$$

We shall use this iterative scheme in our examples. Proof of convergence is given in [?].

5 Test Examples

To compare the performance of the methods for choosing the unknown parameters in a restoration problem we shall use the cameraman image. The digitalized image on a 256 × 256 grid with grey levels between 0 and 255 was blurred with the point spread function

$$D(r) \propto (1 + \frac{r^2}{R^2})^{-\gamma} \tag{20}$$

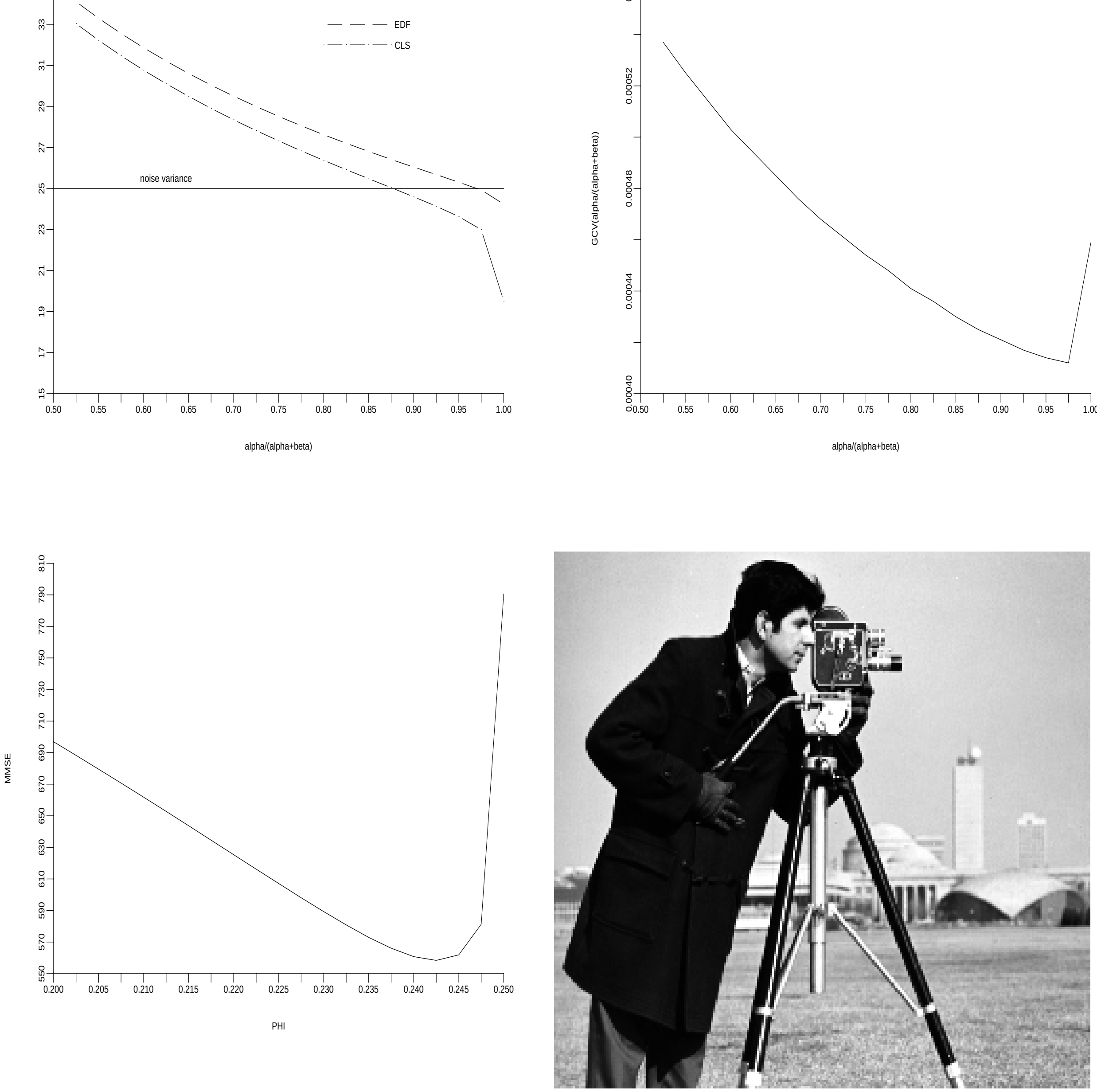

Figure 1: From left to right and top to bottom. (a) Curves of CLS(λ) and EDF(λ). (b) Curve of GCV(λ) . (c) MMSE of the hierarchical model restorations depending on ϕ. (d) Original cameraman image.

with $R = 5$ and $\gamma = 3$ and r denoting the distance to the source. Noise of variance 25 was added. The signal to noise ratio was 23.86 dBs, computed as $10 \log_{10} \frac{\sigma^2}{\sigma_n^2}$, where $\sigma^2 = \frac{1}{p} \sum_i (\mathbf{f}_i - \bar{\mathbf{f}})^2$ and σ_n^2 is the noise variance.

Figure 1 despicts, in (a), $RSS(\lambda)/p$ and $RSS(\lambda)/EDF(\lambda)$ and, in (b), $GCV(\lambda)$ for several values of λ. The optimal values were $\lambda_{CLS} = 0.143$, $\lambda_{EDF} = 0.0257$ and $\lambda_{GCV} = 0.0257$, which correspond to ratios $\frac{\alpha}{\alpha+\beta}$ of 0.875, 0.975 and 0.975 respectively. Figure 1(c) despicts the MMSE (defined as $\frac{1}{p} \parallel \mathbf{f} - \hat{\mathbf{f}} \parallel^2$) of the hierarchical model restorations depending on ϕ. The corresponding ratios $\frac{\alpha}{\alpha+\beta}$ for the MLE and hierarchical model were 0.98 and 0.974, respectively. The original cameraman image is also shown in figure 1(d).

The distorted image and the restorations produced by the CLS, EDF and hierarchical methods to select the unknown parameters are displayed in figure 2. We used a CAR prior model. Only 10 iterations were needed for the hierarchical algorithm to converge. MLE algorithm needed 75 iterations. As can be observed, the hierarchical model produces sharper restorations. The problem of estimating ϕ has been adressed in [**?**] using a different framework.

This work has been supported by the "Comisión Nacional de Ciencia y Tecnología" under contract TIC91-0660.

Figure 2: From left to right and top to bottom. (a) Blurred ($R = 5$) and noisy ($\sigma^2 = 25$) cameraman image. (b) Restoration with the CLS model. (c) Restoration with the EDF model. (d) Restoration with the hierarchical model.