CHOOSING THE REGULARIZATION PARAMETER IN IMAGE RESTORATION

D. M. Titterington
University of Glasgow

ABSTRACT

Many procedures in statistical image restoration can be regarded as regularization techniques involving a scalar smoothing parameter. The paper collates several methods for choosing the smoothing parameter, including model-based maximum likelihood. Minimum risk, generalized cross-validation, choice based on fit to the data, and "equivalent degrees of freedom" choice. Some theoretical and empirical comparisons are summarized.

Keywords: Image restoration; ill-posed inverse problems; cross-validation; ridge regression; regularization; smoothing.

1. Introduction to the image restoration problem

The basic structure of our version of the image restoration problem is as follows. A vector x, of length N, contains the description of a true scene made up of N pixels. Thus, the ith element of x, x_i, provides the true label, color, grey-level or intensity of pixel i, $i = 1, \ldots, N$.

Although x is the quantity of direct interest, it is not directly available. Instead, our data consist of a vector y, the observed record of the true scene. Typically y and x are not the same, y being a blurred and noisy version of x. The image restoration problem is, therefore, to construct some "estimate" of x on the basis of y and any other relevant information. Such a restoration of x will be denoted, generically, by $\hat{x}$.

2. The need for regularization

Statistical image restoration is a model-based activity and we motivate the general discussion of this section by a simple example.

Example: Additive linear model

Suppose x and y are related by

$$y = Hx + \varepsilon,$$

where H is an $N \times N$ matrix and ε represents a vector of noisy disturbances. Thus, H represents the systematic blur (or point-spread matrix) imposed by the sensor. Its elements are usually assumed to be known, and we shall follow this assumption here. For simplicity, we assume that ε represents a Gaussian white noise vector; that is,

$$\varepsilon \sim N(0,\ \sigma^2 I),$$

where 0 is a vector of zeroes, I is an identity matrix and σ^2 may or may not be known.

In the statistical literature, H might be called the design matrix and a natural estimator of x, based on y, is the least-squares estimator (or "direct deconvolution")

$$\hat{x}_0 = H^{-1}y = x + H^{-1}\varepsilon. \tag{1}$$

The motivation for the subscript on $\hat{x}_0$ will become clear later. Equation (1) indicate that $E(\hat{x}_0) = x$, a good property, but that cov $\hat{x}_0 = \sigma^2 H^{-1}(H^T)^{-1}$. The latter could be disastrous if H is ill-posed and unfortunately this is often the case in image restoration, where the dimension of H, indicated by N, is very large.

In this respect, we are being confronted with just one of a very wide range of so-called ill-posed inverse problems. To motivate a general approach to counteract this numerical instability, note that we can interpret the least-square estimators as the solution of the optimization problem

$$\min_x \Delta(y, x), \tag{2}$$

where

$$\Delta = (y - Hx)^T(y - Hx) = \|y - Hx\|^2.$$

Note the following remarks.

(i) $\Delta(y, x)$ represents a measure of "distance" between the data, y, and the scene, x.

(ii) Alternatively, $\Delta(y, x)$, as x varies, gives a measure of lack of fidelity of the restoration to the data.

(iii) $\Delta(y, x)$ is equivalent, as a function of x, to $-\log p(y|x)$, where $p(y|x)$ denotes the probability density of y, given x. Thus, if x is regarded as a set of parameters. $\hat{x}_0$ represents the maximum likelihood estimator.

The unsatisfactory nature of $\hat{x}_0$ can be ameliorated by solving, not (2), but

$$\min\{\Delta(y, x) + \beta\Phi(x)\}, \tag{3}$$

where $\Phi(x)$ is some penalty for "roughness" and $\beta > 0$ is a scalar. This constitutes the method of *regularization*: β is a regularization or smoothing parameter whose value dictates the trade-off between fidelity to the data and smoothness. Construction of a specific restoration involves the choice of Φ, which dictates the manner of smoothing, and of β, which determines the degree of smoothing.

The regularization prescription appears in a wide variety of problems in statistical smoothing and image restoration (Titterington, 1985a, 1985b). The following versions are of particular relevance here.

(i) If $\Delta(y, x) = -\log p(y|x)$ and $\beta\Phi(x) = -\log p(x)$, where $p(x)$ is some marginal density for x, then the minimizer of (3) is the maximum a posterior estimator of x (Geman and Geman, 1984).

(ii) If $\Delta(y, x)$ is quadratic in x, as in the Example, and if $\Phi(x) = x^T C x$, where C is a non-negative definite matrix, then (3) leads to a ridge-regression estimator of x. For the Example as stated, the formula is

$$\hat{x}_\beta = (H^T H + \beta C)^{-1} H^T y,$$

assuming that the inverse exists. Note how the subscript notation $\hat{x}_\beta$ fits in with the earlier case based on zero smoothing.

(iii) If the elements of x are positive and $p_i = x_i / \Sigma x_j$, then the maximum entropy method of Gull and Skilling (1984) is revealed as a special case with

$$\Phi(x) = \Sigma p_i \log p_i.$$

In this paper, we concentrate on the crucial choice of the smoothing parameter, β.

3. General methods for choosing the smoothing parameter

We consider, in the following subsections, several systematic approaches to the choice of β.

(i) Modelling approach

(ii) Minimum risk approach

(iii) Generalized cross-validation

(iv) Choice based on fit to the data

(v) Equivalent degrees of freedom approach.

3.1 Modelling approach

To motivate this method, it is helpful to regard the image restoration problem as a missing-data problem in which, corresponding to the available data, y, there are *complete* data, x. Given a probabilistic structure, in the form of $p(y|x)$ and $p(x)$, with β interpreted as a parameter within the latter, it is natural to base inferences about β on the likelihood corresponding to the observed data. This is, therefore, $p(y) = p(y|\beta)$, where

$$p(y|\beta) = \int p(x, y|\beta)dx$$
$$= \int p(y|x)p(x|\beta)dx$$

where explicit indication is now made of the dependence of $p(x)$ on β and where the integral would be replaced by a summation in discrete problems.

In view of the "incomplete data" interpretation of the problem, a maximum likelihood estimate of β might be computed using the EM algorithm of Dempster, Laird and Rubin (1977). This algorithm generates a sequence $\{\beta^{(r)}\}$ of values by repeating the following double-step.

E-**step**: Given $\beta^{(r)}$, evaluate $H_r(\beta) = E\{\log p(x, y|\beta)|y, \beta^{(r)}\}$.

M-**step**: Choose $\beta = \beta^{(r+1)}$ to maximize $H_r(\beta)$.

The convergence properties of $\{\beta^{(r)}\}$ may or may not be straightforward.

One can dress up this approach in Bayesian terminology. If $p(x|\beta)$ is regarded as a "prior" density for x, then the above method of choosing β constitutes an *empirical Bayes* technique.

If the elements of x are label indicators associated with a finite palette of colors, then the problem can be interpreted as a variation of the analysis of *mixture* data: see Titterington, Smith and Markov (1985).

Were $p(x)$ such that the elements x_i are independently and identically distributed and were the y_i also independent, conditionally on x_i, then the data would be standard mixture data, for which both E- and M-steps are straightforward and explicit. However, although the second assumption is plausible, $p(x)$ typically does not factorize in plausible image models, because of natural, local association among the x_i.

In the case of one-dimensional images, one natural model assumes that the x_i follow a Markov Chain. Maximum likelihood analysis of such a so-called "hidden" Markov Chain is described by Baum et al (1970) and by Pickett and Whiting (1987). For this case the E and M-steps are often explicit, but the E-step is complex, involving forward and backward recursions over the set of pixels.

In the obvious generalization to two-dimensional images, Markov Random Field models are adopted. Now, however, the E-step is computationally intractable and even the M-step is non-trivial; see Chalmond (1988) and, for a treatment of the Markov Mesh model, Devijver (1988). In practice, analytical computation is substituted by simulation, based on the Gibbs Sampler of Geman and Geman (1984). Alternatively, maximum likelihood estimation is abandoned in favour of the method of moments (Chalmond, 1988) or other techniques (Derin and Elliott, 1987, Possolo, 1986).

3.2 Minimum risk approach

The motivation here is to find a restoration that is "as close as possible to the true scene" in some sense. If $\widehat{x}_\beta$ denotes the restoration, then one might choose β to solve

$$\min_\beta E_{y|x}\delta(x, \widehat{x}_\beta),$$

where δ is a measure of distance, now to be interpreted as a loss function. Examples of this general approach are those of minimum total mean-squared error (TMSE) and minimum total prediction mean-squared error (TPMSE).

Recalling our criticism of the unstable, but unbiased least-squares estimator in the example, we see that these risk criteria should achieve some sensible trade-off between bias and "variance".

Example: Additive linear model

Suppose $\Phi(x) = x^T Cx$. Then

$$\widehat{x}_\beta = (H^T H + \beta C)^{-1} H^T y$$

and the TMSE is

$$\text{TMSE}(\beta) = \|\{(H^T H+\beta C)^{-1} H^T H - I\}x\|^2 + \sigma^2 \text{tr}\{(H^T H+\beta C)^{-1} H^T H (H^T H+\beta C)^{-1}\}.$$

Similarly,

$$\text{TPMSE}(\beta) = E\|y - K(\beta)y\|^2$$

where $K(\beta) = H(H^T H + \beta C)^{-1} H^T$. Thus,

$$\text{TPMSE}(\beta) = \|\{I - K(\beta)\}Hx\|^2 + \sigma^2 \text{tr}\{(I - K(\beta))^2\}.$$

The practical difficulty with this approach is the dependence of TMSE (β) and TPMSE (β) on x. In practice a preliminary estimate, $\widetilde{x}$, might be substituted at this point or the criterion function might be averaged, using some weighting measure on the space of x, to create a minimum Bayes risk choice for β.

3.3 Generalized cross-validation (GCV)

The method discussed here avoids the dependence on x experienced in Section 3.2. The motivation now is to choose β to optimize average prediction of individual observed values using all the remaining data. This leads to a criterion function of the form CV(β), where

$$\text{CV}(\beta) = N^{-1} \sum_{i=1}^{N} \delta(y_i, E(y_i|\widehat{x}_\beta^{(i)})),$$

in which $\hat{x}_\beta^{(i)}$ is the restoration computed on the basis of all observations except for x_i. The *generalized* cross-validation function GCV(β) is a slight variation on this and is best illustrated by example.

Example: Additive linear model

If RSS (β) denotes the residual sum of squares and we use a simple quadratic loss for δ, then

$$\mathrm{RSS}(\beta) = \|\{I - K(\beta)\}y\|^2$$

and

$$\mathrm{GCV}(\beta) = \mathrm{RSS}(\beta)/[\mathrm{tr}\{I - K(\beta)\}^2].$$

Note that GCV(β) depends only on β and the data. Generalized cross-validatory choice leads to $\beta = \hat{\beta}_{\mathrm{GCV}}$ to minimize GCV(β). It turns out that $\hat{\beta}_{\mathrm{GCV}}$ has asymptotic optimality properties, in a certain sense, in some problems including many versions of the present example; see Golub, Heath and Wahba (1979).

3.4 Choice based on fit to the data

Suppose, for some measure of distance, δ, $\delta(y, x)$ is regarded as a measure of goodness-of-fit and that, conditional on x, $\delta(y, x)$ has a probability distribution, F_δ. Suppose that $q(F_\delta)$ denotes some measure of the location of F_δ, such as the mean or some percentile. Then the present approach selects β to satisfy

$$\delta(y, \hat{x}_\beta) = q(F_\delta).$$

Example: Additive linear model

Find β to satisfy

$$\mathrm{RSS}(\beta) = N\sigma^2.$$

(The justification for this, of course, is that $E\|y - Hx\|^2 = N\sigma^2$.) If the errors of observation are Gaussian, then $\|y - Hx\|^2 \sim \chi^2(N)$. As a consequence, we call the resulting β, $\hat{\beta}_{\mathrm{CHI}}$.

The criticisms of this approach are that σ^2, or an estimate thereof, is required in order to implement the method, and that $\hat{\beta}_{\mathrm{CHI}}$ tends to oversmooth; see Hall and Titterington (1986, 1987), for theoretical and empirical evidence to this effect. The method has, however, been very popular in the regularization literature.

3.5 Empirical degrees of freedom (EDF) choice

Example: Additive linear model

We choose $\beta = \hat{\beta}_{\mathrm{EDF}}$ to solve

$$\mathrm{RSS}(\beta) = \sigma^2 \mathrm{tr}\{I - K(\beta)\}.$$

The motivation here is twofold:

(i) Wahba (1983) suggests using

$$\mathrm{RSS}(\hat{\beta}_{\mathrm{GCV}})/\mathrm{tr}\{I - K(\hat{\beta}_{\mathrm{GCV}})\}$$

as an estimator of σ^2;

(ii) the method imposes less smoothing than does $\hat{\beta}_{\mathrm{CHI}}$.

As in the case of $\hat{\beta}_{\mathrm{CHI}}$, however, a value for σ^2 must be available.

4. References to theoretical comparisons

In this short section we make brief reference to previous work on theoretical comparisons among some of the above methods. Most of the material refers to the Additive Linear Model, in which

$$\hat{x}_\beta = (H^T H + \beta C)^{-1} H^T x.$$

Hall and Titterington (1987) investigated deterministic versions of the minimum risk, CHI and EDF methods, to compare the degrees of smoothing thereby imposed. Three values of β were defined, as follows:

β_{TP} : minimizer of TPMSE (β)

β_{CHI} : solution of E RSS $(\beta) = N\sigma^2$

β_{EDF} : solution of E RSS $(\beta) = \sigma^2 \mathrm{tr}\{I - K(\beta)\}$.

Example: Special case with $H = C = I$

In this case, $\beta_{\mathrm{EDF}} = \beta_{\mathrm{TP}} = r^{-1}$, where

$$r = x^T x/(N\sigma^2),$$

a signal-to-noise ratio, and

$$\beta_{\mathrm{CHI}} = \beta_{\mathrm{EDF}}^{1/2} + o(\beta_{\mathrm{EDF}}^{1/2}).$$

The summary of this is that, if the signal-to-noise ratio is large, so that β_{EDF} is small, then β_{CHI} is less small, thereby illustrating the over-smoothing characteristics of β_{CHI}.

One way of illustrating the differential effects of the methods is to evaluate $\mathrm{tr}K(\beta)$ which, if regarded as the number of degrees of freedom associated with the smoothed fit, measures the complexity of the fitted regression. For this example, the following results obtain (Titterington, 1986).

r	$\mathrm{tr}K(\beta_{\mathrm{EDF}})$	$\mathrm{tr}K(\beta_{\mathrm{CHZ}})$
1	$0.5N$	$0.29N$
3	$0.75N$	$0.5N$
8	$0.89N$	$0.67N$

Hall and Titterington (1987) also look at the case of a class of periodic smoothing splines and show that

$$\beta_{\mathrm{TP}}/\beta_{\mathrm{EDF}} = 0(1).$$

Their asymptotic results for the additive linear model are extended by Kay (1988).

In Hall and Titterington (1986), models are formulated for second-order, stationary, one-dimensional images and a variety of restoration procedures are described. In a comparison between the corresponding β_{EDF} and β_{CHI}, it is shown that, if the signal-to-noise ratio is large, restorations using β_{EDF} fit the data "more closely", in a well-defined sense, than do those with β_{CHI}. Equivalently, β_{CHI} smooths more strongly than does β_{EDF}.

5. Empirical comparisons

As in the previous section, we restrict ourselves to providing references to more extensive descriptions of numerical studies and to giving a more flavor of the main trends revealed therein.

Hall and Titterington (1987) report a simulation study based on a periodic regression function, fitted using periodic smoothing splines. Figures therein, in particular, show that, for this example, cross-validatory choice produces a fitted curve that follows the true regression fairly well. The curves corresponding to $\widehat{\beta}_{\mathrm{EDF}}$ and $\widehat{\beta}_{\mathrm{CHI}}$ are progressively smoother.

A very detailed study in Thompson, Brown, Kay and Titterington (1988) compares β_{TP}, $\widehat{\beta}_{\mathrm{CHI}}$, $\widehat{\beta}_{\mathrm{GCV}}$ and $\widehat{\beta}_{\mathrm{EDF}}$ on a very simple, piecewise-constant, 47-pixel image defined by

$$\begin{aligned} f_i &= 100, \quad i = 1, \ldots, 14 \\ &= 50, \quad i = 15, \ldots, 36 \\ &= 100, \quad i = 37, \ldots, 47. \end{aligned}$$

The only non-zero elements in H (assumed known) were

$$\begin{aligned} &h_{ii} = 0.6, && i = 1, \ldots, N(= 47) \\ &h_{i,i+1} = h_{i+1,i} = 0.2, && i = 1, \ldots, N-1 \\ &h_{1N} = h_{N1} = 0.2. \end{aligned}$$

The noise variance was $\sigma^2 = 100$ and first order regularization was imposed, so that C had non-zero elements given by

$$\begin{aligned} &C_{ij} = 2 && i = 1, \ldots, N-1 \\ &C_{i,i+1} = C_{i+1,i} = 1, && i = 1, \ldots, N-1 \\ &C_{11} = C_{NN} = 1. \end{aligned}$$

Altogether, 1000 realizations of the observed image were created. Among the many comparative exercises carried out by Thompson, Brown, Kay and Titterington (1988) we restrict attention here mainly to one aspect to performance, namely, the effectiveness in recovering the true scene.

As criterion functions, we use measures of average (per pixel, per simulation) squared errors, and partitions of this into components measuring bias and variance.

Suppose from the rth simulation, a restoration $\hat{x}^{(r)}$ results. (For clarity, the dependence on β is omitted.) Then the indicators of bias, $\hat{B}$, and variance, $\hat{V}$, are given by

$$\hat{B} = N^{-1}\|\overline{x} - x\|^2$$

and

$$\hat{V} = (NR)^{-1}\sum_{r=1}^{R}\|\hat{x}^{(r)} - \overline{x}\|^2,$$

where $R = 1000$ and

$$\overline{x} = R^{-1}\sum_{r=1}^{R}\hat{x}^{(r)}.$$

For this problem the "optimal" β_{TP} takes the value $\beta_{\mathrm{TP}} = 0.928$.

All methods except for $\hat{\beta}_{\mathrm{GCV}}$ rely on knowledge of σ^2. The use of purely data-based estimates of σ^2 is currently under investigation. Here, we provide results based on both σ^2 and σ_D^2, where σ_D^2 is the estimator of σ^2 based on the true residuals associated with a given set of data. In other words, given y and x,

$$\sigma_D^2 = N^{-1}\|y - Hx\|^2.$$

The value obtained for $\hat{V}$, $\hat{B}$ and $\hat{B} + \hat{V}$ are provided in the following table.

TABLE

Comparison of performance among $\beta_{\mathrm{TP}}, \hat{\beta}_{\mathrm{CHI}}, \hat{\beta}_{\mathrm{EDF}}$ and $\hat{\beta}_{\mathrm{GCV}}$.

	Method	$\hat{V}$	$\hat{B}$	$\hat{B}+\hat{V}$
σ	TP	27.3	27.7	55.0
	CHI	19.2	49.7	68.9
	EDF	57.4	23.6	81.0
	GCV	102.0	18.1	120.1
σ_D	TP	27.0	27.6	54.6
	CHI	16.3	50.5	66.8
	EDF	39.6	23.5	63.1

Several interesting points arise from the table.

(i) Although β_{TP} is unattainable in practice, it is included as a benchmark.

(ii) The "over-smoothing" imposed by $\hat{\beta}_{\mathrm{CHI}}$ is betrayed by the large bias contribution.

(iii) EDF choice performs quite well, particularly in terms of bias, if a value of σ^2 can be used that is appropriate to the individual realizations. This provides further encouragement for the search for good, data-based, estimators of σ^2.

(iv) GCV is also good, so far as bias is concerned, but $\widehat{V}$ is disappointing. This is very largely due to a few (about 5%) realizations in which GCV results in grossly undersmoothed restorations. This and other awkward features of GCV are discussed in detail in Thompson, Kay and Titterington (1988).

Acknowledgements

It is hoped that the present exposition makes clear the importance of the collaborative influence of P. Hall, J. C. Brown, J. W. Kay and A. M. Thompson, of whom the last-mentioned is supported by a research grant from the U.K. Science and Engineering Research Council.

References

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains. *Ann. Math. Statist., 41*, 164–171.

Chalmond, B. (1988). An iterative Gibbsian techniques for simultaneous structure estimation and reconstruction of M-ary images. *Preprint 88–28*, Mathematiques, Univ. Paris-Sud.

Dempster, A. P., Laird, T. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the *EM* algorithm (with discussion). *J. R. Statist. Soc. B, 39*, 1–38.

Derin, H. and Elliott, H. (1987). Modelling and segmentation of noisy and textured images using Gibbs random fields. *IEEE. Trans. Pattern Anal. Machine Intell., PAMI-9*, 39–55.

Devijver, P. A. (1988). Image segmentation using causal Markov random field models. *Lecture Notes Comput. Sci., 301*, 131–143.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell., PAMI-6*, 721–741.

Geman, S. and McClure, D. C. (1985). Bayesian image analysis: An application to single photon emission tomography, *Proc. ASA Statist. Comput. Section*, 12–18.

Gull, S. and Skilling, J. (1984). Maximum entropy method in image processing. *Proc. IEE, 131*, F. 646–659.

Golub, G., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*, 215–223.

Hall, P. and Titterington, D. M. (1986). On some smoothing techniques used in image restoration, *J. R. Statist. Soc. B, 48*, 330–343.

Hall, P. and Titterington, D. M. (1987). Common structure of techniques for choosing parameters in regression problems. *J. R. Statist. Soc. B, 49*, 184–198.

Kay, J. W. (1988). Asymptotic comparison factors for smoothing parameter choices in linear regression. Submitted for publication.

Pickett, E. E. and Whiting, R. G. (1987). On the estimation of probabilistic functions of Markov Chains. *Lecture Notes in Economics and Math. Systems., 297*, Springer.

Possolo, A. (1986). Estimation of binary Markov random fields. *Tech. Rep. 77*, Dept. Statist., Univ. Washington, Seattle.

Thompson, A. M., Brown, J. C., Kay, J. W. and Titterington, D. M. (1988). A comparison of methods of choosing the smoothing parameter in image restoration by regularization. Submitted for publication.

Thompson, A. M. Kay, J. W. and Titterington, D. M. (1989). A cautionary note about cross-validatory choice. J. Statist. Comput. Simul., 33, 199–216.

Titterington, D. M. (1985a). General structure of regularization procedures in image processing. *Astron. Astrophys., 144*, 381–387.

Titterington, D. M. (1985b). Common structure of smoothing techniques in statistics. *Int. Statist. Rev., 53*, 141–170.

Titterington, D. M. (1986). Comments on a paper by F. O'Sullivan. *Statist. Sci., 1*, 519–521.

Titterington, D. M., Smith, A. F. M. and Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc. B., 45*, 133–150.

Younes, L. (1988a). Estimation and annealing for Gibbsian fields. *Ann. Inst. Henri Poincare*, to appear.

Younes, L. (1988b). Parametric inference for imperfectly observed Gibbsian fields. *Preprint 99–17*, Mathematiques, Univ. Paris-Sud.