ROBUST PRIORS FOR SMOOTHING AND IMAGE RESTORATION

by

Hans R. Künsch
Seminar für Statistik
ETH-Zentrum
CH-8092 Zürich

Research Report No. 67a
May 1993

Seminar für Statistik
Eidgenössische Technische Hochschule (ETH)
CH-8092 Zürich
Switzerland

ROBUST PRIORS FOR SMOOTHING AND IMAGE RESTORATION

Hans R. Künsch

Seminar für Statistik

ETH-Zentrum

CH-8092 Zürich

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

13 May 1993

Abstract

The Bayesian method for restoring an image corrupted by added Gaussian noise uses a Gibbs prior for the unknown clean image. The potential of this Gibbs prior penalizes differences between adjacent grey levels. In this paper we discuss the choice of the form and the parameters of the penalizing potential in a particular example used previously by Ogata (1990). In this example the clean image is piecewise constant, but the constant patches and the step sizes at edges are small compared with the noise variance. We find that contrary to results reported in Ogata (1990) the Bayesian method performs well provided the potential increases more slowly than a quadratic one and the scale parameter of the potential is sufficiently small. Convex potentials with bounded derivatives perform not much worse than bounded potentials, but are computationally much simpler. For bounded potentials we use a variant of simulated annealling. For quadratic potentials data-driven choices of the smoothing parameter are reviewed and compared. For other potentials the smoothing parameter is determined by considering which deviations from a flat image we would like to smooth out and retain respectively.

Keywords: Gibbs distribution, Gaussian and non-Gaussian smoothness priors, maximum a posteriori estimation, images with discontinuities, simulated annealing.

1 Introduction

Since the pioneering paper by Geman and Geman (1984) there has been much interest in the Bayesian approach to image analysis. The basic procedure is easy to understand. It contains three ingredients: A Gibbs random field model as the prior, a model for image formation consisting typically of blur, degradation and superposition of noise, and Bayes formula to obtain the posterior given the image on which the analysis is based. Still the implementation is complicated by problems like choosing the actual form and the parameters of the prior or the computational difficulties. The aim of this paper is to discuss some of these problems in one concrete example since the tools for a general theoretical analysis seem to be lacking at this moment. The example we have chosen comes from a recent paper by Ogata (1990) where a Monte Carlo method to estimate parameters of a Gibbs prior has been proposed. The clean image in this example is piecewise constant, and the task is to remove added noise without smoothing the edges of the clean image. Since there is no blurring or degradation in the image, this problem is a classical two-dimensional discrete smoothing problem. One can distinguish at least the following classes of priors: quadratic potentials, convex potentials with bounded derivatives, potentials with derivatives redescending to zero and inclusion of an unobservable edge process (Geman and Geman, 1984). Among these classes both the computational difficulties and the potential ability to produce the desired restorations increase. Hence one would like to understand how much is actually gained by using a computationally more demanding prior. The advantage of non-quadratic potentials for these types of problems has been stressed by several authors, e.g. Geman and McClure (1987), Besag (1989), Green (1990), Geman and Reynolds (1992), and Kitagawa (1987) in a one-dimensional setting. The differences between convex and non-convex potentials have not been investigated as far as I know. Ogata (1990) reported that a logarithmic potential did not produce better results than a quadratic one in this example. As a robustnik I believed stronlgy in the advantages of heavy-tailed models and I wanted to know the limits of the methods. Therefore I took up the same example again. This paper contains the main results which were quite surprising in many respects.

In Section 2 we give a precise formulation of our study. In Section 3 we show that for quadratic potentials the computations greatly simplify, allowing us for instance to compare several methods for choosing the smoothing parameter. In Section 4 we discuss convex potentials with bounded derivatives. We show that they produce decent restorations for suitably chosen parameters. In Section 5 we discuss bounded potentials. We show that simulated annealing is able to come very close to the optimal restoration. Parameters are chosen by an ad-hoc argument requiring the knowledge of the noise variance and a prototype of a clean image. In Section 6 we briefly consider two additional clean images in order to see how well the same potentials and parameters perform when the clean image is somewhat different from this prototype. Again convex potentials with bounded derivative produce good restorations. The results are summarized in Section 7.

2 Statement of the Problem

Suppose that a clean image $(\theta^0_{ij}; 1 \leq i,j \leq n)$ is corrupted by additive Gaussian white noise $(\varepsilon_{ij}, 1 \leq i,j \leq n)$, i.i.d. $\sim \mathcal{N}(0,\sigma^2)$. The problem is to estimate (θ^0_{ij}) from the corrupted image

$$y_{ij} = \theta^0_{ij} + \varepsilon_{ij}. \tag{2.1}$$

We consider here the example of Ogata (1990) where $n = 20, \sigma^2 = 1$ and θ^0 is the following step function

$$\theta^0_{ij} = \begin{cases} 1 & 1 \leq i \leq 10, \quad 1 \leq j \leq 10 \\ 2 & 1 \leq i \leq 10, \quad 11 \leq j \leq 20 \\ -1 & 11 \leq i \leq 20, \quad 1 \leq j \leq 10 \\ 0 & 11 \leq i \leq 20, \quad 11 \leq j \leq 20 \end{cases} \tag{2.2}$$

The method we are using is the maximizer of the posterior density (MAP) for a prior of the following form:

$$\pi(\theta) = Z^{-1} \exp(-\frac{1}{2\tau^2} \sum_{(i,j)\sim(k,\ell)} \phi\left((\theta_{ij} - \theta_{k\ell})/\delta\right)) \tag{2.3}$$

where $(i,j) \sim (k,\ell)$ means that the pixels (i,j) and (k,ℓ) are nearest neighbors. This form of the prior seems sufficiently general for the problem at hand. Because (θ^0_{ij}) is piecewise constant, there is no need to consider higher-order models (see Geman and Reynolds, 1992) where the difference $\theta_{ij} - \theta_{k\ell}$ would be replaced by linear combinations filtering out polynomials of degree one or two. Also because the edges are parallel to the axes, there seems to be no need to include diagonal terms, i.e. second nearest neighbors. As concerns the choice of ϕ, we are going to discuss the following three cases

$$\begin{aligned} \phi(x) &= x^2 && \text{(Gaussian)} \\ \phi(x) &= \begin{cases} x^2 & |x| \leq 1 \\ 2|x| - 1 & |x| \geq 1 \end{cases} && \text{(Huber)} \\ \phi(x) &= \begin{cases} x^2 & |x| \leq 1 \\ 1 & |x| \geq 1 \end{cases} && \text{(truncated Gaussian)} \end{aligned}$$

In the Huber case, we obtain as $\delta \longmapsto 0$, $\tau^2\delta \mapsto$ const. the L_1-case. In the truncated Gaussian case $\delta = 0$ gives $\phi(x) = 1_{[x \neq 0]}$ which has been proposed by Leclerc (1989). The Huber prior is a representative of the class of convex ϕ's with bounded derivative, whereas the truncated Gaussian prior represents the class of bounded ϕ's. We expect that our findings will generalize to these larger classes, at least qualitatively.

By Bayes formula the posterior of (θ_{ij}) given (y_{ij}) is

$$\pi(\theta|y) = Z(y)^{-1} \exp\left(-(2\sigma^2)^{-1} \sum_{i,j} (y_{ij} - \theta_{ij})^2 - (2\tau^2)^{-1} \sum_{(i,j)\sim(k,\ell)} \phi\left((\theta_{ij} - \theta_{k\ell})/\delta\right)\right) \tag{2.4}$$

Hence the MAP-estimator of (θ^0_{ij}) is obtained by minimizing

$$H(\theta|y) = \sum_{i,j}(y_{ij}-\theta_{ij})^2 + \beta \sum_{(i,j)\sim(k,\ell)} \phi((\theta_{ij}-\theta_{k\ell})/\delta) \tag{2.5}$$

where $\beta = \sigma^2/\tau^2$. An alternative to the MAP is the posterior mean $E[\theta_{ij}|y]$.
I do not think that the posterior mean is computationally simpler. When $\pi(\theta|y)$ is multimodal, the Markov chain used to simulate the posterior can take a very long time to switch from one mode to another. So the multimodal case poses problems also for computation of the posterior mean. Furthermore compared with the MAP the posterior mean has more difficulties to bring out edges clearly. This is because the location of an edge can vary among different restorations with nonnegligible posterior probability. Averaging over these restorations then blurs this edge. For these reasons we consider only the MAP.

The main difficulty in the present example is to obtain enough smoothness in those regions where (θ^0_{ij}) is constant without blurring the edges. One realization of (y_{ij}) which we are going to use in the sequel is given in Figure 1. The human eye can discover the larger of the two edges, but has difficulties with the smaller one. Ad-hoc techniques with moving robust filters are not satisfactory. Figures 2 and 3 show the result of taking the moving median in a 5×5 window and the moving shorth in a 7×7 window respectively. The shorth is the mean of the shortest intervall containing half of the data, see Rousseeuw and Leroy (1987, Chap.4). It was chosen as a robust estimator which has a small bias even when there is a large fraction of outliers on one side. The median performs poorly, and also the shorth fails to nicely reproduce the smaller edge. So there is a need for more sophisticated techniques.

3 The Gaussian Case

From (2.3) and (2.4) we see that with $\phi(x) = x^2$ both the prior and the posterior are Gaussian with mean zero. Denote the inverse covariance matrix of the prior by $\tau^{-2}A$, i.e.

$$\sum_{ij,k\ell} \theta_{ij} A_{ij,k\ell} \theta_{k\ell} = \sum_{i,j\sim k,\ell} (\theta_{ij}-\theta_{k\ell})^2. \tag{3.1}$$

The computation of the MAP is in this case greatly simplified by the following result giving the eigenvectors and eigenvalues of A.

Theorem 3.1 *The eigenvalues of the matrix A defined in (3.1) are*
$\lambda_{ij} = 2\,(2-\cos(\omega_i)-\cos(\omega_j))$ $(1 \le i,j \le n)$ *where* $\omega_i = \pi(i-1)/n$
and the eigenvector to λ_{ij} *has components* $e_{i,k}e_{j,\ell}$ $(1 \le k,\ell \le n)$ *where* $e_{i,k} = \sin(\omega_i k) - \sin(\omega_i(k-1))$ *for* $i > 1$ *and* $e_{1,k} \equiv 1$.

Proof: A can be written as a Kronecker-product

$$A = A_n^{(1)} \otimes I_n + I_n \otimes A_n^{(1)},$$

where $(A_n^{(1)})_{i,i+1} = (A_n^{(1)})_{i+1,i} = -1,\ (A_n^{(1)})_{11} = (A_n^{(1)})_{nn} = 1,$
$(A_n^{(1)})_{ii} = 2\ (1 < i < n),\ (A_n^{(1)})_{ij} = 0$ otherwise,
and I_n is the identity matrix of dimension n.

It is easily checked that the eigenvalues of $A_n^{(1)}$ are $2\,(1-\cos(\omega_i))$ with eigenvectors $(e_{i,k})_{1\le k\le n}$. Hence the result follows from standard facts about Kronecker products, see e.g. Bellman (1960), Chapter 12. □

Remark 3.1: The fact that eigenvalues of quadratic forms arising for Gaussian Markov random fields can be calculated without imposing toroidal boundary conditions has been observed in the stationary case by Speed (1978). One may ask if a similar result is available also for other choices of neighbors. Obviously we can take

$$A = A_n^{(1)} \bigotimes I_n + I_n \bigotimes A_n^{(1)} + \gamma A_n^{(1)} \bigotimes A_n^{(1)}.$$

This gives diagonal terms $(\theta_{ij} - \theta_{i\pm1,j\pm1})^2$, but also additional terms for nearest neighbor differences at the boundary. In other situations, the eigenvalues of A seem much harder to get in closed form.

Remark 3.2: Because of the free boundary conditions in (2.3), the (θ_{ij}) are not stationary. In the Gaussian case we can easily compute second moments of increments with the help of Theorem 3.1. For instance the variance of the difference between nearest neighbors, $\mathrm{Var}[(\theta_{ij}-\theta_{kl})^2]$ with $(i,j)\sim(k,l)$, is $0.70\tau^2$ for (i,j) at a corner, $0.64\tau^2$ for (i,j) and (k,l) both in the middle of one of the four boundaries, and $0.50\tau^2$ for (i,j) in the center of the square. Thus the θ_{ij}'s are more variable at the boundary than in the center. This property makes the model also appealing for field trials.

Theorem 3.1 allows us to compute the MAP-estimator according to the formula

$$\widehat{\theta} = D\ \mathrm{diag}\ \left((1+\beta\lambda_{\mathbf{i}})^{-1}\right) D^T y \tag{3.2}$$

where $\lambda_{\mathbf{i}} = \lambda_{ij}$ are the eigenvalues of A and the columns of D contain the normalized eigenvectors of A. The results look similar to Fig. 4 of Ogata (1990), so we do not show them here.

So the only remaining problem is the choice of the smoothing parameter $\beta = \sigma^2/\tau^2$. This is a long-studied problem and several proposals for data-dependent choices of β have been made in the literature, see e.g. Hall and Titterington (1986), Kay (1988), Wahba (1990). Some of them assume σ^2 to be known, others allow both σ^2 and τ^2 to be unknown. We will discuss briefly the following methods:

1. Minimizing estimated mean square error: It is easily checked that

$$\sigma^2 \sum_{\mathbf{i}} (1-\beta\lambda_{\mathbf{i}})/(1+\beta\lambda_{\mathbf{i}}) + \sum_{\mathbf{i}} (\beta\lambda_{\mathbf{i}}/(1+\beta\lambda_{\mathbf{i}}))^2\,(D^T y)_{\mathbf{i}}^2.$$

is an unbiased estimator of $\sum_{\mathbf{i}}(\widehat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}})^2$. Hence we can choose β by minimizing the above expression.

2. Variance tuning: By the law of large numbers $\sum \varepsilon_{\mathbf{i}}^2 \approx n^2\sigma^2$. The unknown errors $\varepsilon_{\mathbf{i}}$ can be replaced by the residuals $y_{\mathbf{i}} - \widehat{\theta}_{\mathbf{i}}$. This leads to estimating β by solving $\sum(y_{\mathbf{i}} - \widehat{\theta}_{\mathbf{i}})^2 = n^2\sigma^2$, i.e.

$$\sum_{\mathbf{i}} \left(\beta\lambda_{\mathbf{i}}/(1+\beta\lambda_{\mathbf{i}})\right)^2 (D^T y)_{\mathbf{i}}^2 = n^2\sigma^2.$$

3. Variance tuning with equivalent degrees of freedom: The residuals $y_{\mathbf{i}} - \widehat{\theta}_{\mathbf{i}}$ have smaller variance than the $\varepsilon_{\mathbf{i}}$. This can be taken into account by the so-called equivalent degrees of freedom (Hall and Titterington, 1986) which gives in our situation:

$$\sum_{\mathbf{i}} \left(\beta\lambda_{\mathbf{i}}/(1+\beta\lambda_{\mathbf{i}})\right)^2 (D^T y)_{\mathbf{i}}^2 = \sigma^2 \sum_{\mathbf{i}} \beta\lambda_{\mathbf{i}}/(1+\beta\lambda_{\mathbf{i}})$$

4. Marginal likelihood estimation: The marginal log likelihood of the $y_{\mathbf{i}}$'s is

$$\sum_{\mathbf{i}} \log\left(\sigma^2 + \tau^2/\lambda_{\mathbf{i}}\right) + \sum_{\mathbf{i}} (\sigma^2 + \tau^2/\lambda_{\mathbf{i}})^{-1} (D^T y)_{\mathbf{i}}^2$$

So we can minimize this with respect to τ^2 for given σ^2 or with respect to both σ^2 and τ^2. A small problem arises here because the smallest eigenvalue is zero. We simply exclude this eigenvalue from the sum above.

5. Generalized cross validation (see Wahba, 1990): In our situation, this amounts to choosing β by minimizing

$$\sum_{\mathbf{i}} \left(\beta\lambda_{\mathbf{i}}/(1+\beta\lambda_{\mathbf{i}})\right)^2 (D^T y)_{\mathbf{i}}^2 \left(\sum_{\mathbf{i}} \beta\lambda_{\mathbf{i}}/(1+\beta\lambda_{\mathbf{i}})\right)^{-2}.$$

Each of the proposals 1-5 leads to an estimating equation of the form

$$\sum_{\mathbf{i}} h_{\mathbf{i}}(\beta)(D^T y)_{\mathbf{i}}^2 = g(\beta) \tag{3.3}$$

with a suitable choice of $h_{\mathbf{i}}$ and g. This equation is easy to solve numerically in all cases. In order to obtain some comparison of the methods without doing extensive simulations, we calculated approximate means and variances of $\widehat{\beta}$ with the following argument. Denote by β_0 the solution of

$$\sum_{\mathbf{i}} h_{\mathbf{i}}(\beta) E\left[(D^T y_{\mathbf{i}})^2\right] = g(\beta) \tag{3.4}$$

and set $u_{\mathbf{i}} = (D^T y)_{\mathbf{i}}^2 - E[(D^T y)_{\mathbf{i}}^2]$. If $\sum h_{\mathbf{i}}(\beta_0) u_{\mathbf{i}}$ is small compared with $\sum h_{\mathbf{i}}(\beta_0) E[(D^T y)_{\mathbf{i}}^2]$, it makes sense to consider a Taylor expansion of (3.3) around β_0 . This gives

$$0 = \sum_{\mathbf{i}}(D^T y)^2_{\mathbf{i}} h_{\mathbf{i}}(\widehat{\beta}) - g(\widehat{\beta}) \cong \sum_{\mathbf{i}} h_{\mathbf{i}}(\beta_0)u_i + (\widehat{\beta} - \beta_0)\{\sum_{\mathbf{i}}(D^T y)^2_{\mathbf{i}} h'_{\mathbf{i}}(\beta_0) - g'(\beta_0)\}.$$

Substituting the last term in brackets by its expectation, we finally obtain

$$\widehat{\beta} - \beta_0 \cong -\sum h_{\mathbf{i}}(\beta_0)u_i / \{\sum h'_{\mathbf{i}}(\beta_0)E[(D^T y)^2_{\mathbf{i}}] - g'(\beta_0)\}. \tag{3.5}$$

From this we can see that

$$E[\widehat{\beta}] \cong \beta_0 \text{ and} \tag{3.6}$$

$$\mathrm{Var}[\widehat{\beta}] \cong \sigma^2(\beta_0) = \sum h_{\mathbf{i}}(\beta_0)^2 \mathrm{Var}[(D^T y)^2_{\mathbf{i}}] / \{\sum h'_{\mathbf{i}}(\beta_0)E[(D^T y)^2_{\mathbf{i}}] - g'(\beta_0)\}^2 \tag{3.7}$$

Table 1 gives numerical values for the methods 1-5 together with the optimal β obtained by minimizing the sum of squared errors $\sum_{\mathbf{i}}(\theta_{\mathbf{i}} - \widehat{\theta}_{\mathbf{i}})^2$. Estimates with σ^2 known are closer to the optimal value than those who do not require σ^2. Methods based on the marginal likelihood have smallest variablility, but tend to undersmooth more than generalized cross validation and minimizing estimated MSE. This undersmoothing is probably due to the prior being inappropriate for the true image (2.2). A similar phenomenon has bee noticed by Wahba (1985) for spline smoothing.

In this comparison one should however keep in mind that no choice of β achieves a decent restoration. Either the edges are smoothed away or there is too much noise left in the areas where θ^0 is constant. In view of this and in view of the square errors given in Table 1, any choice of β between 1 and 3 is acceptable, and it is not worthwhile to try to improve the methods.

4 The Huber prior

If ϕ is convex, the quantity $H(\theta|y)$ of (2.5) which we have to minimize is strictly convex in θ. Moreover $H(\theta|y)$ goes to $+\infty$ as any θ_{ij} goes to $\pm\infty$. Hence $H(\theta|y)$ has exactly one minimum which is global. An algorithm to find this minimum which is easy to implement is Besag's (1986) iterated conditional modes (ICM). It visits the pixels periodically in a given order and always minimizes $H(\theta|y)$ by varying only the component of θ at the current pixel. Naively one might think that this algorithm must converge to the MAP. However for a proof one needs differentiability of ϕ as noted by Besag et al. (1991), p. 9-10. In the case $\phi(x) = |x|$, ICM reaches a fixpoint after a finite number of steps, but this fixpoint is usually very far from the MAP. This was a surprise at least for me. The fixpoints (there are many of them) are all the pixelwise minima of $H(\theta|y)$, and a pixelwise minimum need not be a minimum unless ϕ is differentiable. The phenomenon occurs already in the case of two pixels if $|y_1 - y_2| < \beta$. Since the absence of computational problems is the main advantage of convex ϕ's, the choice $\delta = 0$ is not recommended.

We are thus left with two parameters β and δ to be determined. It has been suggested, e.g. in Geman and McClure (1987) or Geman and Reynolds (1992), that the scale parameter δ is less crucial and can be chosen on a priori grounds like the size of a step in the clean image considered relevant. Based on this we put $\delta = 0.5$. The restoration with $\beta = 0.75$ given in Fig. 4 is disappointing. Varying β did not help either. First I believed that this shows a failure of convex ϕ's until by curiosity I once put $\delta = 0.05, \beta = 0.075$ (remember that with our choice of the parameters β/δ should remain constant as $\delta \longrightarrow 0$). The result given in Fig. 5 is much more satisfactory. So the value of δ is crucial here, and for good results it must be much smaller than a step in the image considered to be of interest. For small δ, the ICM needs a rather large number of sweeps until convergence – an indication of the problems that occur when $\delta = 0$. Nevertheless one sweep of ICM is done really fast, so computation was not a problem, even for small δ.

The choice of the smoothing parameter β was made by the following ad-hoc argument. For a lower bound we require that a single spike in a flat region is effectively smoothed out. So assume that $\theta_{k\ell} =$ const. for $(k,\ell) \sim (i,j)$, $y_{ij} =$ const. $+\, x$ and consider $\widehat{\theta}_{ij}$ minimizing $H(\theta|y)$. One obtains

$$\widehat{\theta}_{ij} = \text{ const.} + \begin{cases} x + 4\beta/\delta & \text{if } \; x \le -4\beta/\delta - \delta \\ x\delta/(\delta + 4\beta/\delta) & \text{if } \; |x| \le 4\beta/\delta + \delta \\ x - 4\beta/\delta & \text{if } \; x \ge 4\beta/\delta + \delta. \end{cases} \tag{4.1}$$

So very large spikes will only be reduced by a constant. This is a consequence of assuming normal errors. But for $\delta << \sigma$ and $\beta/\delta \ge 0.75\ \sigma$ all practically occuring spikes will be reduced sufficiently.

For an upper bound we require that reasonably large patches in the image should be retained. Assume first that the true image θ^0 of (2.2) is known. Then we consider the following two restorations $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$

$$\widehat{\theta}^{(1)}_{ij} = \begin{cases} m_1 & 1 \le i,j \le 10 \\ m_2 & 1 \le i \le 10, 11 \le j \le 20 \\ m_3 & 11 \le i \le 20, 1 \le j \le 10 \\ m_4 & 11 \le i,j \le 20 \end{cases} \tag{4.2}$$

and

$$\widehat{\theta}^{(2)}_{ij} = \begin{cases} (m_1 + m_2)/2 & 1 \le i \le 10 \\ m_3 & 11 \le i \le 20, 1 \le j \le 10 \\ m_4 & 11 \le i,j \le 20 \end{cases} \tag{4.3}$$

where the m_k's are the means of the y_{ij}'s in each of the four quarters. So in $\widehat{\theta}^{(2)}$ the upper half of the small edge has been smoothed out. Now we require that with high probability

$$H(\widehat{\theta}^{(1)}|y) \le H(\widehat{\theta}^{(2)}|y). \tag{4.4}$$

Of course (4.4) does not guarantee that the MAP is closer to $\widehat{\theta}^{(1)}$ than to $\widehat{\theta}^{(2)}$, but it should give an indication for reasonable values of the parameters. By a straightforward computation we have for $m_2 > m_1 + \delta, m_1 > m_3 + \delta, (m_1 + m_2)/2 > m_4 + \delta$

$$H(\widehat{\theta}^{(2)}|y) - H(\widehat{\theta}^{(1)}|y) = 200\ (m_1 - m_2)^2/4 - 10\beta(2|m_1 - m_2|/\delta - 1)$$

Since $m_1 - m_2 \sim \mathcal{N}(1, 0.02)$, it follows that

$$P[H(\widehat{\theta}^{(2)}|y) \ge H(\widehat{\theta}^{(1)}|y)] \ge 1 - \Phi\left((4\beta/\delta - 10)2^{-1/2}\right). \tag{4.5}$$

The right hand side is practically one for $\beta/\delta \leq 1.5$. This explains the value $\beta/\delta = 1.5$ we have used. Table 2 shows that values of β/δ between 0.75 and 1.5 give indeed good restorations in our case. Note also that the sum of squared errors is much smaller than with the Gaussian prior.

One might object that our way of choosing the smoothing parameter depends strongly on knowing the clean image. But first, this is not true for the lower bound which is already quite a good value. Second, we can still use similar arguments as long as we have an idea for what kind of clean image our restoration should perform well. We then can compare as above a faithful restoration with one where patches and edges we would like to retain are smoothed out. An alternative is to use marginal likelihood estimation as in Ogata (1990). This requires however extensive computations.

Finally we give here an argument which explains why small values of δ are needed. We note that the MAP $\widehat{\theta}$ is determined as the solution of

$$\begin{aligned} 0 &= -2(y_{ij} - \widehat{\theta}_{ij}) + \beta/\delta \textstyle\sum_{(k,\ell)\sim(i,j)} \phi'\left((\widehat{\theta}_{ij} - \widehat{\theta}_{k\ell})/\delta\right) \\ &= -2(y_{ij} - \widehat{\theta}_{ij}) + 2\beta/\delta^2 \textstyle\sum_{(k,\ell)\sim(i,j)} \chi\left((\widehat{\theta}_{ij} - \widehat{\theta}_{k\ell})/\delta\right)(\widehat{\theta}_{ij} - \widehat{\theta}_{k\ell}) \end{aligned}$$

where $\chi(x) = \phi'(x)/(2x)$. This suggests the following iterative algorithm to calculate $\widehat{\theta}$: Put $\widehat{\theta}^{(0)}$ equal to the restoration with a Gaussian prior and smoothness parameter β/δ^2. Then calculate iteratively for $m = 1, 2, \ldots$

$$\widehat{\theta}^{(m)} = \text{ arg min}_\theta \left\{ \sum_{i,j} (y_{ij} - \theta_{ij})^2 + \beta/\delta^2 \sum_{(i,j)\sim(k,\ell)} b^{(m)}(i,j,k,\ell)(\theta_{ij} - \theta_{k\ell})^2 \right\}$$

where $b^{(m)}(i,j,k,\ell) = \chi\left((\widehat{\theta}_{ij}^{(m-1)} - \widehat{\theta}_{k\ell}^{(m-1)})/\delta\right)$. So each $\widehat{\theta}^{(m)}$ is the restoration with a Gaussian prior, smoothing parameter β/δ^2 and weights for the bonds between neighboring pixels. This iterative procedure is exactly what one obtains from the dual edge model of Geman and Reynolds (1992), Section 3. Namely $\phi(\sqrt{x})$ is concave and with $\beta(b) = 1/b$ we have

$$\phi(x) = \inf_{0<b\leq 1} (bx^2 + \beta(b)) - 1.$$

From this it can be shown that the procedure converges to the MAP. Ideally the weights should be zero at an edge and one otherwise. For $\beta = 0.75, \delta = 0.5$ the weights $b^{(1)}$ are unity for all horizontal bonds except one where it is 0.94. The vertical bonds not crossing the edge are all unity, and those crossing the edge are given in Table 3. This suggests that with these parameters the MAP will bring out the larger edge a bit clearer than the Gaussian restoration, but the smaller edge will not change. This is confirmed in Fig. 4. In contrast, for $\beta = 0.075, \delta = 0.05$ the weights $b^{(1)}$ are less than one for bonds close to either of the two edges, cf. Fig. 6. This explains the relative success of the restoration with these parameters. Unfortunately we are not able to compute analytically which values of δ give good weights $b^{(1)}$, as a function of the size of patches and edges relative to the noise variance.

5 The truncated Gaussian prior

We begin by discussing the choice of the smoothing parameter β, using the same arguments as in Section 4. For the lower bound, the analogue of (4.1) is

$$\widehat{\theta}_{ij} = \text{const.} + \begin{cases} x\delta^2/(4\beta+\delta^2) & \text{if} \quad |x| < (4\beta+\delta^2)^{1/2} \\ x & \text{if} \quad |x| > (4\beta+\delta^2)^{1/2} \end{cases} \tag{5.1}$$

ϕ'. From this we see that for $\delta << \sigma$ we should have $\beta \geq 3\sigma^2$ in order to smooth out spikes occurring under Gaussian noise. On the other hand for the upper bound we again compare the restorations $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ of (4.2)-(4.3). We have

$$H(\widehat{\theta}^{(1)}|y) - H(\widehat{\theta}^{(2)}|y) = 200(m_1 - m_2)^2/4 - 10\beta.$$

Hence

$$P\left[H(\widehat{\theta}^{(2)}|y) \geq H(\widehat{\theta}^{(1)}|y)\right] \geq 1 - \Phi\left((10\beta)^{1/2} - 50^{1/2}\right)$$

In order to have the right hand side close to one, β should be ≤ 2. Hence with this prior it seems difficult to achieve decent local smoothing without removing relevant structure with non-negligible probability. In our specific realization both $m_2 - m_1$ and $m_4 - m_3$ were very close to one, so here β's larger than 2 can be used.

A second problem with this prior is the computation of the MAP because $H(\cdot|y)$ has a large number of local minima. In particular, the result of ICM depends strongly on the starting value. For instance taking the observed image itself as starting value gives poor results. The results with the restoration based on the Huber prior ($\beta = 0.075,\ \delta = 0.05$) as the starting value are shown in Figs. 7-9 for $\beta = 4$ and various choices of δ. With $\beta = 2$ the pictures look very similar except for some additional outliers which are not smoothed out. So with $\delta = 0.5$ the smaller edge disappears whereas with $\delta = 0.05$ many false edges are introduced. The choice $\delta = 0.25$ seems to be about right.

One might wonder whether this sensitivity to the choice of δ is reduced when looking at the global minimum. I have experimented with ICM for different starting values, including the clean image. From this I conjecture that for $\beta \geq 2, \delta = 0.5$ and $\beta = 4, \delta = 0.25$ the global minimum preserves only the larger edge whereas for $\beta \leq 4, \delta = 0.05$ and $\beta = 2, \delta = 0.25$ it preserves both edges. So the choice of δ seems to be important also for this prior, and rather small values are required for a faithful restoration. However it seems that the smaller δ, the more local minima exist. The question thus arises whether there is an algorithm which comes at least close to the global minimum. The algorithm most widely discussed in the imaging literature for this task is simulated annealing, Geman and Geman (1984). We restrict ourselves in the following to the limiting case $\delta = 0$ because small δ's are interesting. Moreover for $\delta = 0$ the problem becomes discrete which simplifies the algorithm to some extent. Namely in order to compute the MAP for $\delta = 0$, we only have to partition the set of pixels $\{1, \ldots, n\}^2$ into connected components where the restoration is constant. Then the restoration is equal to the mean of the y_{ij}'s in each component.

The effect of the prior is to add a roughness penalty equal to β times the length of boundaries to the lack of fit term $\sum(y_{ij} - \theta_{ij})^2$. Hence when looking at a fixed pixel, it may either belong to the same component as one of the neighboring pixels or it may form a component by itself. The change in $H(\theta|y)$ for these possibilities is easily computed. So we used the following variant of simulated annealing with Metropolis' algorithm.

1. Choose a temperature schedule $(T_k)_{k=1,\ldots,N}$.
2. Choose the initial partition by having each pixel form a separate component and put $k = 1$.
3. At step k do the following:
   (a) Choose a pixel (i, j) at random.
   (b) Modify the current partition by letting (i, j) belong to a different component, randomly chosen among all possibilities.
   (c) Compute the change Δ in $H(.|y)$ between the current and the modified partition.
   (d) Make the modified partition to the new current partition with probability $\min(1, \exp(\Delta/T_k))$.
4. Increase k by 1 unless $k = N$ and go back to 3.

Some care is needed at step 3(b) because it might happen that the modified partition has no longer connected components. In such a case we kept the current partition and increased k by 1. A drawback of this is that the algorithm is not parallelizable because we need the whole partition to check whether the components are connected. It is easy to see that with the above transitions we can get from any partition to the partition consisting of one single component and vice versa. This is needed for the application of simulated annealing.

I experimented with the above algorithm, taking $\beta = 2$ and $\beta = 4$ and a linear temperature schedule between $4\beta/\log(2)$ and $\beta/\log(n^2)$. It turned out that for good results N had to be 2000 n^2. With $\beta = 4$ we then came very close to the restoration (4.2), only a few pixels were misclassified. With $\beta = 2$ the results were not as good.The restorations obtained had in addition to the four big components several small groups of outliers which were not smoothed away. Presumably (4.2) is still the MAP for $\beta = 2$, but there are other restorations where there is only a small difference in $H(\theta|y)$. Still with $\beta = 4$ simulated annealing passed this rather difficult test and produced the desired results. The price in terms of necessary iterations is however rather high. The problem with the algorithm is that once we are in a local minimum and T_k is low, then a long time is spent until the current partition is actually changed. It ought to be possible to speed up the algorithm at this stage.

In Leclerc (1989) a different algorithm has been proposed. It chooses a decreasing sequence $\delta_k \downarrow 0$. At step k, ICM with $\delta = \delta_k$ and the current restoration as starting value is used to obtain the next restoration and then k is increased by one. We were unable to make this algorithm work. In order to bring out the smaller edge, δ has to be small. When we finally reached a small enough δ, this smaller edge had already been smoothed out.

6 Two additional examples

Here we briefly report on the results we obtained with our methods in the two cases

$$\theta_{ij}^0 = 0 \tag{6.1}$$

and

$$\theta_{ij}^0 = \begin{cases} 1 + 0.05(j - 0.5) & 1 \leq i \leq 10 \\ -1 + 0.05(j - 0.5) & 11 \leq i \leq 20 \end{cases} \tag{6.2}$$

respectively. The noise (ϵ_{ij}) was the same as the one which was added to (2.2) to produce Fig. 1. The noisy images are shown in Figs. 10 and 12. With (6.1) we wanted to see what happens when we have more smoothness than we expected. Also it gives us a check that the techniques do not produce edges which are not there. With (6.2) we wanted to see whether the techniques can distinguish between smooth changes and jumps. In these two examples we always used the same parameters β and δ as before. The aim was to see how well we can do when the true image is different from the prototype used to determine the parameters.

The results from using the Huber prior with $\delta = 0.05$ and $\beta = 0.0755$ are shown in Figs. 11 and 13. The restorations look quite good. Note that the Gaussian prior would have again difficulties with the edge in (6.2). It could of course do very well with (6.1) when we choose a large smoothing parameter. But in order to achieve the same sum of squared errors as the Huber prior, we have to put $\beta \approx 20$ which is very different from the optimal β for (2.2).

With the prior $\phi(x) = 1_{[x \neq 0]}$ there are two questions: What does the MAP look like, and how close can we come to the MAP with simulated annealing ? Let us first discuss (6.1). Then I am quite sure that the MAP for $\beta = 4$ is the constant restoration equal to the arithmetic mean of the observations. However simulated annealing could not find this restoration even when I increased the number of sweeps to $3000n^2$. It typically ended with 3 or 4 connected components. With $\beta = 2$ annealing left also additional isolated outliers unsmoothed. So in this case the MAP would be the optimal restoration, but I could not solve the computational difficulty.

Finally in the case of (6.2) I believe that the MAP with $\beta = 4$ has two connected components $\{(i,j); i \leq 10\}$ and $\{(i,j); i > 10\}$. With $\beta = 2$ the MAP has presumably the 4 components $\{(i,j); i \leq 10, j \leq 14\}$, $\{(i,j); i \leq 10, j > 14\}$, $\{(i,j); i > 10, j \leq 15\}$ and $\{(i,j); i > 10, j > 15\}$. Again we could not quite reach the MAP with simulated annealing. It produced one or two additional components for $\beta = 4$ and several additional components with $\beta = 2$. So in any case this prior fails for the image (6.2).

7 Discussion

We have shown that in our example it is possible to restore the edges with a Gibbs priors which penalizes large differences between neighboring pixels less severly than a Gaussian prior. There seems to be no need to introduce an unobservable edge process here. Also the main improvement occurs when passing from a quadratic to a convex potential with bounded derivative. In this case the computation of the restoration is rather easy. The absolute value potential which is the limit of the scale parameter going to zero should however be avoided because it makes the computation of the MAP difficult. Using a bounded non-convex potential allows an almost perfect restoration provided one has an algorithm which attains the global maximum of the posterior distribution. We have shown that at least in one case simulated annealing can do this. Still it is doubtful whether the slightly improved restorations are worth the much larger computational effort.

For all our results the good choice of both the scale and the smoothing parameter are crucial. The scale parameter should be much smaller than the steps in image. One possible interpretation for this is that we should use potentials which are concave for all positive arguments. The smoothing parameter has been chosen by considering what happens to constant regions of different size and height in an image. It would be interesting to know whether Bayesian likelihood and ABIC (Akaike 1980) give similar parameter and model choices. This would require high-dimensional integrations as in Ogata (1990). Finally there is the question how our findings generalize, e.g. to more complex images and situations with blur. It is clear that more complex images require additional terms in the prior, but with such modifications we expect a similarly good performance of robust priors. The examples in the last Section give additional evidence for this.

Acknowledgement: This work was begun during a stay as visiting professor of the Institute of Statistical Mathematics. I would like to thank the people there for financial support, hospitality and discussions. I also thank Julian Besag, Donald Geman and two referees for helpful comments on the first version of this paper.

REFERENCES

Akaike, H. (1980). Likelihood and Bayes procedure. *In:* Bayesian Statistics (eds. J. M. Bernardo et al.). University Press, Valencia, Spain.

Bellman, R. (1960). *Introduction to Matrix Analysis.* McGraw-Hill, New York.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). J. Roy. Statist. Soc. B, **48**, 192-236.

Besag, J. (1986). Towards Bayesian image analysis. J. Appl. Statist. **16**, 395-407.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math. **43**, 1-59.

Geman, D. and Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. IEEE Trans. Pattern Anal. Machine Intell., **14**, 367-383.

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

Geman, S. and McClure, D.E. (1987). Statistical methods for tomographic image reconstruction. Bulletin of the Intern. Statist. Inst., Proc. of the 46th Session **52**, Book 4, 5-21.

Green, P.J. (1990). Penalized likelihood reconstructions from emission tomography data using a modified EM algorithm. IEEE Trans. Med. Imag. **9**, 84-93l

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

Kay, J.W. (1988). On the choice of regularisation parameter in image restoration. *In:* Pattern Recognition. (ed. J. Kittler), Lecture Notes in Comp. Sci. **301**, 587-596.

Kitagawa, G. (1987). Non-Gaussian state space modeling of nonstationary time series (with discussion). J. Amer. Statist. Assoc. **82**, 1032-1063.

Leclerc, Y.G. (1989). Constructing simple stable descriptions for image partitioning. Intern. J. Comp. Vision **3**, 73-102.

Ogata, Y. (1990). A Monte Carlo method for an objective Bayesian procedure. Ann. Inst. Statist. Math. **42**, 403-433.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

Speed, T.P. (1978). Relations between models for spatial data, contingency tables and Markov fields on graphs. Suppl. Adv. Appl. Probab. **10**, 111-122.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann. Statist. **13**, 1378-1402.

Wahba, G. (1990). *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series **59**, SIAM Philadelphia.

Table 1:

Comparison of data driven choices of the smoothing parameter with a Gaussian prior. $\hat{\beta}$ is computed for the image of Fig.1. SSE($\hat{\beta}$) is the sum of squared errors. β_0 and $\sigma(\beta_0)$ are approximate means and standard errors of $\hat{\beta}$ as defined in (3.4) and (3.7) respectively.

Method	$\hat{\beta}$	SSE($\hat{\beta}$)	β_0	$\sigma(\beta_0)$
Minimizing estimated mean square error	1.52	57.4	2.03	0.49
Variance tuning	3.59	58.9	5.63	2.09
Variance tuning with equivalent degrees of freedom	1.55	57.2	2.90	1.50
Marginal likelihood with σ^2 known	1.21	60.8	1.36	0.18
Marginal likelihood with σ^2 unknown	0.97	66.0	0.83	0.26
Generalized cross validation	1.43	58.1	1.27	0.50
Minimizing sum of squared errors	2.16	55.5	2.03	0.23

Table 2:

Sum of squared errors for the restoration using the Huber prior with $\delta = 0.05$ and varying β.

β/δ	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
SSE	26.1	22.1	23.1	25.9	31.1	37.8	45.1	53.1

Table 3:

Weights $b^{(1)}$ for vertical bonds crossing the edge in the clean image for $\beta = 0.75$ and $\delta = 0.5$. The weights are defined in the text.

column number	1	2	3	4	5	6	7	8
weight	.73	.75	.56	.70	.81	1.0	1.0	1.0
column number	9	10	11	12	13	14	15	16
weight	.98	.81	1.0	1.0	.88	1.0	.77	.90
column number	17	18	19	20				
weight	.96	.83	.63	.81				

Figure captions

Fig.1: Bird's-eye-view (a) and gray-scale image (b) of one realization from (2.1)-(2.2). The gray-scales split the range $[-3.62, 4.66]$ of the image into equal intervals.

Fig.2: Bird's-eye-view (a) and gray-scale image (b) of the moving median smoother in a 5×5 window. Gray scales as in Fig. 1.

Fig.3: Bird's-eye-view (a) and gray-scale image (b) of the moving shorth smoother in a 7×7 window. Gray scales as in Fig. 1.

Fig.4: Bird's-eye-view (a) and gray-scale image (b) of the MAP with the Huber prior, $\beta = 0.75$ and $\delta = 0.5$. Gray scales as in Fig. 1.

Fig.5: Bird's-eye-view (a) and gray-scale image (b) of the MAP with the Huber prior, $\beta = 0.075$ and $\delta = 0.05$. Gray scales as in Fig. 1.

Fig.6: Gray-scale image of the weights $w^{(1)}$ for horizontal (a) and vertical (b) bonds for $\beta = 0.075$ and $\delta = 0.05$. The weights are defined in the text. Gray scales split the inverval $[0, 1]$ into equal intervals.

Fig.7: Bird's-eye-view (a) and gray-scale image (b) of the result of ICM with truncated Gaussian prior, $\beta = 4$ and $\delta = 0.5$, using the restoration of Fig. 5 as starting value. Gray scales as in Fig. 1.

Fig.8: Same as Fig. 7, but with $\delta = 0.25$.

Fig.9: Same as Fig. 7, but with $\delta = 0.05$.

Fig.10: Bird's eye view (a) and gray-scale image (b) of (6.1) with added noise. The gray-scales split the range [-3.24,3.11] of the image into equal intervals.

Fig.11: Bird's eye view (a) and gray-scale image (b) of the MAP for Fig. 10 with $\beta = 0.075$ and $\delta = 0.05$. Gray-scales as in Fig. 10.

Fig.12: Bird's eye view (a) and gray-scale image (b) of (6.2) with added noise. The gray-scales split the range [-3.24,4.58] of the image into equal intervals.

Fig.13: Bird's eye view (a) and gray-scale image (b) of the MAP for Fig. 12 with $\beta = 0.075$ and $\delta = 0.05$. Gray-scales as in Fig. 12.