Bayesian scale space analysis of differences in images

Lasse Holmström *,1, Leena Pasanen 2

Department of Mathematical Sciences, P.O.Box 3000, 90014 University of Oulu, Finland

Abstract

We consider the detection of image features that appear in different spatial scales or resolutions. In particular, the goal is to capture the scale dependent differences in a pair of noisy images of the same scene taken at two different instants of time. A new approach is proposed that uses Bayesian statistical modeling and simulation based inference. The method can be viewed as a further development of SiZer technology, originally designed for nonparametric curve fitting. A strength of the Bayesian simulation based approach is straightforward inference and modeling flexibility that facilitates the incorporation of domain specific prior information on the images under consideration. The method is believed to have applications for instance in satellite based remote sensing. In addition to artificial test images, a preliminary analysis of a pair of Landsat images used in satellite based forest inventory is therefore also included.

Key words: Image analysis, Bayesian methods, Scale space, SiZer, BSiZer, Satellite images
1 Introduction

The idea of a scale space has its origin in computer vision where it refers to a family of smooths of a digital image [14]. No particular level of smoothing is regarded as “correct” and each smooth is thought to provide information about the underlying image structure at a particular scale, little smoothing revealing small details and heavy smoothing displaying only the coarsest features. The presence of noise makes scale space analysis harder as it may then be difficult to distinguish the actual structure of the image from artifacts caused by noise. Tools of statistical inference therefore become relevant in assessing which of the apparent features are structure and which are just random noise.

Scale space analysis was only relatively recently introduced to statistics by P. Chaudhuri and J.S. Marron [1,2]. Their original approach, called SiZer, considered one-dimensional nonparametric probability density estimation and curve fitting. The statistically significant features of the curve underlying the data were investigated by finding a confidence interval for the derivative of its smooth. When this is done, not just for one particular level of smoothing, but for a whole range of smooths, then features of the underlying curve such as its maxima, minima and local trends are revealed in several different scales simultaneously. Since its inception, the original SiZer has been developed into various directions, including a two-dimensional version that can be applied to images [6,7]. While most of this work has been based on classical statistics, more recently Bayesian approaches have also been proposed. The Bayesian SiZer for curve fitting, BSiZer for short, was introduced in [4]. A somewhat different approach is described in [8]. The purpose of the present paper is to introduce iBSiZer, that is, the Bayesian SiZer for images.

The image SiZer [6,7] treats the image as a surface on the pixel plane and classifies its features based on gradients and curvatures. In this paper we will consider the seemingly simpler problem of just detecting the presence of features in different spatial scales. Such an approach is useful for example when one is interested in finding statistically significant, resolution dependent differences between two noisy images. The proposed approach is believed to have applications for instance in the analysis of remote sensing satellite images for the purposes of forest inventory and environmental monitoring as well in medical imaging. In addition to artificial examples, we discuss a preliminary

* Corresponding author
  
  Email address: lasse.holmstrom@oulu.fi (Lasse Holmström).
  
  URL: http://cc.oulu.fi/~lh/ (Lasse Holmström).

1 Visiting the National Center for Atmospheric Research in Boulder, Colorado.
Work supported by a Senior Scientist Grant no. 119675 from the Academy of Finland.

2 Work supported by the Finnish Graduate School in Stochastics and Statistics.
analysis of a pair of Landsat 7 ETM+ images used in satellite based forest inventory.

The basic idea of iBSiZer is to compute several smooths of an image reconstructed from a noisy observation and then to make inferences about the credible features in those smooths. An early version of this method was proposed in [9]. An example of iBSiZer analysis is shown in Figure 1. Displayed are, from left to right, the original image, a noisy observation of it, and “maps” that visualize the inference of credible features that appear in three different resolutions. As we are mainly interested in difference images, the pixel intensities here take on both positive and negative values and different colors are used to indicate areas of credibly positive and negative pixel values. Note that, although the paper focuses on the detection of differences between images, iBSiZer can also be used to identify areas in images that differ from any given reference field. For example, one can find areas that are credibly brighter or darker than the image mean.

Smoothing amounts to local weighted averaging of the pixel intensities which means that iBSiZer analyzes the images in different resolutions. At smallest scale smoothing, differences in individual pixels or average differences in very small neighborhoods of pixels are considered. Raising the smoothing level corresponds to analyzing the average differences in increasingly large neighborhoods of pixels. Then changes in isolated pixels tend to be smoothed out and what remains are just large scale mean changes. Sometimes also a low intensity, large scale feature which with small scale smoothing is masked by noise becomes visible and credible for the largest scale smooths. Instances of these phenomena will be seen in the examples.

The structure of the rest of the paper is as follows. In section 2 the basic Bayesian framework and the construction of the iBSiZer maps are explained. Then, in section 3, the statistical models used are discussed in detail. We also describe our sampling procedures and explain how some important parameters can be estimated. Example images, both artificial and natural, are analyzed in section 4. A discussion and suggestions for further work are offered in section 5.


2 Multiscale analysis of difference images

2.1 The basic Bayesian framework

We think of a digital image as an $M \times N$ array of real numbers $x_{ij}$. However, in mathematical derivations we treat the image as a vector $x = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$, $n = MN$, obtained by vectorizing the two-dimensional array column-wise. A noisy observed image is modeled as

$$
y = x + \varepsilon,
$$

where $\varepsilon$ is additive Gaussian noise with $\varepsilon \sim \mathcal{N}(0, \Sigma)$, and $\Sigma$ is positive definite. Although this model is used throughout this paper, our methods can be easily generalized to other noise types. Given two such noisy images $y_i = x_i + \varepsilon_i$, $i = 1, 2$, we want to find pixels where $x_1$ and $x_2$ are different. For example, if the images represent the same object viewed at two different instants of time, differences indicate changes that have occurred over time. The challenge is, of course, to find the true differences in the presence of corrupting noise in the observed images. Assuming that $\varepsilon_1$ and $\varepsilon_2$ are independent, the difference image $x = x_2 - x_1$ and its observed version $y = y_2 - y_1$ can be assumed to still satisfy (1).

We will base our inference on a Bayesian approach that models one’s prior beliefs about $x$ and other model parameters $\theta$ by assigning to them prior probability densities $p(x|\theta)$ and $p(\theta)$. Here $\theta$ can include $\Sigma$ and/or some tuning parameters that may or may not be treated as random variables. The likelihood of the data $y$ is given by the conditional density $p(y|x, \theta)$ and the joint posterior density of $x$ and $\theta$ is

$$
p(x, \theta|y) = p(x|\theta)p(\theta)p(y|x, \theta) \propto p(x|\theta)p(\theta)p(y|\theta).
$$

In simple cases, e.g., when $\theta = \Sigma = \sigma^2 I$, with $I$ the $n \times n$ identity matrix, the posterior density $p(x|y)$ can be obtained analytically by integrating $p(x, \theta|y)$ over $\theta$. In more complicated situations integrations can be avoided by resorting to simulation, as will be explained later in subsections 3.2.1 and 3.2.2.

In the simplest kind of inference that treats each pixel independently of others, one can now find (or sample) the marginal posterior density $p(x_s|y)$ of each pixel, and identify those pixels which differ from zero with at least some high threshold posterior probability $\alpha$. We refer to this mode of inference as “pixelwise”. These are then the pixels at which the difference between the two images is taken to be credible, or, “statistically significant”. In all inferences we typically take $0.9 \leq \alpha < 1$. 

2.2 Scale space analysis: iBSiZer

In the scale space approach, instead of just the difference image $x$ itself, we are in fact interested in its credible features that appear in a range of scales, or image resolutions. We therefore consider a smoothing operator $S_\lambda$ and smooths $x_\lambda = S_\lambda x$ of $x$ for several different smoothing levels $\lambda > 0$. Bayesian inference about the smooth $x_\lambda$ is based on its posterior density $p(x_\lambda | y)$. This posterior density can be obtained by applying the transformation $S_\lambda$ to the density $p(x|y)$. In practice, a sample is first simulated from $p(x|y)$ and then transformed by applying $S_\lambda$ to the sample images and inference is based on this transformed sample.

In principle, the smoother $S_\lambda$ used in iBSiZer could be for example a convolution with a Gaussian smoothing kernel, as suggested in the general scale space theory [14], or a nonparametric regression as in the original SiZer [1]. In fact, we only want that, as $\lambda \to \infty$, $S_\lambda x$ should approach the pixel intensity mean $(1/n) \sum_{s=1}^{n} x_s$. We in fact use a smoother of the form $S_\lambda = (I + \lambda Q)^{-1}$, where $Q$ is a prior precision matrix (inverse covariance matrix) used for $x$ (cf. section 3.1). As will be seen in section 3.2.1, this smoother suggests itself in a natural way through the structure of the posterior mean $E(x|y)$ (cf. (11)). The advantages of using $S_\lambda$ instead of, say, Gaussian convolution, are that by utilizing sparse matrix algorithms (e.g. LU-factorization of a band matrix) $S_\lambda x$ can be computed fast and that there is no need for special handling of pixels outside the image proper, such as zero padding. The particular $Q$ we use in $S_\lambda$ is associated with the second order Neumann prior discussed in section 3.1. With this precision matrix $S_\lambda$ turns out to be an approximate kernel smoother where a small portion of the weights are slightly negative.

2.3 The iBSiZer atlas of maps

The results of iBSiZer inference are displayed in the form of an “atlas” of color maps, with each map corresponding to a given level $\lambda$ of smoothing. An example of such an atlas is shown in Figure 4. We have also developed an interactive tool which allows the user to explore the maps through a graphical interface. This tool and the Matlab code for iBSiZer are available at http://cc.oulu.fi/~lpasanen/iBSiZer.

In a map, blue and red are used for pixels at which, with posterior probability at least $\alpha$, the difference in the two images is positive and negative, respectively. Gray indicates that the difference is not significant at this level of credibility. However, instead of the simple pixelwise (PW) inference described in section 2.1 we in fact perform joint, that is, simultaneous inference.
over all pixels of the image. This results in maps that show the global pattern of credible features at the chosen level $\alpha$ of posterior probability. We adapted the simultaneous inference methods of “highest pointwise probabilities” (HPW) and “simultaneous credible intervals” (CI) of [4] to the present context of images. The changes needed to extend the algorithms designed for curves to images are straightforward. The HPW method (now “highest pixel-wise probabilities”) is a greedy algorithm that, starting from the pixels with the highest individual credibility, assembles image segments that are credibly jointly positive or negative. The CI method, on the other hand, computes the joint posterior credible intervals for the posterior mean and flags pixels blue or red according to whether the intervals are positive or negative.

3 The models

3.1 The image priors

For two pixel locations $s$ and $t$, write $s \sim t$ if they are neighbors. An interior pixel location $s = (i, j)$ has the four neighbors $(i - 1, j), (i + 1, j), (i, j - 1)$ and $(i, j + 1)$. The boundary and the corner pixels have three and two neighbors, respectively. Our first prior for the true underlying image $x$ is

$$p(x|\lambda_0) \propto \lambda_0^{n_i/2} \exp \left( -\frac{1}{2} \lambda_0 x^T Q_i x \right), \quad i = 1, 2,$$

(3)

where

$$x^T Q_1 x = \sum_{s \sim t} (x_s - x_t)^2, \quad x^T Q_2 x = \sum_t \left( \sum_{s \sim t} x_s - 4x_t \right)^2,$$

(4)

and $\lambda_0 > 0$ is a parameter. In the definition of $Q_1$, the summation is over all unordered pairs of neighboring pixels. The $n \times n$ matrix $Q_1$ is symmetric and positive semidefinite with rank $n - 1$ and we therefore take $n_1 = n - 1$ in the normalizing constant of the prior. In fact, it is easy to see that the null space of $Q_1$ consists of constant images. It follows that this prior is sensitive only to the intrinsic variation of the image pixels relative to their mean, not the absolute level of the pixel values. The integral of (3) over $x$ is not finite and therefore this function cannot represent a proper prior. However, $p(x|\lambda_0)$ can be used express our belief about the smoothness of the image $x$. The quantity $x^T Q_1 x$ measures the variability of neighboring pixel values and a small value of $\lambda_0$ allows very rough images to have a significant prior probability, whereas for a large value of $\lambda_0$ only very smooth images are likely. Such a “smoothing prior” is an example of a intrinsic Gaussian Markov random field [19].
Fig. 2. Sample images from the smoothing priors. Each image is scaled to the interval [0, 255]. The leftmost image is from the first difference prior and the rest of the images are from the second difference prior with the torus, Dirichlet and Neumann boundary conditions, respectively. In all four $\lambda_0 = 1$ and the same random number generator seed was used.

For an interior pixel $t$, the sum over $s$ in $x^T Q_2 x$ of (4) can be recognized as the discrete Laplacian computed at $t$. The prior therefore penalizes for image roughness as measured by the second differences of neighboring pixel intensities. In our experiments we found it useful to modify $Q_2$ slightly by defining four neighbors for the boundary pixels $t$, too. In the first modification the neighbors $s \sim t$ outside the image proper are taken to have zero intensity. In the second modification the boundary values themselves are extended beyond the actual image to produce the required neighbors. The third modification is to assume that $Q_2$ is a circulant matrix. This means that the left and right edges of the image are neighbors and a pixel $(M, j)$ on the bottom edge is a neighbor of the pixel $(1, j + 1)$ on the top edge. With these assumptions the image topologically resembles a torus. We refer to these three modifications as Dirichlet, Neumann and torus boundary conditions, respectively. The Dirichlet-modified prior is also a proper probability density and therefore $n_2 = n$ in (3). The Neumann- and torus-modified priors have the same one-dimensional null space as the first difference prior and therefore in both cases $n_2 = n - 1$.

The differences between the priors are illustrated in Figure 2, where a sample image from each is displayed. The sampled intensities have been rescaled to the interval [0, 255] separately for each image. For the first order, Neumann and torus priors we fixed the value of the upper left hand corner pixel at zero and then sampled from the resulting proper conditional density. In all five images $\lambda_0 = 1$ and the same random number generator seed was used. The images generated from the second order prior have a wider pixel intensity range (masked by scaling here) and, as expected, they appear to be much smoother than the image generated from the first order prior.

Many images include edges, features with large intensity changes that are not noise artifacts. In such images the above simple smoothing priors may not be able to reduce the noise sufficiently without inadvertently smoothing out the actual edges. Then an edge preserving prior is needed. Winkler [21] has
proposed a prior of the form

\[ p(x|\rho, \tau) \propto \exp \left( -\sum_{s \sim t} \varphi_{\rho,\tau}(x_s, x_t) \right), \] (5)

where, for \( a, b \in \mathbb{R} \), \( \varphi_{\rho,\tau}(a, b) = \min(\tau, \rho^2(a - b)^2) \) and summation is over all unordered pairs of neighboring pixels. This prior sets a penalty proportional to the square of the difference between neighboring pixels if their absolute difference is smaller than \( \delta \) and a constant penalty \( \tau \) if it is larger than \( \delta \). It therefore smooths small differences between neighboring pixels but preserves large ones assuming that they are actual features of the underlying image.

Depending on the problem at hand, instead of the exponents (4) in the prior (3), other quadratic forms or, say, absolute values instead of squares could also be used allowing one to better take into account application specific prior beliefs about the structure of the image \( x \). More complicated priors can be introduced by using the techniques described for example in [11], [13], and [21].

3.2 The posteriors

By (1), the likelihood of \( y \) given \( x \) and \( \Sigma \) is

\[ p(y|x, \Sigma) \propto |\Sigma|^{-1/2} \exp \left( -\frac{1}{2} (y - x)^T \Sigma^{-1} (y - x) \right). \] (6)

If \( \Sigma \) is known, substituting (6) and (3) in (2) one gets

\[ p(x|y) \propto \exp \left[ -\frac{1}{2} \left( \lambda_0 x^T Q_i x + (y - x)^T \Sigma^{-1} (y - x) \right) \right]. \]

Note that even though the priors of \( x \) may be improper, the posterior is a proper density, the multivariate normal distribution

\[ N \left( m_i, (\Sigma^{-1} + \lambda_0 Q_i)^{-1} \right), \] (7)

where

\[ m_i = S_{\Sigma,i,\lambda_0} y \equiv (I + \lambda_0 \Sigma Q_i)^{-1} y. \] (8)

In most cases \( \Sigma \) probably cannot be treated as a known quantity. In principle, one could then use an inverse Wishart prior for it as in [5]. However, in practice, the size of the covariance matrix makes sampling from the posterior infeasible. In the next two subsections we discuss some special cases of \( \Sigma \) for which posterior sampling is in fact possible.
3.2.1 Independent and identically distributed noise

If $\Sigma = \sigma^2 I$ the likelihood (6) reduces to

$$p(y|x, \sigma^2) \propto \sigma^{-n} \exp \left[ -\frac{1}{2\sigma^2} (y - x)^T (y - x) \right].$$

When the noise variance $\sigma^2$ is unknown, we use a scaled inverse $\chi^2$-prior for it,

$$p(\sigma^2) = \text{Inv-}\chi^2(\sigma^2 | \nu_0, \sigma_0^2) \propto \sigma^{-n} \exp \left[ -\frac{1}{2\sigma^2} \nu_0 \sigma_0^2 \right]. \quad (9)$$

The parameters $\nu_0$ and $\sigma_0^2$ are chosen to reflect our prior beliefs about the level of noise in the observed image $y$.

To facilitate analytical integration we modify the prior of $x$ so that it depends also on $\sigma^2$. The prior of $x$ is now

$$p(x | \sigma^2, \lambda_0) \propto \left( \frac{\lambda_0}{\sigma^2} \right)^{n_i/2} \exp \left( -\frac{1}{2\sigma^2} x^T Q_i x \right),$$

where $Q_i$ is either the first or the second difference matrix. The posterior of $(x, \sigma^2)$ is then

$$p(x, \sigma^2 | y) \propto (\sigma^2)^{-\left(\nu_0 + n + n_i + 2\right)/2} \exp \left[ -\frac{1}{2\sigma^2} \left( \lambda_0 x^T Q_i x + \|x - y\|^2 + \nu_0 \sigma_0^2 \right) \right], \quad (10)$$

$i = 1, 2$, $n_1 = n - 1$, $n_2 = n$ for the Dirichlet and $n_2 = n - 1$ for the Neumann and the torus versions of the prior. Note that, again, even though the priors of $x$ may be improper, the posteriors are proper densities. As we are mainly interested in $x$, the parameter $\sigma^2$ can be integrated out in (10) to get the marginal posterior of $x$,

$$p(x | y) \propto \left( \lambda_0 x^T Q_i x + \|x - y\|^2 + \nu_0 \sigma_0^2 \right)^{-\nu_0 + n + n_i}/2.$$

For $\lambda > 0$, let $S_{i,\lambda} = (I + \lambda Q_i)^{-1}$, $i = 1, 2$. Completing the square we see that the posterior of $x$ is actually a multivariate t-distribution,

$$x|y \sim \text{t}_{df_i}(m_i, \Sigma_i), \quad (11)$$

with mean $m_i = S_{i,\lambda_0} y$ and covariance

$$\Sigma_i = \frac{\|y\|^2 - y^T m_i + \nu_0 \sigma_0^2}{df_i} S_{i,\lambda_0},$$

where $df_i = \nu_0 + n_i$ are the degrees of freedom. The matrix inversion $(I + \lambda Q_i)^{-1}$ can be computed fast by noting that when Neumann or torus boundary conditions are used, the eigenvalue decomposition of $Q_2$ can be obtained with
the help of the discrete cosine (DCT) or Fourier transformations (DFT), respectively [10]. DCT can also be used with the first difference prior precision matrix \( Q_1 \).

The posterior image \( x|y \) can be directly sampled from this multivariate t-distribution. However, with the edge prior (5), hybrid Gibbs sampling from the joint posterior of \( x \) and \( \sigma^2 \) must be used since the marginal posterior of \( x \) is not any standard distribution. Hence, one first samples pixelwise the values of \( x \) given \( \sigma^2 \) by using the Metropolis-Hastings algorithm and then uses Gibbs sampling to obtain a new value for \( \sigma^2 \) given \( x \).

So far we have considered the smoothing parameter \( \lambda_0 \) in the prior of \( x \) as fixed. While we sometimes may be able choose such a fixed value in a reasonable manner, a fully Bayesian approach is to define a hyperprior for \( \lambda_0 \). To make Gibbs sampling possible, we use a Gamma hyperprior,

\[
p(\lambda_0) = \text{Gamma}(\lambda_0|\eta, \beta) \propto \lambda_0^{(\eta - 1)} \exp\left(-\beta \lambda_0\right).
\]

More details on finding a useful value for \( \lambda_0 \) or for the hyperparameters \( \eta \) and \( \beta \) can be found in section 3.3 and in Appendix A.

For random \( \lambda_0 \), the joint posterior distribution of \( x, \sigma^2 \) and \( \lambda_0 \) is

\[
p(x, \sigma^2, \lambda_0|y) \propto p(x|\sigma^2, \lambda_0)p(\sigma^2)p(\lambda_0)p(y|x, \sigma^2)
\]

\[
\propto \lambda_0^{\frac{n_i + \eta - 1}{2}} (\sigma^2)^{-\left(\frac{n_i + n + n_i + 2}{2}\right)} \exp\left[-\frac{1}{2\sigma^2} \left(\lambda_0 x^T Q_1 x + ||x - y||^2 + \nu_0 \sigma_0^2\right) - \beta \lambda_0\right].
\]

The full conditionals are standard distributions and therefore Gibbs sampling can be used,

\[
x|\sigma^2, \lambda_0, y \sim N\left(m_i, \sigma^2 S_{\lambda_0}\right),
\]

\[
\sigma^2|x, \lambda_0, y \sim \text{Inv-Gamma}\left((n_0 + n + n_i)/2, [\lambda_0 x^T Q_1 x + ||x - y||^2 + \nu_0 \sigma_0^2]/2\right),
\]

\[
\lambda_0|x, \sigma^2, y \sim \text{Gamma}\left(n_i/2 + \eta, x^T Q_1 x/(2\sigma^2) + \beta\right).
\]

### 3.2.2 Isotropic noise

If the noise is assumed to be stationary and isotropic the correlation between two pixels intensities depends only on the distance of the pixel locations. If also toroidal boundary conditions (cf. (3.1)) are used, the covariance matrix
\( \Sigma \) of the noise is circulant. For simplicity, we also assume that \( n \) is odd. Thus,

\[
\Sigma = \begin{bmatrix}
\rho_0 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\
\rho_{n-1} & \rho_0 & \rho_1 & \cdots & \rho_{n-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & \rho_0
\end{bmatrix},
\]

where \( \rho_i = \rho_{n-i}, \; i = 1, \ldots, n-1 \). This covariance matrix can be decomposed as \( \Sigma = \mathbf{V} \text{diag}(\gamma) \mathbf{V}^* \), where \( \mathbf{V} = n^{-1/2}[e^{-2\pi i (k-1)(l-1)/n}]_{k,l=1}^n \) is the DFT transformation matrix, \( \mathbf{V}^* \) is the complex conjugate of \( \mathbf{V} \), and \( \gamma = [\gamma_1, \ldots, \gamma_0]^T = \sqrt{n} \mathbf{V} [\rho_0, \ldots, \rho_{n-1}]^T \). We have that \( \mathbf{V}^* = \mathbf{V}^{-1} \) and, because of symmetry of \( \Sigma \), its eigenvalues are real and \( \gamma_i = \gamma_{n-i-2}, \; i = 2, \ldots, n \) (cf. [18]). For a fixed \( \Sigma \), by using the toroidal boundary conditions also in the prior of \( \mathbf{x} \), the posterior expectation and covariance in (7) are now easy to calculate using the computation rules of circulant matrices. Posterior sampling is fast using the methods explained e.g. in [19, section 2.6.3].

We have also tried the fully Bayesian approach that treats both \( \Sigma \) and \( \lambda_0 \) as random variables. The covariance matrix \( \Sigma \) is determined by its \( m = \lfloor n/2 \rfloor \) first eigenvalues. Hence, instead of setting a prior for whole covariance matrix, it is sufficient to set a prior for these eigenvalues. By examining the form of the posterior distribution of \( (\mathbf{x}, \gamma, \lambda_0) \), we see that Gibbs sampling is possible if inverse Gamma distributions are used as priors for \( \gamma_j, \; j = 1, \ldots, m \),

\[
p(\gamma_j) = \text{Inv-Gamma}(\gamma_j|l_j, a_j) \propto (1/\gamma_j)^{l_j+1} \exp (-a_j/\gamma_j), \; j = 1, \ldots, m,
\]

and the Gamma prior distribution (12) is chosen for the smoothing parameter \( \lambda_0 \). Then the joint posterior distribution of \( \mathbf{x}, \gamma \) and \( \lambda_0 \) is

\[
p(\mathbf{x}, \gamma, \lambda_0|\mathbf{y}) \propto p(\mathbf{x}|\lambda_0)p(\gamma)p(\lambda_0)p(\mathbf{y}|\mathbf{x}, \gamma) \propto \lambda_0^{m+\eta-1} \prod_{j=1}^m \gamma_j^{-(l_j+2)} \exp \left[ -\frac{1}{2} \left( \lambda_0 \mathbf{x}^T \mathbf{Q}_i \mathbf{x} + (\mathbf{y} - \mathbf{x})^T \Sigma^{-1}(\mathbf{y} - \mathbf{x}) \right) - \sum_{j=1}^m \gamma_j^{-1} a_j - \beta \lambda_0 \right]
\]

\[
\propto \lambda_0^{m+\eta-1} \prod_{j=1}^m \gamma_j^{-(l_j+2)} \exp \left[ -\frac{1}{2} \left( \lambda_0 \mathbf{x}^T \mathbf{Q}_i \mathbf{x} + \sum_{j=1}^n \gamma_j^{-1} \mathbf{f}_j^* \mathbf{f}_j \right) - \sum_{j=1}^m \gamma_j^{-1} a_j - \beta \lambda_0 \right],
\]

where \( \mathbf{f} = [f_1, \ldots, f_n]^T = \mathbf{V}(\mathbf{y} - \mathbf{x}) \), and the \( \gamma_j \)'s are assumed independent a priori, \( p(\gamma) = \prod_{j=1}^m p(\gamma_j) \). The full conditionals for Gibbs sampling therefore
are
\[ x|\gamma, \lambda_0, y \sim N \left( (I + \lambda_0 \Sigma Q_i)^{-1} y, (\Sigma^{-1} + \lambda_0 Q) \right), \]
\[ \gamma_1|x, \lambda_0, y \sim \text{Inv-Gamma} \left( l_1 + 1, \frac{f_i f_j + f_{n-j+2} + a_j}{2} \right), j = 2, \ldots, m, \]
\[ \lambda_0|x, \lambda, y \sim \text{Gamma} \left( \frac{n}{2} + \eta, \frac{x^T Q x}{2} + \beta \right). \]

3.3 Estimation of parameters

The priors suggested above depend on parameters and hyperparameters that need to be chosen somehow. Under the isotropic model, a parametric noise covariance matrix can be fitted using for example the estimated variogram and this fit can be used as a basis for a prior. For the iid noise model a maximum likelihood estimate of \( \sigma^2 \) can be used to set \( \nu_0 \) and \( \sigma_0^2 \) because, for the prior expectation of \( \sigma^2 \) we have \( E(\sigma^2) = \frac{\nu_0}{\nu_0 - 2} \sigma_0^2 \). A plausible way to find a reasonable fixed value for \( \lambda_0 \) is to compute the mean of \( x|y \) for different choices of \( \lambda_0 \) and to pick a value that appears to produce a good reconstruction of \( x \) from the noisy observation \( y \). If such a simple approach is not feasible, one can try to obtain a reasonable value for \( \lambda_0 \) by using the observed image \( y \) for maximum likelihood estimation or cross validation. The estimated value of \( \lambda_0 \) can then be used as such or it may be used to suggest suitable values for the hyperparameters \( \eta \) and \( \beta \) in (12). Location and width of the prior and hyperprior distributions should reflect our prior beliefs and uncertainty about the values of \( \sigma^2 \) and \( \lambda_0 \).

In the maximum likelihood (ML) method we consider the data generating mechanism (1) where \( x \) and \( \epsilon \) are independent. The computation of the maximum likelihood estimates of \( \sigma^2 \) and \( \lambda_0 \) is discussed in Appendix A. Generalized cross validation (GCV) was tried for the estimation of \( \lambda_0 \) (e.g. [13,17]). As explained in Appendix A, by using the GCV-estimate of \( \lambda_0 \) an estimate for \( \sigma^2 \) can then also be obtained. In ordinary cross validation each pixel is left out in turn and its value is then predicted by a fit to the rest of the image. We employed the posterior mean fits \( m_i = S_{i, \lambda_0} y \) and \( m_i = S_{\Sigma i, \lambda_0} y \) (cf. (8), (11)) and tried to select \( \lambda_0 \) so that the fit is optimal in the sense of the cross validated prediction error. In the generalized cross validation approach used here we actually minimized with respect to \( \lambda_0 \) a related quantity defined by

\[ \text{GCV}(\lambda_0) = \frac{\frac{1}{n} \| y - S_{i, \lambda_0} y \|^2}{\left[ \frac{1}{n} \text{tr}(I - S_{i, \lambda_0}) \right]^2}, \]

and similarly for \( S_{\Sigma i, \lambda_0} \). As before, Fourier methods can be used to accelerate ML- and GCV-estimation. In our experiments we did not try to estimate the
<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\lambda}_0$</td>
<td>$\hat{\sigma}$</td>
</tr>
<tr>
<td>1st diff</td>
<td>0.97</td>
<td>78.1</td>
</tr>
<tr>
<td>Dirichlet</td>
<td>2.79</td>
<td>86.82</td>
</tr>
<tr>
<td>Neumann</td>
<td>3.75</td>
<td>87.3</td>
</tr>
</tbody>
</table>

Table 1
Estimates of $\lambda_0$ and $\sigma$ with different priors and different estimation methods for the Blobs test image.

parameters $\rho$ and $\tau$ for the edge prior (5) but picked values that appeared to produce a good edge preserving reconstruction of the underlying image.

4 Experiments

All computations were carried out on a Dell OptiPlex 755 PC with Intel Core2 Duo E6750 CPU and 8Gb of RAM using Matlab 2008B under the Linux Ubuntu operating system. Scripts and more examples can be found at http://cc.oulu.fi/~lpasanen/iBSiZer.

4.1 The Blobs

The 100 $\times$ 100 “Blobs” test image is shown in the middle panel of the top row of Figure 3. It was obtained by first sampling from normal distributions within the segments depicted in the left panel and then blurring the result. The normal distribution means in the segments were -250, -30, 0, 30 and 250. The “observed” image $y$ on the right was formed by adding to $x$ iid Gaussian noise with standard deviation $\sigma = 90$. Because the corrupted pixel values also must lie in the interval [-255, 255], values outside this range were discarded and therefore the random error in reality has a truncated normal distribution and our model (1) for the observed data is only an approximation. Note that only the smallest areas with the highest absolute intensities are clearly visible in this noisy image.

The results of parameter estimation are displayed in Table 1. Estimations took less than a second each. The slight underestimation of $\sigma$ indicates that some of the noise is probably misinterpreted as image structure. The second row of Figure 3 shows reconstructions of the Blobs image using the iid noise model with different priors and GCV-parameter estimates. The corresponding ML-estimate based reconstructions looked very similar. The more
heavily smoothing second order difference prior is more consistent with the structure of the true image and should therefore be used instead of the first difference prior that leaves the reconstructed image too noisy. By examining the second difference reconstructions closely one observes that the Dirichlet boundary conditions cause an unwanted tendency towards zero intensities near the boundary. This is undesirable because the Blobs image contains important features in the boundary areas. The GCV-based second difference prior with the Neumann boundary condition was therefore used in the scale space analysis shown in Figure 4. The sample size was 25000 and the parameter values used were $\lambda_0 = 2.3$, $\nu_0 = 50$, and $\sigma^2_0 = 84.0^2$ (note that $85.7^2 \approx 84.0^2(50/48)$).

Posterior sampling took about 90 seconds and the three maps (PW, HPW and CI) about 2 minutes per smoothing level. For the finest resolution, the pixelwise (PW) maps exhibit many false features whereas the simultaneous maps (CI and HPW) perform well. In the HPW map with $\lambda = 1$ some high-amplitude noise combines with the underlying large low intensity (-30 and 30) features to produce a few scattered red and blue areas that could be mistaken for small scale features. The corresponding CI map does not have this problem. Note how, for all methods, increasing $\lambda$ brings out these large scale, low intensity features quite effectively.

Next, in a fully Bayesian approach we treated also the smoothing parameter $\lambda_0$ as random. Using the Neumann prior GCV-estimate 2.33 for $\lambda_0$ as a reference value, we used the prior Gamma($2.3^2$, 2.3) for $\lambda_0$. The mean and the variance of this distribution was 2.3, 1, respectively. Hence the prior is very vague. The burn-in period for the Gibbs sampler was 20000 iterations and af-
Fig. 4. iBSiZer analysis of the Blobs image. The Blobs image $x$ and its noisy version $y$ are displayed in the first row. The Smooths column displays posterior means $\mathbb{E}(S_{\lambda}x|y)$ computed as sample means for different values of the smoothing parameter $\lambda$. Reconstruction of $x$ was based on a second difference Neumann prior with smoothing parameter $\lambda_0 = 2.3$ chosen on the basis of a GCV-estimate of $\lambda_0$. This reconstruction corresponds to the $\lambda = 0$ panel in the Smooths column. The yellow circle in the smooths indicates the approximate size of the smoothing kernel. The three other columns show scale space analyses of the smooths using three different methods to flag credible features: pixelwise (PW), highest pixelwise probability (HPW), and simultaneous credible intervals (CI). The level of credibility $\alpha = 0.95$. 
that every 50th \((x, \sigma^2, \lambda_0)\) triplet was picked for the final random sample. After 10000 samples had been obtained, iterations were re-initialized and the sampling process was started anew to gather the remaining 10000 samples for the final sample of 20000. In all, sampling took an hour. The posterior mean and standard deviation of \(\lambda_0\) and \(\sigma^2\) were 3.66, 0.34, 87.2\(^2\) and 120 respectively. The mean value of \(\lambda_0\) is now close to its ML-estimate 3.75 and the mean value of \(\sigma^2\) is closer to the true image noise variance 90\(^2\). The reconstruction is somewhat smoother than in Figure 4 but the maps themselves look very similar (result not shown). Note also that, although the estimates of \(\sigma^2\) for the Blobs image suggests that some noise is interpreted as structure, this actually leads to very few erroneous flaggings of small scale features in the iBSiZer analyses as shown in Figure 4.

4.1.1 Isotropic noise

To explore the performance of iBSiZer with correlated noise, instead of iid errors we next added isotropic Gaussian noise to the Blobs image. However, in order to have an odd number \(n\) of pixels the last row and the last column on the image matrix were first deleted. The added noise had the spherical covariance [3]

\[
\Sigma(h, c) = \begin{cases} 
  c_0 + c_s, & h = 0, \\
  c_s - c_s\{(3/2)(||h||/c_r) - (1/2)(||h||/c_r)^3\}, & 0 < ||h|| \leq c_r, \\
  0, & ||h|| \geq c_r,
\end{cases}
\]

where \(h\) is the distance between two pixel locations, \(c = [c_0, c_r, c_s]^T\), \(c_0, c_s, c_r \geq 0\), and we took \(c_0 = 0\), \(c_r = 4\), and \(c_s = 50^2\). The noisy image is shown in the top row of Figure 5. The corrupted pixel values need to lie in the interval [-255, 255] but, instead of discarding the values outside this range, they were now truncated and therefore the random error in reality has a censored normal distribution and our model (1) for the observed data is again only an approximation. Note in Figure 5 how the correlated noise gives rise to clusters of pixels that could easily be misinterpreted as structure of the true image.

For the estimation of the noise covariance matrix we first subtracted from \(y\) its Nadaraya-Watson kernel smooth and then fitted the parametric spherical variogram model to the residuals. The Nadaraya-Watson smoothing parameter value was selected so that the covariance parameters estimated reasonably close to their known true values. The estimates for \(c_0, c_s\) and \(c_r\) were 0, 3.56 and 48.5\(^2\). For image reconstruction, the GCV- and ML-estimates of \(\lambda_0\) were 1.9 \(\cdot\) 10\(^{-11}\) and 5.0 \(\cdot\) 10\(^{-4}\), respectively. In the iBSiZer analysis presented in Figure 5 we used the value 5.0 \(\cdot\) 10\(^{-4}\). Posterior sampling took about 45 seconds. Note how, despite a reconstruction \((\lambda = 0)\) which is worse than in Figure 4, the scale space analysis finds essentially the same features as before.
For a fully Bayesian analysis we treated both $\lambda_0$ and the eigenvalues $\gamma_j$ (cf. section 3.2.2) as random with hyperpriors $\lambda_0 \sim \text{Gamma}(5 \cdot 10^{-8}, 5 \cdot 10^{-4})$ and $\gamma_j \sim \text{Inv-Gamma}(l_j, a_j)$, where $l_j = \hat{\gamma}_j^2 / 10^8 + 2$, $a_j = \hat{\gamma}_j^3 / 10^8 + \hat{\gamma}_j$, and $\hat{\gamma}_j$ is the eigenvalue of the covariance matrix estimate corresponding to $\gamma_j$. With these choices, $E(\gamma_j) = \hat{\gamma}_j$ and $\text{Var}(\gamma_j) = 10^8$, a priori, which corresponds to a rather vague prior. A sample of 20000 (10000 + 10000) was generated from the joint posterior distribution using a burn-in of 20000 iterations and keeping every 50th triplet $(x, \gamma, \lambda_0)$. This took about 2 hours. From two leftmost columns on Figure 6 we can see that the extra uncertainty introduced by these priors results in a loss of credible features, especially in the larger scales.

For comparison, we finally repeated the analysis by assuming (erroneously) that the noise is uncorrelated and estimated $\sigma$ and $\lambda_0$ by the GCV- and ML-methods. ML-estimates were $\hat{\lambda}_0 = 0.0239$ and $\hat{\sigma} = 14$ and GCV-estimates were $\hat{\lambda}_0 = 0.034$ and $\hat{\sigma} = 12$. Given that the marginal pixelwise error standard deviation is in fact 50, it is obvious that the reconstruction would interpret much of the noise as structure. We therefore used instead the values $\lambda_0 = 2.5$, $\sigma^2_0 = 50^2$ (the true pixel noise variance) and $\nu_0 = 10$ in the priors. However, the corresponding iBSiZer-maps presented in the two rightmost columns of Figure 6 still flag by far too many features as credible clearly demonstrating the dangers of model misspecification.

4.2 The Desktop

The second artificial example is the difference between two $475 \times 501$ digital camera images of an office desktop. Independent Gaussian noise with standard deviation 10 was added to the two images so that the noise in the difference image has standard deviation $\sigma = \sqrt{10^2 + 10^2} \approx 14$. The two images and their difference are displayed in the first row of Figure 7. Because of the several sharp, small scale features present, we used the edge prior (5) for this test image. After some experimenting the parameter values $\rho = 0.0389$ and $\tau = 1.5$ were found to produce a reasonably good reconstruction. This means that differences smaller than $\sqrt{\tau}/\rho \approx 31.5$ are interpreted as noise. The prior for $\sigma^2$ was $\text{Inv-}\chi^2(50, 14^2)$ and hybrid Gibbs sampling was used to produce a sample of 700 $(x, \sigma^2)$ pairs from their posterior distribution. Posterior sampling took several hours.

For comparison, the Neumann prior was also tried. Large image size forced us to estimate $\lambda_0$ and $\sigma^2$ based only on the 300 first image columns. Estimation took 90 seconds with both methods. ML- and GCV-estimates for $\lambda_0$ and $\sigma$ were 0.89, 0.13 and 17.2, 13.6, respectively. The GCV-estimate of $\sigma$ is close to true value, whereas ML-overestimates it. We therefore used the GCV-based values $\lambda_0 = 0.13$ and $\sigma^2_0 = 12.16^2$, $\nu_0 = 10$. A sample of 1000 was drawn in
Fig. 5. iBSiZer analysis of the Blobs image with correlated noise. See the text and the caption of Figure 4 for more details.

The true difference image \( x \) along with the posterior means of \( x \) are shown on the second row of Figure 7. Posterior means are only slightly smoother than the original image. The third and fourth rows of Figure 7 present the iBSiZer analyses based on the edge prior and the second order Neumann prior. Computation of the credibility maps took about 90 seconds per smoothing level. The difference between the two priors shows in the smallest scale where analysis with the smoothing Neumann prior misses some of the finest detail. With \( \lambda = 0 \) only the highest intensity differences are credible. Increasing \( \lambda \) suppresses the smaller details and for \( \lambda = 5 \cdot 10^4 \) both maps are dominated by the positive change caused by the white paper while the keys and part of the desk mostly account for some average local negative change.

4.3 Landsat images

Our final example is a preliminary analysis of a pair of Landsat ETM+ satellite images kindly provided to us by Professor Erkki Tomppo and Dr. Kai Mäkisara of the Finnish Forest Research Institute. The two images are shown in Figure 8.
Fig. 6. On left, iBSiZer analysis of the Blobs image with correlated noise and hyperpriors for $\lambda_0$ and the eigenvalues of $\Sigma$. On right, noise is erroneously assumed to be iid. See the text and the caption of Figure 4 for more details.

They were taken over eastern Finland on August 2, 1999 (leftmost panel) and May 29, 2002 (middle panel), respectively. The images are in fact small $124 \times 176$ subimages of the original full Landsat images in spectral band 5 (in the mid-infrared area). To better visualize intensity variation, the images and their difference are scaled to the interval $[0,255]$ while keeping the zero intensity level fixed.

From Figure 8 one sees that the image difference exhibits features in many spatial scales. The black area in the middle of the two images is part of a lake. One might assume that the lake area has not changed between the two satellite images and that therefore the difference image intensity variations here are just noise. The mean and the standard deviation of the intensities in the whole difference image are 2.3 and 11.3, respectively, while for the lake area they are 0.4 and 1.2. If iid noise is assumed, we could use the estimated lake variance $1.2^2$ as an estimate of $\sigma^2$. The ML-estimates for $\lambda_0$ from (A.5) for the first and the second order Neumann priors are then $\hat{\lambda}_0 = 0.008$ and $\hat{\lambda}_0 = 0.002$, respectively. The reconstructions (see Figure 9 for the first difference prior reconstruction) look very much like the original unsmoothed difference image and the iBSiZer atlases (not shown) appear rather noisy with the lake disappearing already at moderate smoothing levels. Results were similar for the edge prior. We then estimated the parameters from the full difference image and obtained the values shown in Table 2. The first and the second difference
Fig. 7. Reconstructions and iBSiZer analyses of the Desktop image. First row: two images of an office desk and the noisy Desktop test image. Second row: the true difference image and its reconstructions using an edge prior and the second order Neumann prior. Third row: CI-maps based on an edge prior. Fourth row: CI-maps based on the second order Neumann smoothing prior. See the text and the caption of Figure 4 for more details.

reconstructions based on the GCV-estimates are displayed in Figure 9.

Scale space analyses of the Landsat difference image are shown in Figure 10. In the left column the first order difference prior with $\lambda_0 = 0.09$ and the vague Inv-$\chi^2(10, 4.3^2)$ prior for $\sigma^2$ are used. The posterior distribution sample size was 10000 and it was drawn in about 60 seconds. At the lowest level of smoothing even the tiniest changes are flagged as credible and, as smoothing is increased, aggregate features corresponding to small-area averages start to show up. However, the lake area appears to fade away too soon making the
Fig. 8. The original Landsat subimages and their difference. Images are scaled to the interval [0,255] while keeping zero intensity level fixed. Copyright 1999 ESA and 2002 ESA.

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\lambda}_0$</td>
<td>$\hat{\sigma}$</td>
</tr>
<tr>
<td>1st diff</td>
<td>0.068</td>
<td>4.5</td>
</tr>
<tr>
<td>Neumann</td>
<td>0.36</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Table 2
Estimates of $\lambda_0$ and $\sigma$ with different priors and different estimation methods for the Landsat difference image.

Fig. 9. First row from left: the original difference image, the reconstruction based on the first difference prior and $\hat{\lambda}_0 = 0.008$ obtained using the value $\hat{\sigma} = 1.2$ estimated from the lake area, and a first difference prior reconstruction using GCV-estimates. Second row: second difference reconstruction using the GCV-estimates, second difference reconstruction with a random $\lambda_0$, and a reconstruction assuming isotropic noise. Scaling is used to emphasize intensity variation.
level $\lambda = 400$ hard to interpret. The middle column uses the second order Neumann prior together with the vague hyperprior Gamma$(0.14^2, 0.14)$ for $\lambda_0$ and the prior Inv-$\chi^2(10, 4.7^2)$ for $\sigma^2$. A sample of 10000 ($5000 + 5000$) was generated from the joint posterior distribution in a couple of hours using a burn-in of 20000 iterations and keeping every 50th triplet ($x, \sigma^2, \lambda_0$). The posterior means and the standard deviations for $\lambda_0$ and $\sigma$ were 0.36, 0.018 and 6.01, 0.57, respectively. Reconstruction is now rather smooth (cf. Figure 9) and the smallest features are not visible at any smoothing scale. Still, the major features have been found quite well and the maps have a clean, uncluttered appearance that facilitates their interpretation.

Finally, we moved away from the iid noise setting assuming isotropy, instead. The last row and the last column were deleted from the image in order to make the number of pixels $n$ odd. For covariance estimation we first subtracted from $y$ its Nadaraya-Watson kernel smooth choosing the smoothing parameter so that only noise, not features in the image appeared to be smoothed out. The residuals were then used for parametric estimation of the covariance matrix $\Sigma$ using the variogram. The spherical covariance model (13) was employed as it seemed to fit the data best among the parametric models we tried. Estimates for $c_0, c_r$, and $c_s$ were $4.78^2, 2.32$, and $5.77^2$, respectively. The second order difference prior with circulant boundary conditions was used. The estimated value for $c_r$ suggests some short range spatial correlation under the Gaussian spherical covariance model. The ML- and GCV-estimates for $\lambda_0$ were $3.3 \cdot 10^{-2}$ and $1.3 \cdot 10^{-3}$, respectively. The iBSiZer analysis produced by this approach, however, was unsatisfactory (results not shown). We then treated $\lambda_0$ and $\Sigma$ as random by setting hyper priors $\lambda_0 \sim \text{Gamma}(1.3^2 \cdot 10^{-6}, 1.3 \cdot 10^{-3})$ and $\gamma_j \sim \text{Inv-Gamma}(l_j, a_j)$, where $l_j = \hat{\gamma}_j^2 / 10^6 + 2$, $a_j = \hat{\gamma}_j^3 / 10^6 + \hat{\gamma}_j$ and $\hat{\gamma}_i$ is the eigenvalue of the estimated covariance matrix. A sample of 10000 ($5000 + 5000$) was generated from the joint posterior distribution using a burn-in of 20000 iterations and keeping every 50th triplet ($x, \gamma, \lambda_0$). This sample was drawn in about three hours and its scale space analysis shown in the right column of Figure 10. As expected, both the increased uncertainty from random $\lambda_0$ and $\Sigma$ together with a fitted noise model that allows more structure lead to fewer features being flagged as credible. Compared with analysis based on the iid model using the Neumann prior and random $\lambda_0$ (middle column of Figure 10), there are fewer credible features, especially in the smallest scales.

Looking at the scale space analyses in Figure 10, we note that some of the finest scale credible changes between the two Landsat images lie in forest areas and may well correspond to small stands of trees cut down during the period between the acquisition of the two satellite images. The prominent medium scale negative (red) feature in the lower corner is credible in most scales and can therefore safely be assumed to be “really there”, and not just an artifact resulting from random noise. In fact, inspection of a map of the region in question reveals that this feature corresponds to a marsh area and
Fig. 10. iBSiZer analysis of the Landsat difference image using CI-maps. Left column: first order difference prior. Middle column: second order Neumann prior with a random $\lambda_0$. Right column: noise assumed isotropic, random $\lambda_0$ and $\gamma$. See the text and the caption of Figure 4 for more details.

its significance may therefore well be related to a seasonal change at this site. For coarser spatial resolutions, statistically significant mean changes in increasingly large areas are detected and this is potentially useful for forest inventory purposes where estimation of regional averages of forest resources is of interest.
5 Discussion and further work

In this paper we have proposed a method for extracting credible or “statistically significant” features from a digital difference image. In particular, the goal was to detect credible features in different spatial scales or image resolutions. Bayesian inference was used, because it facilitates flexible modeling, straightforward inference and the possibility to incorporate in the analyses useful prior knowledge about the images at hand. Although we focused on the detection of differences, our method can be used for more general feature extraction where pixel intensities are compared to some reference field.

Many different ways to reconstruct a degraded image have been proposed in the image analysis literature, including numerous Bayesian and frequentist approaches (see e.g. [11,16,21] and the references therein). For the problem of finding image features, Bayesian image segmentation was used e.g. in [15] to classify pixels in DMRI breast images into benign and malignant tissue. Multiscale image segmentation has also been discussed in many papers, including [12] and [20]. The novelty of our approach is the combination of image reconstruction with joint posterior probability based inference for the detection of credible image features for a range of image resolutions simultaneously. As opposed to, say, simple thresholding of a reconstructed image, iBSiZer offers a principled approach to making probability statements about the scale dependent features of the underlying true image. Visualization of the analysis using credibility maps provides an effective means for summarizing the results of statistical inference. The experiments demonstrated the relevance of the scale space approach as analysis in just one scale usually does not allow one to discover all the interesting features of the image because its salient structure in fine and coarse resolutions is usually very different.

An edge preserving prior was tried in two examples (the Desktop and the Landsat images) but mostly we have used a Gaussian smoothing prior for the underlying image. The noise, either iid or isotropic, was also assumed to be Gaussian. This setting allows one to utilize fast computation techniques based on Fourier methods and sparse matrix algorithms. In many cases one then achieves a speed which facilitates reasonably interactive data analysis. Still, in more complicated settings, such as in the fully Bayesian analyses of our examples, the computations can take hours making interactive work impractical. Also, although our model can be easily adjusted to allow non-Gaussian noise and more complex priors, unless one is able to utilize fast computational methods, slow posterior sampling may be the price one has to pay for such generalizations. Speeding up computations in these situations is therefore one area of future work. One should note, however, that once the reconstruction sample has been produced, the scale space analysis itself remains as fast as before, even in these more complicated settings.
The prior distributions involve parameters whose values need to be set. As one often lacks firm prior knowledge about the “true” values, we suggested methods to estimate reasonable values for them. These parameter values can be used as such or they can be used to find reasonable hyperpriors for the parameters in question. This approach appeared to work well in our experiments.

The last example consists of two Landsat satellite images. The data were analyzed assuming that the noise is either iid or, more generally, isotropic. Finding statistically significant mean changes in areas of various sizes is important for forest inventory and this makes our multi-scale smoothing based approach useful in this context. More detailed assessment of the results of this preliminary analysis would clearly require knowing more about the ground truth but such information was not available for us. Still, even at this stage of its development, iBSiZer clearly shows promise as a tool for satellite based forest inventory purposes.

One area of future work is to increase the performance of the proposed method by fine-tuning it for the purposes of forest inventory and other specific applications. Improvements could be gained for example from more realistic noise modeling and from utilizing priors that are better tailored to a particular image domain. Another line of research is to extend iBSiZer to more general spatial settings. A natural next step is the analysis of spatial fields defined on a regular lattice. Our preliminary experiments in this direction have been promising.

Acknowledgement

We would like to thank the Referees and the Associate Editor for useful comments that led to improvements of the original manuscript. The Associate Editor’s suggestion for computational speed-ups were particularly useful.

A Appendix: Maximum likelihood estimation

Here we consider maximum likelihood estimation of the parameters. For clarity, denote the prior densities of \( x \) and \( \varepsilon \) by \( p_x(x|\Sigma, \lambda_0) \) and \( p_\varepsilon(\varepsilon|\Sigma) \), respectively. Taking \( x \) and \( \varepsilon \) to be independent in (1), the density of \( y = x + \varepsilon \) is given by the convolution

\[
p(y|\Sigma, \lambda_0) = \int p_\varepsilon(y - x|\Sigma)p_x(x|\Sigma, \lambda_0)dx
\]

\[
\propto \int \lambda_0^{n/2}|\Sigma|^{-1/2} \exp \left[ -\frac{1}{2} \left( \lambda_0 x^T Q_i x + (y - x)^T \Sigma^{-1}(y - x) \right) \right] dx. \tag{A.1}
\]
Of course, in order for this to be precise, \( x \) should have a true density and this may not be the case. We will deal with this shortly but, for now, let us work with (A.1). Completing the square in the exponent and integrating over \( x \) we can write the likelihood of the data \( y \) as

\[
L(\Sigma, \lambda_0 | y) = \lambda_0^{\frac{n_i}{2}} |\Sigma|^{-1/2} |\tilde{S}_{\Sigma,i,\lambda_0}|^{\frac{1}{2}} \exp \left[ -\frac{1}{2} \left( -(\Sigma^{-1}y)^T \tilde{S}_{\Sigma,i,\lambda_0} (\Sigma^{-1}y) + y^T \Sigma^{-1}y \right) \right],
\]

where \( \tilde{S}_{\Sigma,i,\lambda_0} = (\Sigma^{-1} + \lambda_0 Q_i)^{-1} \). From this we get the log likelihood as

\[
\ell(\Sigma, \lambda_0 | y) = \log L(\Sigma, \lambda_0 | y) = \frac{n_i}{2} \log \lambda_0 - \frac{1}{2} \log |\Sigma| + \frac{1}{2} \log |\tilde{S}_{\Sigma,i,\lambda_0}| - \frac{1}{2} y^T (\Sigma^{-1} - (\Sigma^{-1}) \tilde{S}_{\Sigma,i,\lambda_0} \Sigma^{-1}) y.
\]  

(A.2)

Now, as pointed out above, \( p_x(x|\Sigma, \lambda_0) \) may not be a true density – it takes on a constant value on any linear submanifold of \( \mathbb{R}^n \) parallel to the null space of \( Q_i \). The dimension of this null space is \( n - n_i \) so that \( Q_i \) has the eigenvalue decomposition

\[
Q_i = U \text{diag}(0, \ldots, 0, \delta_{i,n-n_i+1}, \ldots, \delta_{i,n}) U^T,
\]

with the first \( n - n_i \) eigenvalues equal to zero and \( \delta_{i,n-n_i+1}, \ldots, \delta_{i,n} > 0 \). We can then make \( p_x(x|\Sigma, \lambda_0) \) a true density by replacing \( Q_i \) by

\[
Q_i^\kappa = U \text{diag}(\kappa, \ldots, \kappa, \delta_{i,n-n_i+1}, \ldots, \delta_{i,n}) U^T,
\]

where \( \kappa > 0 \). For small \( \kappa \) this density is close in spirit to what the possibly improper density signifies. We can therefore derive the log likelihood with \( Q_i^\kappa \) in place of \( Q_i \) and then let \( \kappa \to 0^+ \) to justify (A.2).

After some straightforward calculation one finds that

\[
\frac{\partial \ell(\Sigma, \lambda_0 | y)}{\partial \lambda_0} \propto \frac{n_i}{2 \lambda_0} - \frac{1}{2} \text{tr}(\tilde{S}_{\Sigma,i,\lambda_0} Q_i) - \frac{1}{2} (\tilde{S}_{\Sigma,i,\lambda_0} \Sigma^{-1}) y^T Q_i (\tilde{S}_{\Sigma,i,\lambda_0} \Sigma^{-1}) y.
\]  

(A.3)

If an estimate for \( \Sigma \) is known, it can be substituted here and an estimate for \( \lambda_0 \) is obtained by solving (A.3) numerically for \( \lambda_0 \) (c.f. sections 4.1.1 and 4.3).

With iid noise the likelihood is

\[
L(\sigma^2, \lambda_0 | y) \propto \left( \frac{\lambda_0}{\sigma^2} \right)^{\frac{n_i}{2}} |S_{i,\lambda_0}|^{\frac{n_i}{2}} \exp \left[ -\frac{1}{2 \sigma^2} \left( -y^T S_{i,\lambda_0} y + ||y||^2 \right) \right].
\]

Now, setting the partial derivative \( \partial \ell(\sigma^2, \lambda_0 | y) / \partial \sigma^2 \) equal to zero one obtains \( \sigma^2 \) as a function of \( \lambda_0 \),

\[
\sigma^2 = \frac{1}{n_i} y^T (I - S_{i,\lambda_0}) y.
\]  

(A.4)
The partial derivative $\frac{\partial \ell(\sigma^2, \lambda_0 | y)}{\partial \lambda_0}$ is

$$\frac{\partial \ell(\sigma^2, \lambda_0 | y)}{\partial \lambda_0} = \frac{n_i}{2\lambda_0} - \frac{1}{2} \text{tr}(S_{i,\lambda_0} Q_i) - \frac{1}{2\sigma^2} (S_{i,\lambda_0} y)^T Q_i (S_{i,\lambda_0} y). \quad (A.5)$$

One can now substitute (A.4) into the right hand side of (A.5), set it equal to zero and solve numerically to get an estimate $\hat{\lambda}_0$ for $\lambda_0$. Substituting then $\hat{\lambda}_0$ into (A.4) one obtains an estimate $\hat{\sigma}^2$ for $\sigma^2$. A GCV-estimate of $\lambda_0$ can similarly be used to get an estimate for $\sigma^2$.

References


