1-1-1992

A Generalized Gaussian Image Model for Edge-Preserving MAP Estimation

Charles Bouman
Purdue University School of Electrical Engineering

Ken Sauer
University of Notre Dame, Department of Electrical Engineering

A Generalized Gaussian Image Model for Edge-Preserving MAP Estimation

Charles Bouman
Ken Sauer

TR-EE 92-1
January 1992

A Generalized Gaussian Image Model for Edge-Preserving MAP Estimation

Charles Bouman
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907-0501
(317) 494-0340

Ken Sauer
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
(219) 239-6999

March 19, 1991

Abstract

We present a Markov random field model intended to allow realistic edges in maximum a *posteriori (MAP)* image estimates, while providing stable solutions. Similar to the generalized Gaussian distribution used in robust detection and estimation, we proposed the generalized Gaussian Markov random field (GGMRF). This model satisfies several desirable analytical and computational properties for MAP estimation, including continuous dependence of the estimate on the data, invariance of the character of solutions to scaling of data, and a solution which lies at the unique local minimum of the a *posteriori* log likelihood function. The GGMRF is demonstrated to be useful for image reconstruction in low dosage transmission tomography.

1 Introduction

Many important problems in image processing and computer vision require the estimation of an image or other 2D field, X, from noisy data Y. For example, tomographic reconstruction and 2D depth estimation are two seemingly dissimilar problems which fit into this structure. When the data is of good quality and sufficient quantity, these problems may be solved well by straightforward deterministic inverse formulae. However, when data is sparse or noisy, direct inversion is usually excessively sensitive to noise. If the data is sufficiently sparse, the inverse problem will be underdetermined or ill-posed. In such cases, the result can be significantly improved by exploiting prior information about X's behavior.

Bayesian estimation is a statistical approach for incorporating prior information through the choice of an *a priori* distribution for the random field X. While many Bayesian estimation techniques exist, a common choice for image estimation problems is the maximum *a posteriori* (MAP) estimator. The MAP estimate has the appealing attribute that it yields the most likely image given the observed data. In addition, it results in an optimization problem which may be approached using a variety of established techniques.

The specific choice of prior distribution for X is, of course, a critical component in MAP estimation. The Markov random field (MRF) has been applied widely during the recent past[1, 2, 3, 4], due to its power to usefully represent many image sources, and the local nature of the resulting estimation operations. A variety of distinct models exist within the class of MRFs, depending on the choice of the *potential functions*. Each potential function characterizes the interactions among a local group of pixels by assigning a larger cost to configurations of pixels which are less likely to occur. In particular, we will restrict our attention to potential functions $\rho(x_i - x_j)$, which act on pairs of pixels. The shape of $\rho(\Delta)$, where Δ is the difference between pixel values, then indicates the attributes of our model for X.

One of the more troublesome elements of applying MRFs to image estimation is coping with edges. Because most potential functions penalize large differences in neighboring pixels, sharp edges are often discouraged. This is especially true for the Gaussian MRF, which penalizes the square of local pixel differences. Many approaches to ameliorate this effect have been introduced. Geman and Geman[2], incorporated a "line process" into their MRF to describe sharp discontinuities. Others limited the penalty of any local difference at some prescribed threshold[5, 6], or created other potential functions which become flat at large magnitudes of their arguments[7, 8, 9]. Since such functions are non-convex, the global optimization required in MAP estimation can not be exactly computed, and an approximate MAP estimate must be used. In addition, we show that there is a second equally important liability to using MRFs with non-convex potential functions: the MAP estimate may not be a continuous function of the input data. This means that the position of the $\hat{X}$ with globally minimal cost may undergo a large shift due to a small perturbation in Y. Therefore, the MAP estimator is an unstable and ill-posed inverse operation.

Several researchers have proposed the use of convex potential functions. Stevenson and Delp[10] used the convex Huber function[11], which is quadratic for small values of A, and linear for large values. The point of transition between the quadratic and linear regions of the function is a predetermined threshold, T. Green[12] and Lange[13] included the strict convexity criterion, also for the sake of computational tractability. Green's choice of $\log\cosh(\Delta)$ has a shape similar to that of the Huber function, but with the transition point from approximately quadratic to approximately linear at $T = 1$. Lange also derived several other potential functions in [13], each satisfying convexity and several other desired properties.

The restriction to convex potential functions makes the computation of the exact MAP estimate feasible, but the approaches listed above still exhibit a limitation: their effect in MAP estimation is dependent on the scaling of X and Y. The transition threshold for the Huber function, for example, should be related to the magnitude of edges expected in X. If

this magnitude is unknown, or edges of widely varying magnitudes are expected, then the smoothing of these edges may be inconsistent. Typically, fine edges may be blurred and large edges accentuated. Similar difficulties hold for the other non-quadratic functions mentioned.

In this paper, we introduced an MRF model for Bayesian estimation, which is intended to ameliorate both of the problems discussed above. The general form of the potential function is $|\Delta|^p$, with $1 \leq p \leq 2$. The resulting form of the probability density function for X is similar to the generalized Gaussian distribution commonly used as a noise model in robust detection and estimation[14]. Due to this similarity, we use the name generalized Gaussian Markov random field (GGMRF) to describe these images. The parameter p controls the cost of abrupt edges. When $p = 1$ sharp edges are no more costly than smooth edges, and when $p = 2$ the familiar Gaussian assumption holds.

The log of the GGMRF has two important properties. It is convex, and it scales with the data. Convexity makes minimization efficient and leads to a stable MAP estimator. The scaling property leads to a homogeneous MAP estimator when the observation noise has the generalized Gaussian distribution with a corresponding form. We also give the canonical form for all distributions which have the convexity and scaling properties.

We briefly explore the connection between median filtering and MAP estimation using the GGMRF prior together with the generalized Gaussian noise model. The recursive weighted median filter results as the local update operation for computation of MAP estimate when $p = 1$. However, it is shown that the local median filter updates do not converge to the global MAP estimate. This connection is of interest since median filters are a useful class of homogeneous edge preserving nonlinear filters for image processing.

In the experimental section of this paper, the GGMRF is applied to the problem of image reconstruction from integral projections. Bayesian techniques have been applied to similar problems, but most previous assumptions for prior distributions have been Gaussian[15, 16, 17]. We consider the transmission tomographic case, with low X-ray dosage, and attendant

high photon counting noise. This noise is especially problematic in projection rays passing through highly absorptive regions; in the limit these regions are effectively radio-opaque, and present the equivalent of the hollow projection (a.k.a. "bagel") problem. Reconstructions using the convolution backprojection algorithm suffer from a trade-off between excessive blurring and noise artifacts. A similar trade-off, with better results, can be made in using a Gaussian MRF as a prior density on X. The GGMRF, however, with smaller values of p, allows the formation of sharp edges, while more effectively suppressing the photon counting noise in the estimate. The success of the GGMRF in regularization of the tomographic reconstruction offers hope that it will be useful in many other image restoration and reconstruction problems.

2 Statistical Framework

2.1 MAP Estimation and Gibbs Distributions

We first define some basic notation. A random field X will be defined on the set of N points S, and each pixel, X_s for $s \in$ S, takes values in $\mathbb{R}$. The neighbors of X_s will be denoted by $X_{\partial s}$ where $\partial s \subset$ S. Further, the neighbors of each point must be chosen so that they have the property that Vs, $r \in$ S $s \notin \partial s$ and $r \in \partial s \Leftrightarrow s \in \partial r$.

There are two important criteria which must be met by estimates of an image, X, from data, Y. First, X must accurately fit the data. A natural measure of this fit is the log likelihood function

$$L(y|x) = \log \mathrm{P}(\mathrm{Y} \in dy | X = \mathrm{x})$$

where $\mathrm{P}(\mathrm{Y} \in dy | X = \mathrm{x})$ is the density function of Y given X. The maximum likelihood (ML) estimate is the estimate which best fits the data.

$$\hat{X} = \arg \max_{\mathrm{z}} \mathrm{L}(\mathrm{Y}|\mathrm{x})$$

The ML estimate often does not satisfy the second criterion, which is the incorporation

of reasonable prior information about the the behavior of the image. In practice this can cause undesirable behavior or nonuniqueness[18] of the result. In the reconstruction from projections problem, the ML estimator may contain excessive high frequency variation which we would not expect to exist in the original image. If the number of data samples in Y is smaller than the number of unknown pixel values in X, the solution will be underdetermined. Similar problems of underdetermined or ill-posed solutions occur in a wide variety of problems in motion estimation[19], surface reconstruction[20] and edge detection[21].

One approach to incorporating prior information into the solution is to adopt a prior distribution for the unknown image, $g(x) = \mathrm{P}(\mathrm{X} \in \mathrm{dx})$. The logarithm of the a *posteriori* distribution of X given Y may then be computed using Bayes' formula.

$$\begin{aligned} L_p(x|y) &\triangleq \log P(X \in dx|Y = \mathrm{y}) \\ &= L(y|x) + \log g(x) - \log P(Y \in \mathrm{dy}) \end{aligned}$$

This posterior distribution may then be incorporated into a Bayesian estimation technique such as maximum a *posteriori* (MAP) estimation.

$$\begin{aligned} \hat{X} &= \arg\max_x L_p(x|Y) \\ &= \arg\max_x \{L(Y|x) + \log g(x)\} \\ &= \arg\max_x \{L(Y, \mathrm{x})\} \end{aligned} \tag{1}$$

The last equation indicates that the MAP estimate also maximizes the log of the joint distribution, $L(y, \mathrm{x}) = \log \mathrm{P}(\mathrm{Y} \in \mathrm{dy}, \mathrm{X} \in \mathrm{dx})$.

When the prior distribution of X is Gaussian, the log likelihood $\log g(x)$ will be a quadratic function of x. If $\mathrm{P}(\mathrm{Y} \in dy|X = \mathrm{x})$ is also Gaussian, the MAP estimate corresponds to $\mathrm{E}\{\mathrm{X}|Y\}$, and is therefore the minimum mean squared error estimate[22]. When the prior distribution is not Gaussian, the optimality properties of the MAP estimator are less clear[18]. However, MAP estimation is computationally direct and has experimentally been shown to work well in a variety of problems[1, 2, 3, 4, 23].

A critical issue is the choice of prior distribution for X. We will use Markov random fields (MRF) since they restrict computation to be local but still include a very wide class of possible models. Gibbs Distributions are used to explicitly write the distributions of MRF's. A Gibbs distribution is any distribution which can be expressed in the form

$$p(x) = \frac{1}{Z} \exp\left\{ - \sum_{c \in C} V_c(x) \right\}$$

where Z is a normalizing constant, $V_c(\cdot)$ is any function of a local group of points c and C is the set of all such local groups. The key to the definition of the Gibbs distribution is the specification of these local groups of points. A local set of points, c, is called a clique if Vs, $r \in c$, s and r are neighbors. If Gibbs distributions are restricted to use functions of cliques induced by the neighborhood system ds, then the random field X will have the property that

$$\forall s \in S \quad p(x_s | x_r \; r \neq s) = p(x_s | x_{\partial s}) \; .$$

This is the fundamental property of an MRF. In fact, the Hammersley-Clifford theorem states that under some technical conditions, a random field is a MRF if and only if it has a probability distribution corresponding to a Gibbs distribution[24, 25].

2.2 Gaussian Markov Random Fields

A common choice for the prior model is a Gaussian Markov random field (GMRF)[15, 16, 17]. The distribution for a Gaussian random field has the form

$$g(x) = \frac{\lambda \sqrt{2}}{(2\pi)^{N/2}} |B|^{1/2} \exp\left\{ -\lambda^2 x^t B x \right\} \; . \tag{2}$$

where B is a symmetric positive definite matrix, λ is a constant, and x^t is the transpose of x. If the model is homogeneous, then the matrix B is Toeplitz-block-Toeplitz. In this case, we assume that the diagonal elements of B are unity, and, therefore, λ^{-2} is equal to twice the prediction variance of a pixel, X,, given its neighbors, $X_{\partial s}$. In order for this to correspond to a Gibbs distribution with neighborhood system ds, we also impose the constraint that

$B_{sr} = 0$ when $s \notin \partial r$ and $s \neq r$. This distribution may then be rewritten to form the log likelihood

$$\log g(x) = -\lambda^2 \left(\sum_{s \in S} a_s x_s^2 + \sum_{\{s,r\} \in C} b_{sr} |x_s - x_r|^2 \right) + c$$

were $a_s = \sum_{r \in S} B_{sr}$ and $b_{sr} = -B_{sr}$. Notice that the second sum is now over all distinct pairs of neighboring pixels. MAP estimation of X then results from minimization of the following cost function:

$$\hat{x} = \arg\min_x \left\{ -L(y|x) + \lambda^2 \left(\sum_{s \in S} a_s x_s^2 + \sum_{\{s,r\} \in C} b_{sr} |x_s - x_r|^2 \right) \right\}.$$

The GMRF prior has a number of analytical advantages when the conditional distribution of Y given X is also Gaussian. First, since the function to be optimized is quadratic it is also convex. Therefore, the global minimum will be unique and feasible to compute. This model also has the advantage that the minimum mean squared error estimate of X, the conditional expectation of X given Y, and the MAP estimate of X all coincide. Finally, the Gaussian structure inherits a wealth of techniques for choosing and estimating model parameters.

Unfortunately, the Gaussian prior distribution results in estimates of X which are either excessively noisy or generally blurred. The blurring effect is particularly undesirable along the edges that often occur in real images. In fact, this same dilemma arises in many problems which require the estimation of local image properties. Examples of such local properties are depth in surface reconstruction[20], or velocity in motion estimation[19]. In general, these problems are underdetermined and require the incorporation of prior information.

2.3 Nonconvex Log Prior Distributions

Non-Gaussian MRF's are interesting because they can potentially model both the edges and smooth regions of images. Many initial approaches have utilized non-Gaussian MRF's with highly nonconvex log likelihood functions. These models incorporate an additional unobserved random field called a line process which determines the location of edges[2, 26].

While this approach is intuitively appealing and innovative, it makes minimization very difficult and introduces a variety of additional model parameters to choose or estimate.

More recently, many approaches have focused on MRF's with simpler Gibbs distributions of the form

$$\log g(x) = - \sum_{\{s,r\} \in C} b_{sr} \rho(|x_s - x_r|) \quad + \quad \text{constant} \tag{3}$$

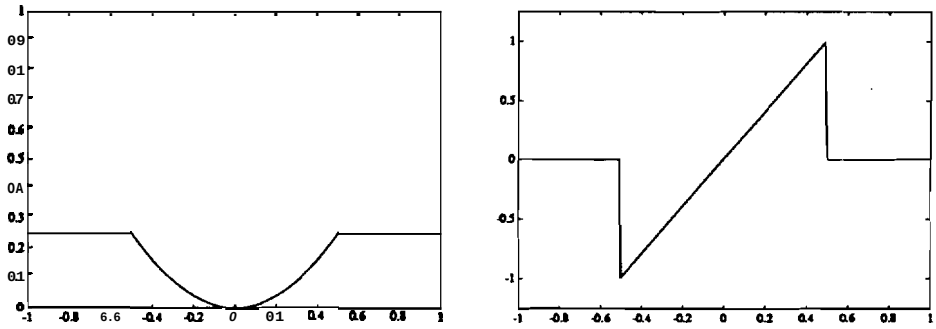
where p is a monotone increasing, but not convex function[7, 8, 5, 6, 9, 12, 13]. A typical function used by Blake and Zisserman[5] is

$$\rho(\Delta) = (\min\{|\Delta|, T\})^2 \ .$$

This function is shown in Fig. 1a for T = 0.5. Notice that the function is quadratic near zero, but the flat region beyond the value T allows sharp edges to form in the reconstructed image. The derivative of p determines the tendency of neighbors in X to be attracted and plays a role analogous to the influence function of robust statistics[11, 27]. This function is shown in Fig. 1b. Notice that for values greater than T the model does not encourage pixels to be closer in value. If two pixels differ by a value greater than T, then it is likely that they lie on opposite sides of an edge, and therefore their values should not be required to be close. However, we will see in Section 3.2 that convex functions can also achieve this desirable goal.

For the purposes of modeling the prior distribution on images, this distribution has some significant practical and theoretical disadvantages. Since the function is nonconvex it is generally impractical to globally minimize. Instead, the MAP estimate can only be approximated using a number of different techniques[2, 5, 6]. In fact, the solution often depends substantially on the method used to perform the minimization.

In addition to this computational issue, there is a disadvantage to the quality of reconstruction that results from such a nonconvex prior. In practice, once the size of an image edge significantly exceeds the value T, there is no longer any tendency for the estimated edge in $\hat{x}$ to be smooth. This means that the MAP estimate may abruptly change as the



(a) $\rho(\Delta) = (\min\{|\Delta|, 0.5\})^2$ (b) $\frac{d\rho(\Delta)}{d\Delta}$ = influence function

Figure 1: a) A typical nonconvex cost function. p is a function of A the difference between neighboring pixel values. b) The derivative of p represents the attraction between two points separated by A.

magnitude of an edge in the input data Y increases. This often leads to a unnatural quality in the reconstruction, in which reconstructed edges greater than T are sharp, yet edges less than T are smooth.

Another undesirable quality in these reconstructions is due to the fact that the MAP estimate, $\hat{x}$, is not continuously dependent on the input, y. To illustrate this point consider the nonconvex functions shown in Fig. 2. Fig. 2a shows a function with two local minima at positions x_1 and x_2. Fig. 2b show a small perturbation on the first function with the same two local minima. Notice that while the difference between the two functions is small the difference between the two global minima is large. In addition, there is clearly an intermediate point between the two function for which the solution is not unique. This phenomenon can produce drastic effects in MAP estimates of a random field. For example, a small spatial cluster of pixels with larger values than their surrounding neighbors may be suppressed by the cost of their border. If the potential function is nonconvex, this cluster may abruptly appear in the MAP estimate, due to a very small change in data.

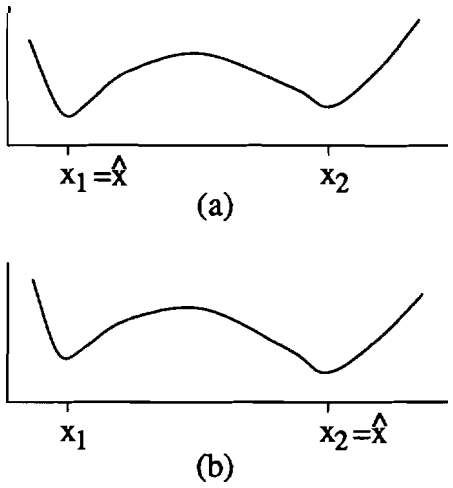


Figure 2: This figure illustrate how small changes in a nonconvex function can result in large changes in the location of the functions minimum value.

2.4 Regularization and MAP Estimation

The discontinuous dependence of a solution on data has long been viewed as a undesirable property of any inverse problem. Hadamard originally defined a problem to be well posed if its solution:

- exists,
- is unique,
- depends continuously on the data.

When the solution does not depend continuously on the data, the problem is often called unstable since the solution can change dramatically with small perturbations of the data.

The problem of regularizing ill-posed (not well posed) problems has been the subject of much research[18, 19, 20, 21]. In particular, Tikhonov has introduced methods for regularizing deterministic problems by introducing stabilizing functionals which play a role analogous to the log prior distribution of MAP estimation[28]. In this work, Tikhonov also determined

that these stabilizing functionals must meet certain conditions to guarantee that the resulting problem is well posed. In particular, the stabilizing functionals are required to be "quasimonotone" to eliminate the possibility that the solution can discontinuously jump. A quasimonotone function is defined by Tikhonov to be one which contains no local minima other than the global minima. We modify Tikhonov's original definition slightly and define the following.

Definition 1 *A local minimum of a function is any point which is the minimum on some local open neighborhood of the point.*

Definition 2 *The functional* $h : U \rightarrow \mathbb{R}$ *is called quasimonotonic if* $h(\cdot)$ *has a unique global minimum,* $h(x_o)$*, at the point* $x_o \in U$*,* $h(\cdot)$ *contains no other local minima, and there exists a number* $b > h(x_o)$ *such that* $\{x \in U : h(x) \leq b\}$ *is compact.*

We may now use the concept of quasimonotonicity to prove a theorem which gives sufficient conditions for the MAP estimator to be well posed. These theorems are in the spirit of Tikhonov's work but are specifically designed for our purposes, and are based on unrelated methods of proof given in Appendix A and B.

Theorem 1 *Let* $f(\cdot,\cdot)$ *be a continuous functional* $f : U \times V \rightarrow \mathbb{R}$ *such that for all* $y \in V$ $f(\cdot, y)$ *is quasimonotonic then*

$$\arg\min_{x \in U} f(x, y)$$

is a continuous function of y*.*

Theorem 2 *Any strictly convex function* $f: \mathbb{R}^N \rightarrow \boldsymbol{R}$ *with a local minimum is quasimonotonic.*

These two theorems may then be combined with the properties of probability density functions to yield the following result.

Corollary 1 Let X and Y be finite dimensional random objects. If $L(y, x) = \log P(X \in dx, Y \in dy)$ exists and is a strictly convex function of (x, y), then the MAP estimate is a well posed operation.

2.5 Convex Log Prior Distributions

The practical difficulties of minimization and the implications of instability are both disadvantages of using nonconvex log likelihood functions for a prior distribution. Stevenson and Delp considered a similar issue in the reconstruction of surfaces from range information[10]. The problem of surface reconstruction differs slightly from our problem since the possibility of both edge and roof discontinuities (first and second derivatives) must be included in the prior distribution. Stevenson and Delp's search for an alternative convex energy function was largely motivated by the intractable nature of nonconvex minimization. They chose the the Huber function first introduced in robust statistics[11].

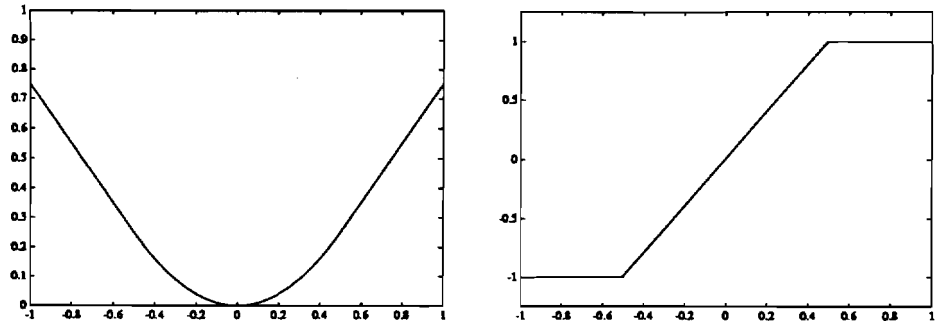
$$\rho(\Delta) = \begin{cases} \Delta^2 & \text{if } |\Delta| \leq T \\ T^2 + 2T|\Delta - T| & \text{if } |\Delta| > T \end{cases}$$

This function and its corresponding influence function are shown in Fig. 3 for T = 0.5. For values greater than T the linear region of this function also allows sharp edges, yet convexity makes the MAP estimate efficient to compute.

In separate but related work, Green[12] employed the function

$$\rho(\Delta) = \log \cosh(\Delta),$$

which produces useful Bayesian estimates of emission tomograms, while providing the aforementioned computational advantages. This $\rho(\Delta)$ is qualitatively similar to that of Huber's since it is quadratic near zero and linear for large A. Lange derived several other strictly convex potential functions in in a study of convergence of the expectation maximization algorithm[13].



(a) $\rho(\Delta)$ = Huber function (b) $\frac{d\rho(\Delta)}{d\Delta}$ = influence function

Figure **3:** a) The convex cost function used by Stevenson and Delp. ρ is a function of A the difference between neighboring pixel values. b) The derivative of ρ represents the attraction between two points separated by A.

While these methods worked well in their applications, and represent important conceptual steps, the behavior of their MAP estimates depends on the absolute scale of data. For example, the fact that a single value of T must be chosen for the Huber function is still a significant limitation when modeling images. Even if the value of T could be estimated accurately, it is not clear that a single value of T can accurately describe real images. In practice, all edges in an image do not have a single size. Therefore, real edges of magnitude less than T are smoothed while those edges greater than T are sharply reconstructed. This often has the effect of suppressing important but small detail, and can result in a reconstructed image with an unnatural appearance.

3 A Stable Scale Invariant Approach

The conclusion of the previous sections is that it is desirable for the log of the prior distribution to be convex and not depend on an absolute parameter of scale such as T. In fact, many image processing operations that have been widely adopted such as linear and median filtering[29] are homogeneous operations, and do not have dependence on absolute scale. Homogeneous operations have the property that scaling of the input data results in proportional scaling of the output image. Our approach is to look for prior distributions

which yield a homogeneous **MAP** estimation operator. This will result in a model that will reconstruct edges accurately without prior knowledge of their size.

3.1 Homogeneous MAP Estimation

The **MAP** estimator is homogeneous if for all real constants a and for all inputs y the following holds.

$$\arg\max_x L(\alpha y, x) = \mathrm{a}\arg\mathrm{rnax}_x \mathrm{L}(y, x)$$

This is equivalent to the relationship

$$\arg\max_x L(\alpha y, a\, x) = \arg\max_x L(y, x)\ .$$

Such an equality may be insured by requiring that L have the functional behavior

$$L(\alpha y, \alpha x) = \beta(\alpha, y)L(y, x) + \gamma(\alpha, y) \tag{4}$$

where β and y are functions of a and y. **A** reasonable method for assuring that (4) holds is to require that the prior distribution and likelihood of Y given X have the form

$$L(\alpha y|\alpha x) = \beta(\alpha)L(y|x) + \gamma_1(\alpha) \tag{5}$$

$$\log g(\alpha x) = \beta(\alpha)\log g(x) + \gamma_2(\alpha)\ . \tag{6}$$

These are the basic relations which we will use to enforce homogeneity in the **MAP** estimator. We call such functions scalable.

Definition 3 A *strictly positive function $g(x)$ is called scalable if for each constant* a, *there exist two constants β and y so that for almost every x*

$$\log g(\alpha x) = \beta \log g(x) + \gamma\ .$$

The form of the function $L(y|x)$ is usually determined by the physics of a problem. However, the restriction that it be scalable is not unreasonable. To see this, consider the

random vector, Z of independent and identically distributed random variables, Z_s, with the generalized Gaussian distribution[14]

$$P(Z_s \in \mathrm{dz}) = \frac{q}{2\Gamma(1/q)} \exp(-|z|^q) \tag{7}$$

parameterized by q. When q = 2 the components of Z have a Gaussian distribution. When q = 1 they have a Laplacian distribution, and for $1 < q < 2$ the distribution has intermediate behavior. This noise model is commonly used in robust statistical estimation since it captures the heavy tailed behavior that is often exhibited by real noise distributions. If Y has the form

$$\mathrm{Y} = \mathbf{A}X + \mathbf{D}^{-1}Z \tag{8}$$

where A and D are matrices, then $L(y|x)$ will be scalable. To see this notice that

$$\log L(y|x) = ||\mathbf{D}(Y - \mathbf{A}X)||_q^q + \text{constant} \tag{9}$$

where $||\cdot||_q$ is the l_q norm, and

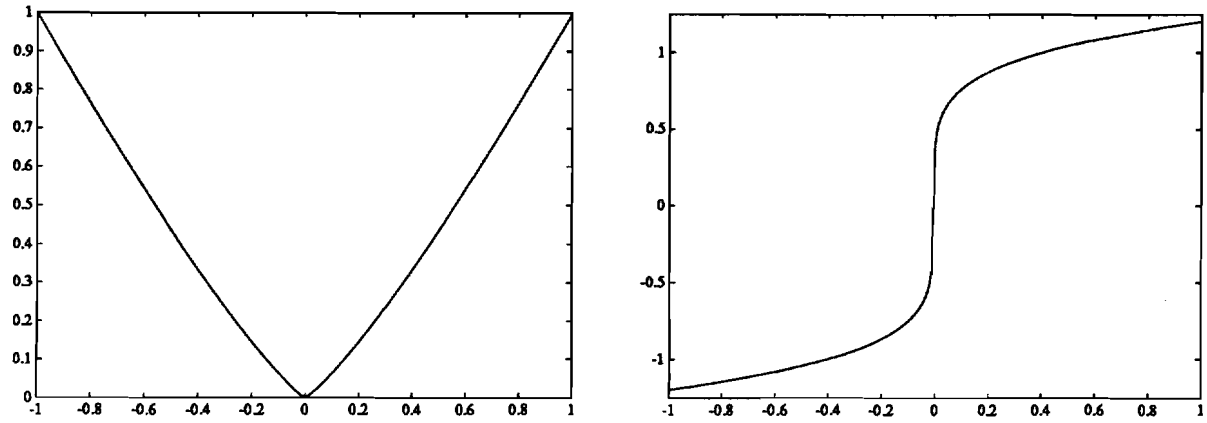
$$||\mathbf{D}(\alpha Y - \mathbf{A}\alpha X)||_q^q = \alpha^q ||\mathbf{D}(Y - \mathbf{A}X)||_q^q .$$

If in addition to $L(y|x)$ being scalable, $g(x)$ is scalable with the same constants α and β, then the MAP estimator will be homogeneous. In addition, we argued in Section 2.3 that the $\log g(x)$ should be a strictly convex function in order to insure stability of the solution to perturbations of the data. Enforcing these two requirements leads to the following theorem proved in Appendix C.

Theorem 3 The function g : $\#\mathrm{I}^N \rightarrow \mathbb{R}$ is a scalable density function with a convex log density function $\log g(x)$ if and only *if*

$$-\log g(x) = ||x||^p + c$$

for some norm $||\,.\,||$ and constants $p \geq 1$, and c.



(a) $\rho(\Delta) = |\Delta|^{1.2}$ (b) $\frac{d\rho(\Delta)}{d\Delta}$ = influence function

Figure 4: a) An example of the proposed scale invariant convex cost function when p = 1.2. p is a function of A the difference between neighboring pixel values. b) The derivative of p represents the attraction between two points separated by A.

3.2 Generalized Gaussian MRF

Theorem 3 leaves available a wide variety of possible choices for $g(x)$. However, we propose a simple generalization of Gaussian MRF's base on the concept of generalized Gaussian noise. This model has the functional form similar to (3), but uses $\rho(\Delta) = |\Delta|^p$,

$$\log g(x) = -\lambda^p \left(\sum_s a_s |x_s|^p + \sum_{\{s,r\} \in C} b_{s,r} |x_s - x_r|^p \right) + \text{constant}, \tag{10}$$

where $1 \leq p \leq 2$, and λ is a parameter which is inversely proportional to the scale of x. We call the class of random fields with this distribution generalized Gaussian Markov random fields (GGMRF) since this model is contained within the more general class of MRF's and includes all Gaussian MRF's when p = 2. As in the case of the GMRF, not all values of the parameters a, and b,,, will lead to a consistent model. In fact, $g(x)$ will be well defined only if $\log g(x)$ is a positive definite function of x. A sufficient condition for positive definiteness is that a, > 0 and b,,, > 0. This condition also insures that the $-\log g(x)$ is convex. In practice, we may choose a, $= 0$, which results in an ill defined density. However, this is not a practical difficulty since the function $L(y|x)$ causes the MAP estimate to be unique.

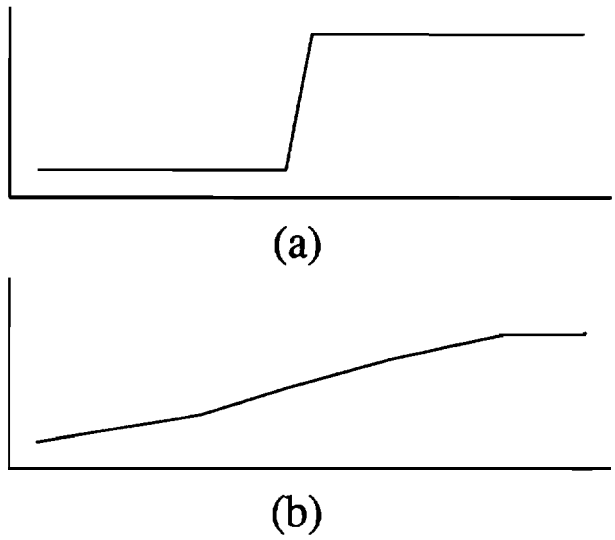


Figure 5: When p = 1, any monotone function which starts and ends at the same value has the same total cost. Therefore, both the sharp edge and the smooth edge have the same cost.

The choice of p is critical in determining the character of the model, Larger values of p discourage abrupt discontinuities while smaller values of p allow them. Figure 4a shows the function $\rho(\Delta) = |\Delta|^{1.2}$ and the corresponding influence function $1.2|\Delta|^{0.2}$. Notice that the tendency of pixels to be close in value increases more slowly as the separation increases. The special case of a one dimensional "image" x with p = **1** provides insight into edge reconstruction. For this case, the prior distribution has the form

$$\log g(x) = -\sum_{s=1}^{N-1} |x_s - x_{s+1}| + \text{constant} .$$

As long as x is a monotone (increasing or decreasing) function, then

$$\sum_{s=1}^{N-1} |x_s - x_{s+1}| = |x_1 - x_N| .$$

Therefore, the total cost is simply the difference between the starting and ending values. This means that abrupt edges in the reconstruction have no greater cost than smooth edges. Fig. 5 illustrates this fact and indicates that nonconvex functions are not required for the reconstruction of sharp edges.

Let us assume that the observed distortion has the form of (9) and the prior distribution is from a GGMRF. Then both the prior distribution of X, and the conditional distribution of Y given X will be scalable. If in addition, p = q then the α and β parameters for both distributions will be the same, and the MAP estimator will be a homogeneous operation. More generally, if we write the MAP estimate, $\hat{x}$, explicitly as a function of the input data, y, and the prior scale parameter, λ, it is easily shown that

$$\hat{x}(\alpha y, \lambda) = \alpha \hat{x}(y, \alpha^{1-q/p}\lambda) \ . \tag{11}$$

When p = q, the relation $\hat{x}(\alpha y, \lambda) = \alpha \hat{x}(y, \lambda)$ holds for all α, and the MAP estimator is homogeneous.

When p $\neq$ q, the MAP estimator is not homogeneous, since the distributions for the prior and observation noise no longer coincide. However, (11) indicates the qualitative behavior of the MAP estimate does not change as the input is scaled since the result is proportional to a MAP estimate using a different regularization constant, $\alpha^{1-q/p}\lambda$. More formally, we may define the set of all regularized solutions for each input y.

$$\hat{\mathcal{X}}(y) = \{\hat{x}(y, \lambda) : \lambda > 0\}$$

Then we say that for p $\neq$ q, the MAP estimate, $\hat{x}(y, \lambda)$, has the generalized homogeneity property since the set of all solutions scales with the input size.

$$\hat{\mathcal{X}}(\alpha y) = \alpha \hat{\mathcal{X}}(y) \tag{12}$$

4 Optimization Techniques

In this section, we discuss the minimization techniques which we will use to compute the MAP estimator. These methods are of fundamental importance for two reasons. First, they provide basic intuition for understanding MAP estimation using the GGMRF prior. Second, the minimization techniques connect the area of MAP estimation to the literature

in weighted median filtering[29, 30, 31]. Since median filtering has been shown to be of broad practical importance in image filtering, we believe this suggests that methods based on the GGMRF prior can also be practically useful in a variety of image estimation applications.

We will adopted a simplified problem for illustrating the issues of minimization. Assume that Y is formed by adding white noise to X,

$$Y = X + \sigma Z \tag{13}$$

where Z is defined in (7) and a is a scale parameter (not equal to the standard deviation). We will also assume that the prior model is a homogeneous MRF (i.e. $b_{s-r} = b_{r-s}$ is used in place of $b_{s,r}$), and the coefficients $a, = 0$. The MAP estimate is then given by

$$\hat{x} = \arg\min_{x} \left\{ \sum_{s \in S} |y_s - x_s|^q + \sigma^q \lambda^p \sum_{\{s,r\} \in C} b_{s-r} |x_s - x_r|^p \right\} . \tag{14}$$

Since this cost function is convex for p, q ≥ 1, finding a global minimum will be computationally feasible.

In general, (14) may be minimized using either global or local iterative techniques. Two examples of global iterative techniques are gradient descent and conjugate gradient[32]. At each iteration, these methods compute the multidimensional gradient of the cost function being minimized. The result is then used to determine a new solution which is closer to the global minimum. Local minimization methods iteratively minimize the cost function at each pixel, x_s, of x. Since X is a MRF, minimization of the cost function with respect to x_s results in the following simple local computation.

$$\hat{x}_s = \arg\min_{x_s} \left\{ |y_s - x_s|^q + \sigma^q \lambda^p \sum_{r \in \partial s} b_{r-s} |x_s - x_r|^p \right\} \tag{15}$$

This is equivalent to the local operation used in the method ICM proposed by Besag[3].

The discussion of minimization methods will be broken down into distinct cakes depending on the values of p and q. When p = q = 2 the well-known Gaussian case occurs. Here the reconstruction may be thought of as the best linear estimate with the resulting edge blurring

and nonrobustness to noise. The local minimization operation reduces to a linear average of the observed value y, and the neighbors of x_s.

$$\hat{x}_s = \frac{y_s + (\sigma\lambda)^2 \sum_{r\in\partial s} b_{r-s} x_r}{1 + (\sigma\lambda)^2 \sum_{r\in\partial s} b_{r-s}}$$

4.1 $p = q = 1$

When p = q = 1 the cost function is not strictly convex so corollary 1 does not apply.[1] However, this still represents an important limiting case as the distributions become heavy tailed. For p = q = 1 the cost function is a convex polytope in a high dimensional space. Along the edges of the polytope, the function is not differentiable. If for all r, $\sigma^q \lambda^p b_r = 1$, then the local minimization operation reduces to the median of the observed pixel value, y_s, and the pixel's neighbors,

$$\hat{x}_s = \text{median}\{y_s, \quad x_{r_1}, x_{r_2}, \ldots, x_{r_i}\}$$

where each pixel has i neighbors. Notice that this operation always includes the information from the original data, y_s. This keeps the MAP estimate from drifting too far from the original data.

As the signal-to-noise ratio goes to zero, $\sigma^q \lambda^p \to \infty$. If b_r is assumed constant, then the local minimization operation which results is

$$\hat{x}_s = \text{median}\{x_{r_1}, x_{r_2}, \cdots, x_{r_i}\} \ . \tag{16}$$

This is the recursive median filter which has been extensively studied in nonlinear filtering literature. Notice that this filter assigns no weight to the original data.

In the most general case of arbitrary coefficients, the solution of (15) is known as the weighted median. The weighted median is the value, $\hat{x}$, such that the total weight of pixels greater than $\hat{x}$ is as close as possible to the total weight of pixels less than $\hat{x}$. Since the

[1] In fact, counter examples to continuity may be easily given whenever p or q are 1.

weighted median has the flexibility to treat pixels differently as a function of position, it has attracted attention as a nonlinear filter for image processing[33, 34].

Median filters are known to be robust homogeneous filtering operations which preserve edges in practical image processing applications. So it is encouraging that they result as the local minimization operations of our **MAP** estimation problem. Surprisingly however, **MAP** estimation and median filtering are actually quite distinct because the local operations generally *do not* converge to the global **MAP** estimate! This happens because the local operations become "stuck" on the edges of the polytope. In fact, it is well known that the recursive median filter of (16) converges to a root signal (which is generally not constant)[30, 31]. However, since this operation results from assuming that $\sigma^q \lambda^p = \infty$, we know that the global **MAP** estimate must have the form x_s =constant for all s. In practice, we have found that the global **MAP** estimate may be computed by alternating a complete pass of local minimization with a single iteration of a gradient-based method. Since the cost function is not differentiable, a question arises as to how to compute the gradient. We use the approximation that

$$\frac{d|x|}{dx} = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} .$$

In general, the local minimization will not globally minimize the posterior distribution whenever p or q equals 1. This is due to the fact that the cost function is not differentiable in these cases. While it is possible to perform the minimization solely using gradient based methods, we have found the alternation of local minimization with gradient methods to be important in making minimization computationally efficient.

4.2 $1 < p, q < 2$

When $1 < p = q < 2$, the cost function is differentiable, and it may be easily shown that the iterative local minimization of (15) converges to the global **MAP** estimate. In this case, the global minimum is the only point for which the gradient is zero. This local operation is, of

course, nonlinear. When p = q, it is also homogeneous (as is the entire MAP estimator). The operation of (15) has the form of a least powers M-estimator used in robust statistics[27]. In practice, a value of q = 1.2 has been found to yield a good compromise between asymptotic efficiency and robustness for M-estimation in real data[27].

Due to the physical nature of a problem, we may often have $1 \leq p \neq q \leq 2$. In this case, the local operation for minimization is not homogeneous, though it does have the generalized homogeneity property described in Section 3.2.

5 Statistical Tomographic Reconstruction

In this section, we briefly describe the specific problem of statistical reconstruction of 2D cross-sections from integral projections. This inversion problem has been approached within the Bayesian framework for both emission and transmission tomography[16, 17, 7, 12].

The 2-D Radon transform maps a function of two variables, which we denote by $x(s_1, s_2)$, into a function indexed by (θ, t) according to

$$p(\theta, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(s_1, s_2) \delta(t - s_1 \cos\theta - s_2 \sin\theta) ds_1 ds_2 \qquad (17)$$

where $\delta()$ is an impulse function. Fig. 6 illustrates the collection of projection data for a single value of 0. The value of $p(\theta, t)$ represents the integral of $x(s_1, s_2)$ along the ray at orientation $8 + \frac{\pi}{2}$, at a displacement t from the center of the field.

In practice, reconstruction requires finite-dimensional representation of both the projection data, p, and the modeled image, x. The projections may be discretized by computing them for only a finite set of M projection rays, $\{(\theta_i, t_i)\}_{i=0}^{M}$. The i^{th} projection is then written as $p_i = p(\theta_i, t_i)$. The Radon transform equations may now be written in the discrete form

$$p = \mathbf{A}x$$

where $\mathbf{A}$ is a sparse matrix whose $(i,j)th$ entry indicates the contribution of modeled pixel j to the i^{th} projection measurement.

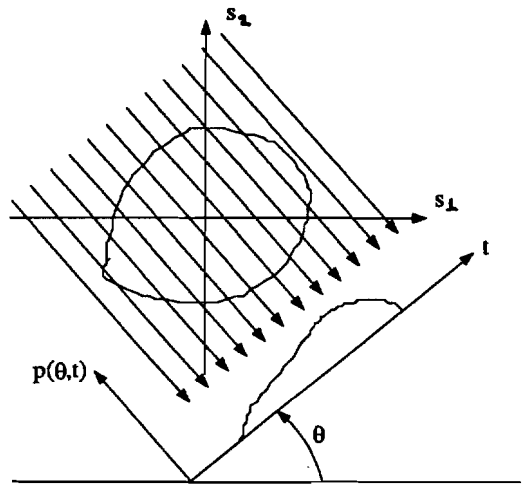


Figure 6: Projection data for angle θ, resulting in the one-dimensional function $p(\theta,t)$.

In transmission tomography the projections, p, are not measured directly. Instead, raw data are in the form of the number of photons, y_i, detected after passing through an absorptive material. We will use a quadratic approximation of the log likelihood of photon counts, y given the image x [36]:

$$L(y|f) \approx -\frac{1}{2}(\hat{p} - \mathbf{A}x)^t \mathbf{D}^2(\hat{p} - \mathrm{Ax}) + c(y), \tag{18}$$

where $\hat{p}_i$ and D are defined by

$$\begin{aligned} \hat{p}_i &= \log(y_T/y_i) \\ \mathbf{D} &= \mathrm{diag}\{\sqrt{y_1}, \sqrt{y_2}, \cdots, \sqrt{y_M}\} \end{aligned}$$

for input photon count y_T. The matrix D more heavily weights errors corresponding to projections with large values of y_i. These projections pass through less dense objects, and consequently have higher signal-to-noise ratio. In the limit of opaque projections where no photons pass through the material, the approximation simply applies no weight to the measurement.

In order to apply the MAP estimation techniques described above, we will require computationally efficient methods for implementing the global and local minimization methods

described in Section 4. In fact, these methods have already been developed in [35, 36] for the quadratic case and discrete-valued priors. This work shows that both the local and global minimization methods require approximately equal amounts of computation per iteration through the data. However, when a Gaussian prior is used, the local method is analytically shown to suppress high frequency error components more rapidly; whereas the global methods suppress low frequencies more rapidly. We believe that this same qualitative behavior will hold for other prior distributions, and that the best strategy is to combine these techniques.

6 Experimental Results

Under the approximation of the conditional log likelihood of the photon counts given in (18), we are restricted to $q = 2$ for the present experimental work, and will show the character of the results' dependence on the choice of p in the GGMRF. The results presented here were achieved using Gauss-Seidel(GS) type iterations[36], with pixel-by-pixel updates, combined with gradient ascent each third iteration. In preliminary trials, convergence of this technique was faster than that of either method independently. As mentioned in Section 4, the GS iterations will in general not find the global minimum for $p = 1$, and will be slow in converging for other small values of p.

The test phantom, shown in Fig. 7, consists of two distinct densities, 0.22cm^{-1} and 0.48cm^{-1}, both of which are within the range of human tissue in X-ray absorptivity. Increasing intensity in Fig. 7 represents higher absorptivity. The physical diameter is approximately 20cm. Projections are collected using only $y_T = 2000$ photons per ray, far below typical clinical dosages, making the lighter regions nearly opaque to the X-rays. With these values for y_T and object composition, photon counting noise may dominate the corruption of the reconstruction if conventional techniques such as convolution backprojection (CBP) are used. The best (by visual ihspection) CBP reconstruction resulted from relatively severe

lowpass filtering of projection data before inversion, and can be seen in Fig. 7. This case is similar to the hollow projections problem, but note that these methods require no estimation of the dense regions' locations, or interpolation of projections. The algorithm can be applied directly to other limited data problems such as the limited-angle reconstruction.

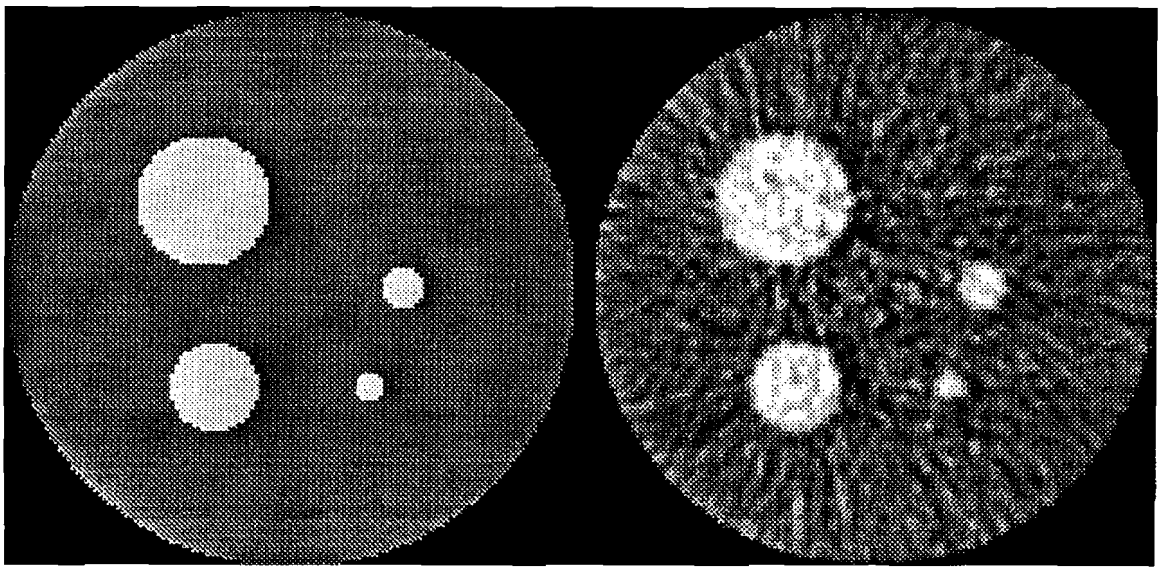
The MRFs used featured only 4-pixel neighborhoods, with equal weighting of nearest horizontal and vertical neighbors. For each p, we have chosen λ yielding the best experimental result. For the Gaussian case, when $p = 2$, we chose $\lambda = 11.2$. This is equivalent to a standard deviation for each pixel, given its neighbors, of 0.03cm^{-1}. For $p = 1.2$, we chose a value of $\lambda = 46.4$, and for $p = 1$, $\lambda = 300$.

The quality of reconstruction for each p did not exhibit great sensitivity to λ, but convergence rates depend heavily on both parameters. Initial stages of convergence proceeded rapidly in each case, but the approximately 450 iterations required for convergence of the estimate when $p = 1$ was over an order of magnitude higher than in the Gaussian case. However, the optimization technique for estimation with the GGMRF ha. not been the focus of this research thus far, and we plan to study improvements for faster convergence. One possibility includes a variable step size for the gradient portion of the iterations.

The reconstruction using the Gaussian prior ($p = 2$) suffers from the smoothing of edges, as illustrated in Fig. 8a. Smaller values of λ can sharpen object boundaries but at the expense of larger noise artifacts.

Fig. 8b and c show the results when a GGMRF prior is used with $p = 1$. Since Fig. 8b only uses the local Gauss-Seidel updates, the image does not converge to the true MAP estimate. The blocky noise artifacts are analogous to the root signals of a median filter. Fig. 8c shows the true MAP estimate for $p = 1$.

The value of $p = 1.2$, as suggested by Rey[27], also produces an improved reconstruction, with intermediate qualities shown in Fig. 8d. Each of these examples featuring non-Gaussian priors exhibits some boundary blockiness, and tendency toward very straight edges, due to

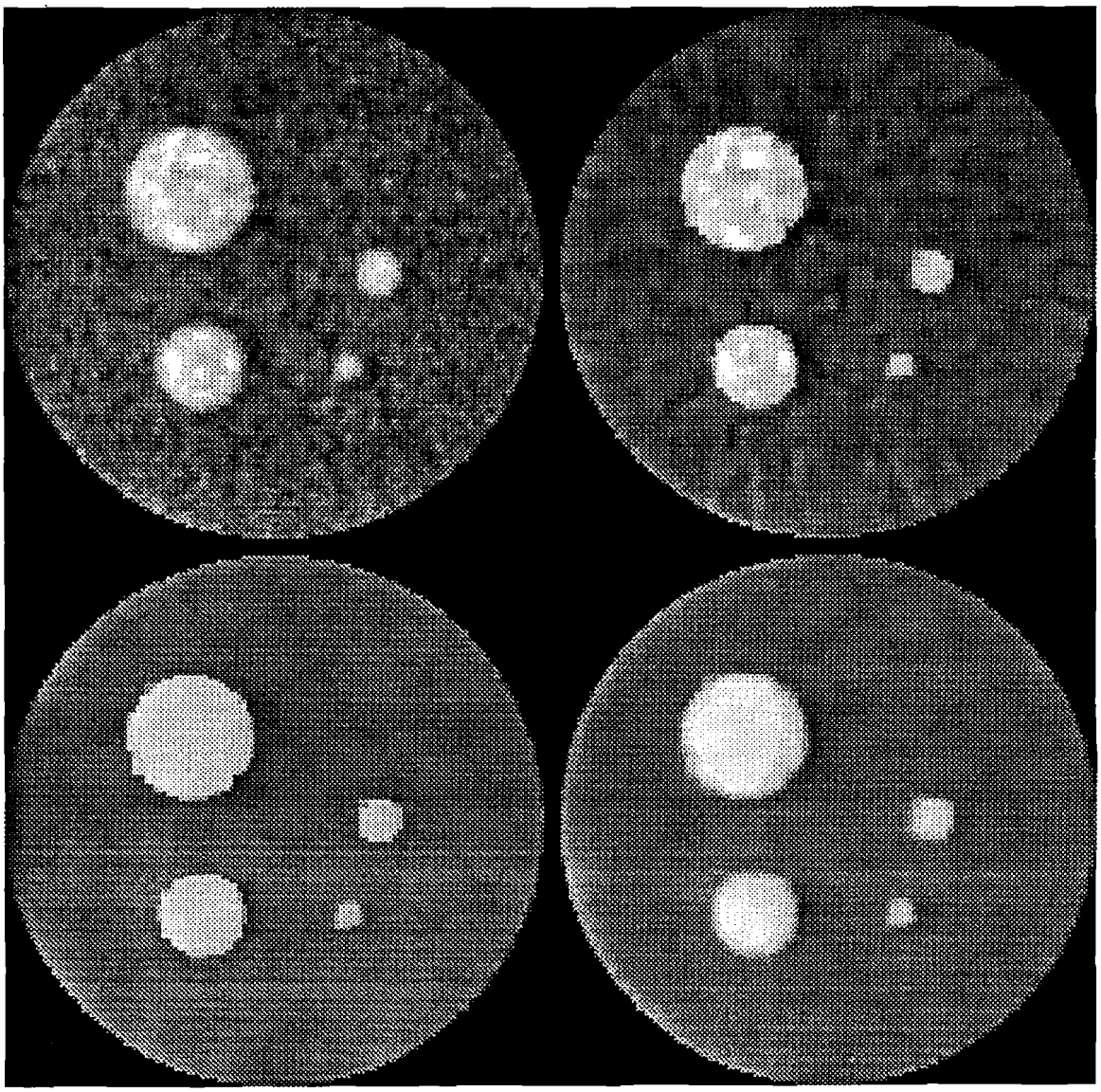


a | b

Figure 7: a) Original phantom (left); b) convolution backprojection reconstruction in low photon dosage with 128 projections at each of 128 angles (right); All images are presented at a resolution of 128 x 128 pixels.

the simple 4-pixel neighborhood. The size of the neighborhood is a minor consideration in computational cost in the tomographic problem[36], and improved results are expected with larger neighborhoods.

The GGMRF MAP estimate with small values of p has substantially lower mean-squared error than the CBP image, or the MAP estimate with the Gaussian prior. But because the mean-squared error tends to be dominated by pixels at the edges of the high intensity regions, we have found it to be a misleading measure of performance. Alternatively, Fig. 9 shows a histogram of the absolute error in the reconstructed images for $p = 1$ and $p = 2$. Note the much greater concentration of errors at the lower magnitudes for the case $p = 1$.



a	b
c	d

Figure 8: a) MAP estimate using Gaussian prior, $p = q = 2$ (upper left); b) Result of using only local minimization with $p = 1$, $q = 2$ (upper right); c) MAP estimate using Generalized Gaussian MRF, $p = 1$, $q = 2$, with alternating gradient ascent and local optimization (lower left); d) MAP estimate for $p = 1.2$ (lower right).

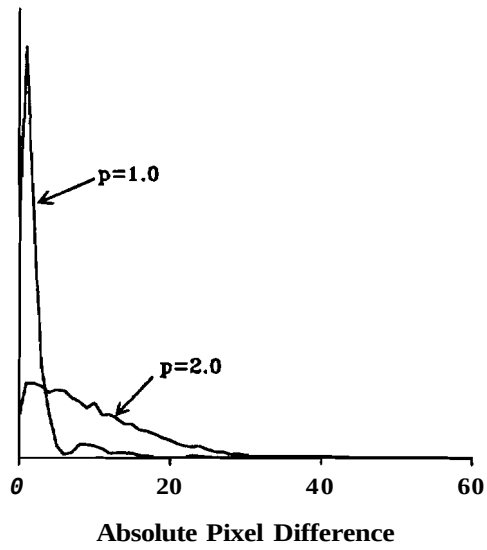


Figure 9: Histograms of absolute error in the reconstructed images for the cases $p = 1.0$ and $p = 2.0$.

7 Conclusion

The GGMRF has demonstrated analytical properties and experimental results which offer promise for applications in many problems of image estimation. In particular, the GGMRF prior leads to a MAP estimator which may be uniquely computed. Moreover, when $1 < p \leq 2$, this MAP estimator is guaranteed to be a continuous function of the input data. For any problem in which the noise parameter q equals the prior parameter p, the MAP estimator will be invariant to scaling of the data. In particular, this means that edge magnitudes need not be predetermined. When $p \neq q$, variations in the data scale are equivalent to variations in the signal-to-noise ratio, $\sigma^q \lambda^p$, used in the MAP estimate.

The simulations presented here, in computed tomography, have dealt with a simple phantom, with only two densities. Such an image is ideal for priors which encourage sharp transitions in estimated reconstructions. The suitability of the GGMRF to more smoothly varying images is, as yet, unclear. However, it is promising that median filters, which have

been successfully applied in images processing, are closely related to MAP estimation with the GGMRF prior. As noted earlier, the Bayesian approach has the advantage of retaining the original data in its recursions.

The very slow convergence of the MAP estimate with small p is an impediment to the efficient application of these techniques. A major effort of our coming research will be directed toward speeding the MAP estimation process.

Acknowledgment

The authors would like to thank Professors R. Stevenson and E. Delp for their many useful ideas and comments.

A Appendix

Proof of theorem 1:

We will prove this theorem for $f(\cdot,\cdot)$ defined on any general metric space, U x V. The appropriate induced metrics on U and V will be denoted by $d_u(x,\tilde{x})$ and $d_v(y,\tilde{y})$ respectively. This is equivalent to $||x-\tilde{x}||$ and $||y-\tilde{y}||$ when U and V are vector spaces.

Choose any $y \in V$. By assumption, there is a unique global minimum.

$$\begin{aligned} \hat{x} &= \arg\min_{x\in U} f(x,y) \\ c &= f(\hat{x},y) \end{aligned}$$

Our objective is then to show that for any $\epsilon > 0$, there is a $6 > 0$, so that for all $\tilde{y}$ with $d_v(y,\tilde{y}) < \delta$

$$\arg\min_{x\in U} f(z,\tilde{y}) \in E$$

where

$$E = \{x : d_u(x,\hat{x}) < \epsilon\} .$$

By assumption, there exists a $b > 0$ such that

$$A_1 = \{x \in U : f(x,y) \leq b + c\}$$

is a compact set. If we define the sequence of sets

$$A_n = \{x \in U : f(x,y) \leq b/n + c\}$$

then each A_n contains $\hat{x}$ and must be compact since it is a closed subset of A_1. If $\bar{\mathcal{E}}$ denotes the closed set given by the complement of E, then

$$O_n = A_n \cap \bar{\mathcal{E}}$$

is a sequence of compact sets none of which contain $\hat{x}$.

We will next show that for some N, $A_N \subset$ E or equivalently O_N is empty. To show this, assume that O_n is not empty for any n. Since O_1 is compact, there is a sequence of points, $x_n \in O_n \subset O_1$, with a subsequence, x_{n_k} that converges in O_1. Since $f(\cdot, y)$ is continuous and $x_n \in A_n$, the limit of this subsequence must also be a global minimum of $f(\cdot, y)$. This contradicts the assumption that there is a unique global minimum to the function $f(\cdot, y)$.

Define the following three subsets of A_N.

$$\begin{aligned} \tilde{A}_N &= \{x \in U : f(x,y) < b/N + c\} \\ \mathbf{I} &= \text{the interior points of } A_N \\ \mathrm{B} &= \text{the boundary points of } A_N \end{aligned}$$

Then by the continuity of $f(\cdot, y)$, it may be shown that $\tilde{A}_N \subset \mathbf{I}$. Therefore,

$$A_N - \tilde{A}_N \supset A_N - I = B \ .$$

This implies that for all x E B, $f(x,y) = b/N + c$.

Since $f(\cdot,\cdot)$ is a continuous function, it is uniformly continuous on any compact set. Therefore, there exists a $\delta > 0$ such that for all $\tilde{y}$, with $d_v(y,\tilde{y}) < \delta$

$$\sup_{x \in A_N} |f(x,y) - f(x,\tilde{y})| < \frac{b}{4N} \ . \tag{19}$$

We will use this fact to show that for any choice of $\tilde{y}$, with $d_v(y,\tilde{y}) < 6$ the global minimum of $f(\cdot,\tilde{y})$ is still a member of $A_N \subset E$. Since A_N is compact and $f(\cdot,\tilde{y})$ is continuous, $f(\cdot,\tilde{y})$ must take on its minimum value at some point, $\tilde{x} \in A_N$. If we can show that the point $\tilde{x}$ is in the interior of A_N, then it must be the global minimum since it is a local minimum and $f(\cdot,\tilde{y})$ is assumed to have no local minima other than the global minimum.

Using the uniform continuity property of 19, we know that for all boundary points $x \in B$

$$f(x,\tilde{y}) \geq \frac{3b}{4N} + c \,, \tag{20}$$

and at the point $\hat{x}$

$$f(\hat{x},\tilde{y}) \leq \frac{b}{4N} + c \tag{21}$$

Therefore, $f(\cdot,\tilde{y})$ is less at the point $\hat{x}$ than at any point on the boundary of A_N. Since there is at least one point in the interior of A_N which is less than any point on the boundary, the minimum of $f(\cdot,\tilde{y})$ must fall on the interior of A_N.

B Appendix

Proof of Theorem 2:

Strict convexity implies that at most one local minimum of f exists. Without loss of generality, assume the minimum occurs at $x = 0$, and the minimum value is $f(0) = 0$.

In order to show that $C = \{x \in \mathbb{R}^N : f(x) \leq b\}$ is compact for any $b > 0$, we invoke the Heine-Borel Theorem, which states that in $\mathbb{R}^N$, every closed and bounded set is compact.

Convexity implies continuity of f. Since the mapping $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is continuous, $f^{-1}(S) \in \mathbb{R}^N$ is closed for every closed set S in $\mathbb{R}$. Therefore, $C = f^{-1}((-\infty, b])$ is a closed set.

By definition of the unique local minimum, there is an N-ball, $B = \{x : ||x|| \leq 1\}$, about 0 such that for $x \neq 0$ with $x \in B$, $f(x) > 0$. Given that f is continuous, the latter inequality

holds on the surface of the ball, $D = \{x : ||x|| = 1\}$, a compact set. The continuous function f must attain a minimum in D, and we denote any point at which the minimum occurs as x_m. If we choose $b = f(x_m) > 0$, then $C \subset B$ and C is compact. To see this assume that $x_o \in C$, but $x_o \notin B$. Then defining $\lambda = \frac{1}{||x_o||}$, we have

$$\begin{aligned} \lambda f(x_o) + (1-\lambda) f(0) &> f(\lambda x_o) \\ &= b \end{aligned}$$

and this implies the contradiction $f(x_o) > b$.

C Appendix

Proof of theorem **3:**

($\Leftarrow$) We must prove that **(a)** $\exp\{-||x||^p\}$ defines a proper density function, **(b)** $||x||^p$ is convex, and **(c)** scalable.

a) Any norm has the property that in a finite dimensional space

$$\int_{R^N} \exp\{-||x||^p\} < \infty .$$

Therefore, this forms a proper probability distribution.

b) We may use the triangle inequality and the convexity of $|.|^p$ to show that $||x||^p$ is convex. For all $0 < \lambda < 1$,

$$\begin{aligned} ||\lambda x + (1-\lambda) y||^p &\leq (\lambda ||x|| + (1-\lambda)||y||)^p \qquad (22) \\ &\leq \lambda ||x||^p + (1-\lambda)||y||^p \qquad (23) \end{aligned}$$

c) X is scalable since

$$\begin{aligned} \log g(\alpha x) &= ||\alpha x||^p + c \\ &= |\alpha|^p \{\log g(x)\} + (1 - |\alpha|^p) c \end{aligned}$$

($\Rightarrow$) We must determine that $-\log g(x) = (f(x))^p$ +constant where f has the properties **(a)** for all c $\geq$ 0, $f(cx) =$ cf (x), **(b)** for all x, f (x) = 0 implies x = θ where 8 is the zero vector, and ***(c)*** f(x) obeys the triangle inequality.

a) Define the function $\tilde{u}(x) = -\log g(x)$. Since $\tilde{u}(x)$ is convex, it is a continuous function of x, and $\tilde{u}(\theta)$ exits where θ is the vector of zeros. By assumption we have that for any a there are β and γ so that

$$\tilde{u}(\alpha x) = \beta\tilde{u}(x) - \gamma \ .$$

If we define the new function, $u(x) = \tilde{u}(x) - \tilde{u}(\theta)$ then for all x

$$u(\alpha x) = \beta u(x)$$

Choose any a > 1. Since $g(x)$ must integrate to 1, there must be an x such that $u(x) = u_o > 0$. Consider the three points $u(0x) = 0$, $u(x) = u_o$ and $u(\alpha x) = \beta u_o$. Convexity implies that $\beta \geq$ a. Therefore, we can find a p $\geq$ 1 so that $\beta = \alpha^p$.

Define f (x) = $(u(x))^{1/p}$. Then for all integers n $\geq$ 1, f $(\alpha^n x) = \alpha^n$f (x). Choose 6 = $a^{1/m}$, then similarly

$$u(\alpha x) = u(\delta^m x) = \beta_\delta^m u(x)$$

where β_δ is chosen so that $u(\delta x) = \beta_\delta u(x)$. From these relationships, we may infer that

$$\begin{aligned} \beta_\delta^m &= \beta = \alpha^p \\ \beta_\delta &= \alpha^{p/m} \\ \text{f}(\delta^n x) &= \delta^n f(x) \\ f(\alpha^{1/m} x) &= \alpha^{1/m}\text{f (x)} \\ f(\alpha^{n/m} x) &= \alpha^{n/m} f(x) \ . \end{aligned}$$

Since m and n are arbitrary integers and f is continuous, we have that for all c $\geq$ 1, $f(cx) = cf(x)$. Let 0 < c < 1, then $(1/c)f(cx) = f(x)$, and therefore $f(cx) = cf(x)$. Therefore, we have that for all c $\geq$ 0, f (cx) = cf (x).

b) Assume that there exists $f(x_o) = 0$ but $x_o \neq 8$. For all y,

$$\begin{aligned} u(x_o + y) &\leq (1/2)u(2x_o) + (1/2)u(2y) \\ &= (1/2)u(2y) \\ &= 2^{p-1}u(y) \end{aligned}$$

where the first inequality is by convexity, and the second by the fact that $f(2x_o) = 0$. By continuity of f, we may define the set

$$A_\epsilon = \{y : y^t x_o = 0,\ ||y|| < \epsilon\}$$

such that for all y $\in A_\epsilon$, $u(y) < 1$. Using the above inequality, we have that

$$\begin{aligned} \int_R g(x)dx &\geq \int_{-\infty}^{\infty}\int_{A_\epsilon} \exp\{-u(\alpha x_o/||x_o||_2 + y) - u(\theta)\}\, dyd\alpha \\ &\geq \int_{-\infty}^{\infty}\int_{A_\epsilon} \exp\left\{-2^{p-1} - u(\theta)\right\} dyd\alpha \\ &= \infty . \end{aligned}$$

c) First choose any x and y so that $f(x) = f(y) = c \neq 0$. Then for any $0 < \lambda < 1$,

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= cf\left(\frac{\lambda}{c}x + \frac{(1-\lambda)}{c}y\right) \\ &= c\left(u\left(\frac{\lambda}{c}x + \frac{(1-\lambda)}{c}y\right)\right)^p \\ &= c(\lambda u(x/c) + (1-\lambda)u(y/c))^p \\ &= c\left(\frac{\lambda}{c}u(x) + \frac{(1-\lambda)}{c}u(y)\right)^p \\ &= c \\ &= \lambda f(x) + (1-\lambda)f(y) . \end{aligned}$$

Now choose any $x, y \neq 8$, then $f(x), f(y) \neq 0$. Define,

$$\lambda = \frac{f(x)}{f(x) + f(y)}$$

and also define $x' = x/\lambda$ and $y' = y/(1-\lambda)$. Then since $f(x) = f(y')$, we may apply the above result to yield the triangle inequality for $f(\cdot)$.

$$\begin{aligned} f(x+y) &= f(\lambda x' + (1-\lambda)y') \\ &\geq \lambda f(x') + (1-\lambda) f(y') \\ &= f(x) + f(y) \end{aligned}$$

References

[1] H. Derin, H. Elliot, R. Cristi, and D. Geman, "Bayes Smoothing Algorithms for Segmentation of Binary Images Modeled by Markov Random Fields," ***IEEE Tmns. Trans. Pattern Anal. and Mach. Intell.***, vol. PAMI-6, no.6 , pp. 707-720, Nov. 1984.

[2] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," ***IEEE Tmns. Pattern Anal. and Mach. Intell.***, vol. PAMI-6, no.6, pp. 721-741, Nov. 1984.

[3] J. Besag, "On the Statistical Analysis of Dirty Pictures," **J. *Roy. Statist. Soc. B,*** vol. 48, no. 3, pp. 259-302, 1986.

[4] H. Derin and H. Elliott, "Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields," ***IEEE Trans. Pat. An. Mach. Intell.,*** vol. PAMI-9, pp. 39-55, Jan. 1987.

[5] A. Blake and A. Zisserman, ***Visual Reconstruction,*** MIT Press, Cambridge, Massachusetts, 1987.

[6] C. Bouman and B. Liu, "A Multiple Resolution Approach to Regularization," ***Proc. SPIE Conf. on Visual Comm. and Image Proc.,*** pp. 512-520, Cambridge, MA, Nov. 9-11, 1988.

[7] S. Geman and D. McClure, "Bayesian Image Analysis: An Application to Single Photon Emission Tomography," in ***Proc. Statist. Comput. Sect. Amer. Stat. Assoc.,*** Washington, DC, pp. 12-18, 1985.

[8] S. Geman and D. McClure, "Statistical Methods for Tomographic Image Reconstruction," ***Bull. Int. Stat. Inst.***, vol. LIZ-4, pp. 5-21, 1987.

[9] T. Hebert and R. Leahy, "A Generalized EM Algorithm for 3-D Bayesian Reconstruction from Poisson data Using Gibbs Priors," ***IEEE Trans. Med. Im.,*** vol. 8, no. 2, pp. 194-202, June 1989.

[10] R. Stevenson and E. Delp, "Fitting Curves with Discontinuities," ***Proc. of the First International Workshop on Robust Computer Vision,*** pp. 127-136, Seattle, WA, Oct. 1-3, 1990.

[11] P. Huber, ***Robust Statistics,*** John Wiley & Sons, New York, NY, 1981.

[12] P. J. Green "Bayesian Reconstructions from Emission Tomography Data Using a Modified EM Algorithm,: ***IEEE Tmns. Med. Im.,*** vol. 9, no. 1, pp. 84-93, March 1990.

[13] K. Lange, "Convergence of EM Image Reconstruction Algorithms with Gibbs Priors," ***IEEE Tmns. Med. Im.,*** vol. 9, no. 4, pp. 439-446, Dec. 1990.

[14] S. A. Kassam, ***Signal Detection in Non-Gaussian Noise,*** Springer-Verlag, New York, 1988.

[15] G. T. Herman, H. Hurwitz, A. Lent, and H-P. Lung, "On the Bayesian Approach to Image Reconstruction," ***Info. and Cont.,*** vol. 42, pp. 60-71, 1979.

[16] K. M. Hanson and G. W. Wecksung, "Bayesian Approach to Limited-Angle Reconstruction in Computed Tomography," ***J. Opt. Soc. Am.,*** vol. 73, no. 11, pp. 1501-1509, Nov. 1983.

[17] E. Levitan and G. T. Herman, "A Maximum ***A Posteriori*** Probability Expectation Maximization Algorithm for Image Reconstruction in Emission Tomography," ***IEEE Tmns. Med. Imag.,*** vol. MI-6, No. 3, pp. 185-192, 1987.

[18] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic Solution of Ill-Posed Problems in Computational Vision," ***J. of the Am. Stat. Assoc.*** vol. 82, pp 76-89, March 1987.

[19] B. Horn and B. Schunck, "Determining Optical Flow," ***Artificial Intelligence,*** vol. 17, pp. 185-203, 1981.

[20] K. Ikeuchi and B. Horn, "Numerical Shape from Shading and Occluding Boundaries," ***Artificial Intelligence,*** vol. 17, pp. 141-183, 1981.

[21] V. Torre and T. Poggio, "On Edge Detection," ***IEEE Tmns. Pat. An. Mach. Intell.,*** vol. PAMI-8, no. 2, pp. 147-163, March 1986.

[22] H. Van Trees, ***Detection, Estimation, and Modulation Theory,*** John Wiley & Sons, New York, 1968.

[23] C. Bouman and B. Liu, "Multiple Resolution Segmentation of Textured Images," ***IEEE Trans. on Pat. An. Mach. Intell.***, vol. 13, no. 2, pp. 99-113, Feb. 1990.

[24] J. Besag, "Spatial Interaction and the Statistical Analysis of Lattice Systems," ***J. Royal Stat. Soc. B,*** vol. 36, pp. 192-326, 1974.

[25] R. Kindermann and J. L. Snell, ***Markov Random Fields and their Applications.*** Providence: American Mathematical Society, 1980.

[26] J. Hutchinson, C. Koch, J. Luo and C. Mead, "Computing Motion Using Analog and Binary Resistive Networks," ***Computer,*** vol. 21, pp. 53-63, March 1988.

[27] W. Rey, ***Introduction to Robust and Quasi-Robust Statistical Methods,*** Springer-Verlag, Berlin, 1980.

[28] A. Tikhonov and V. Arsenin, ***Solutions of Ill-Posed Problems,*** Winston and Sons, New York, 1977.

[29] N. Gallagher and G. Wise, "A Theoretical Analysis of the Properties of Median Filters," ***IEEE Trans. Acoust., Speech, Signal Processing,*** vol. ASSP-29, no. 6, pp. 1136-1141, Dec. 1981.

[30] T. Nodes and N. Gallagher, "Median Filters: Some Modifications and Their Properties," ***IEEE Tmns. Acoust., Speech, Signal Processing,*** vol. ASSP-30, no. 5, pp. 739-746, Oct. 1982.

[31] J. P. Fitch, E. Coyle and N. Gallagher, "Root Properties and Convergence Rates of Median Filters," ***IEEE Tmns. Acoust., Speech, Signal Processing,*** vol. ASSP-33, no. 1, pp. 230-239, Feb. 1985.

[32] F. Beckman, "The Solution of Linear Equations by the Conjugate Gradient Method," in eds. A. Ralston, H. Wilf and K. Enslein, ***Mathematical Methods for Digital Computers,*** Wiley, 1960.

[33] O. Yli-Harja, J. Astola and Y. Neuvo, "Analysis of the Properties of Median and Weighted Median Filters Using Threshold Logic and Stack Filter Representation," ***IEEE Tmns. Signal Processing,*** vol. 39, no. 2, Feb. 1991.

[34] J. Astola and Y. Neuvo, "Matched Median Filtering," to appear in ***IEEE Trans. on Communications.***

[35] K. Sauer and C. Bouman, "Bayesian Estimation from Projections with Low Photon Dosages," to appear in the ***Proc. of IEEE Int'l Conf. on Acoust., Speech and Sig. Proc.,*** 1991.

[36] K. Sauer and C. Bouman, "A Local Update Strategy for Iterative Reconstruction from Projections," submitted to the ***IEEE Tmns. Sig. Proc..***