9-29-2011

Implicit Priors for Model-Based Inversion

Eri Haneda
School of Electrical and Computer Engineering, Purdue University, haneda@purdue.edu

Charles A. Bouman
Purdue University, bouman@purdue.edu

Implicit Priors for Model-Based Inversion

Eri Haneda

Charles A. Bouman

TR-ECE-11-18

September 29, 2011

School of Electrical and Computer Engineering

1285 Electrical Engineering Building

Purdue University

West Lafayette, IN 47907-1285

IMPLICIT PRIORS FOR MODEL-BASED INVERSION

Eri Haneda and Charles A. Bouman

Purdue University
School of Electrical and Computer Engineering
West Lafayette, IN 47907

ABSTRACT

While MRF models have been widely used in the solution of inverse problems, a major disadvantage of these models is the difficulty of parameter estimation. At its root, this parameter estimation problem stems from the inability to explicitly express the joint distribution of an MRF in terms of the conditional distributions of elements given their neighbors. The objective of this paper is to provide a general approach to solving maximum a posteriori (MAP) inverse problems through the implicit specification of a MRF prior. In this method, the MRF prior is implemented through a series of quadratic surrogate function approximations to the MRF's log prior distribution. The advantage of this approach is that these surrogate functions can be explicitly computed from the conditional probabilities of the MRF, while the explicit Gibbs distribution can not. Therefore, the Gibbs distribution remains only implicitly defined. In practice, this approach allows for more accurate modeling of data through the direct estimation of the MRF's conditional probabilities. We illustrate the application of our method with a simple experiments of image denoising and show that it produces superior results to some widely used MRF prior models.

Index Terms— Markov random fields, Inverse problems, Maximum a posteriori estimation.

1. INTRODUCTION

Model-based inversion methods, first introduced decades ago for the solution of ill-posed inverse problems [1] have continued to gain importance as their value in the solution of difficult and widely used inverse problems grows [2, 3, 4]. A classical approach to model-based inversion is the computation of the maximum a posteriori (MAP) estimate which is given by

$$\hat{x} = \arg\max_{x \geq 0} \{\log p(y|x) + \log p(x)\} \ ,$$

where $p(y|x)$ is the forward model of the data vector y given the unknown vector x, and $p(x)$ is the prior model for x. In many important applications, x is an image or 3D volume and $p(x)$ is a prior model for the image.

An alternative to the model-based inversion approach is direct inversion methods which attempt to directly model the relationship between the data, y, and the unknowns, x. So for example, methods such as bilateral filters, kernel regression [5], non-local means [6], BM3D [7], conditional Markov random fields [8], and scale mixtures [9] tend to focus on the direct estimation of x from the available data. In other denoising approaches, such as dictionary [10] and subspace-based methods, the image is modeled as belonging to a low-dimensional manifold, and the estimate x is recovered by constraining y to be from the appropriate estimated subspace.

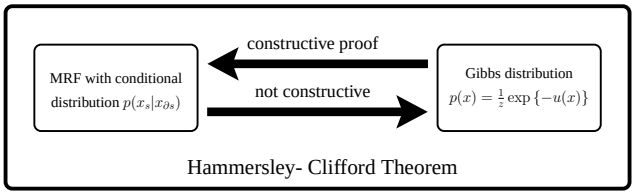


Fig. 1. The Hammersley-Clifford theorem states that X is an MRF with a strictly positive density if and only if it has a Gibbs distribution. However, the value of this very important theorem is limited by the fact that there is generally no tractable method to construct the Gibbs distribution from the conditional distributions.

While direct inversion methods can produce amazing results, model-based inversion approaches have continued to be enormously valuable in applications such as 3D reconstruction from X-ray computed tomography (CT) data [11]. The great advantage of model-based inversion over alternative methods is that it allows for the explicit incorporation of a forward model. In applications such as X-ray CT, microscopy, or astronomy, this is very important because it is often possible to quantify the very complex forward model for the physical measurements very precisely. And this precise forward model can greatly improve the quality of inversion.

However, a weakness of the model-based inverse approach is the fact that one must adopt an explicit, tractable, and accurate prior model $p(x)$. Perhaps the most common choice of prior model has been the Markov random field (MRF) because it limits the dependencies in x so that

$$p_\theta(x_s|x_r \ r \neq s) = p_\theta(x_s|x_{\partial s}) \ ,$$

where x_s is an element of x, ∂s is the index set for the neighbors of s, and θ is a parameter vector which can be used to fit the prior distribution for specific application at hand.

Unfortunately, the conditional distribution $p_\theta(x_s|x_{\partial s})$ does not provide an explicit form for the prior distribution $p_\theta(x)$ of the associated MRF.[1] The partial solution to this dilemma comes from the celebrated Hammersley-Clifford Theorem [12] which states that X is an MRF[2] if and only if its distribution can be expressed as a Gibbs

This research was supported by the U. S. Army Research Office and the Army Research Laboratory under contract number W911NF-09-1-0540, and ALERT DHS center Northeastern University.

[1] This is in contrast to a Markov chain in which the conditional probabilities can be simply multiplied to form the prior distribution.

[2] Technically, it must also have a strictly positive density, but in practice some values of X can occur with extremely low probability.

distribution

$$p(x) = \frac{1}{z_\theta} \exp\{-u_\theta(x)\}$$

where $u_\theta(x)$ is the Gibbs energy function which is formed by a sum of potential functions over neighborhood cliques, and z_θ is the so-called partition function. However, Figure 1 illustrates the limitation of the Hammersley-Clifford theorem. While the proof that a Gibbs distribution is an MRF is constructive, the proof that the MRF must have a Gibbs distribution with potential functions limited to cliques is not. Therefore, the Hammersley-Clifford Theorem does not provide a general trackable method for constructing the Gibbs distribution from the known conditional distribution, $p_\theta(x_s|x_{\partial r})$. It simply says that such a mapping exists.

Moreover, a major disadvantage of the Gibbs distribution is that estimation of the parameters of a Gibbs distribution tends to be difficult due to the[3] intractable nature of the partition function z_θ. In practice, this means that model-based inversion methods typically use very simplistic prior models with a very small number of parameters that can be estimated. This severely limits the expressiveness of prior models, which consequently limits the accuracy of model-based inversion.

The objective of this paper is to introduce a novel method for the construction of prior models which is based on direct estimation of the conditional distribution, $p_\theta(x_s|x_{\partial s})$, of a MRF model. The advantage of this approach is that it offers the opportunity to construct much more informative and accurate prior models through the use of a wide array of modern methods for the accurate estimation of conditional densities from training data. This is in contrast to traditional MRF approaches which typically assume a Gibbs distribution with simple potential functions controlled by a small number of parameters.

In this paper, we introduce a novel approach to computing the MAP estimate which does not require the Gibbs distribution to be explicitly computed. Instead, we compute local successive approximations to the energy function $u_\theta(x)$ using a surrogate energy function $u(x; x')$ where x' is the point of approximation. As it turns out, this surrogate energy function can be explicitly computed from the MRFs conditional probabilities, whereas the underlying Gibbs distribution remains implicit. The surrogate energy function works in the same way as widely used majorization methods for optimization [13, 14, 15, 16, 17]. The key novelty to our method is in how we compute the surrogate energy function's form from the known (e.g. typically estimated from training data) conditional probabilities, $p_\theta(x_s|x_{\partial s})$. Using this method, we can then compute the MAP inversion with a nested iteration: Each "outer loop" updates the point of approximation x', and each "inner loop" maximizes a surrogate MAP optimization problem with a quadratic prior term. Moreover, we show that, intuitively, the implicit prior can be thought of as a non-homogeneous Gaussian MRF prior in which the neighborhood weights of the GMRF are adapted to the local structure of the image, x.

2. BAYESIAN INVERSION USING IMPLICIT PRIOR

Our approach to Bayesian inversion with an implicit prior is illustrated in Figure 2. First, a the conditional distribution, $p_s(x_s|x_{\partial s})$, is either estimated from training data or simply selected based on knowledge of the application. In our example, we assume that the conditional density is shift invariant (i.e. the MRF is homogeneous and $p_s(x_s|x_{\partial s}) = p(x_s|x_{\partial s})$), so that the conditional density can be easily estimated from training data. However, the method is fully applicable to the non-homogeneous case. Any number of methods can be used to model this conditional distribution, but we will adopt a mixture distribution in this work.[4]

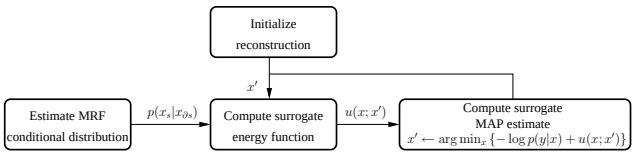


Fig. 2. Flow diagram illustrating the operations in MAP estimation with an implicit prior.

MAP estimation with implicit prior

1. Initialize x'
2. Repeat until x' has converged { //Outer Loop
   (a) Update surrogate energy function $u(x; x')$
   (b) $x' \leftarrow \arg\min_x \left\{-\log p(y|x) + u(x; x')\right\}$ //Inner Loop
   }

The challenge is then to use this conditional density function as the basis for a prior model in MAP inversion. To do this, we will use the known conditional density to compute a surrogate energy function, $u(x, x')$, with the properties that

$$u(x') = u(x'; x') \quad (1)$$

$$u(x) \leq u(x; x')\,, \quad (2)$$

where $u(x)$ is the energy function of the MRF's unknown Gibbs distribution. Once we obtain this surrogate energy function, we may use it to iteratively minimize the MAP cost function with the following procedure. Based on the theory of majorization, it is well known that the two properties of equations (1) and (2) insure monotone convergence of the MAP cost function with each iteration of the algorithm; and if the functions are C^2, then any fixed point of the algorithm will be a point at which the gradient of the log posterior distribution is zero.

However, the problem remains of how to determine the surrogate energy function, $u(x; x')$, from the known conditional densities, $p_s(x_s|x_{\partial s})$. Since we do not have an explicit form for the energy function, $u(x)$, we will need to determine the surrogate energy function's form implicitly. To do this, we first adopt a surrogate energy function with the form

$$u(x; x') = \frac{1}{2}(x - x')^t B(x - x') + d^t(x - x') + c\,, \quad (3)$$

where B is a symmetric matrix, d a column vector, and c a scalar, all of which are assumed to be functions of x'. We note that the

[3]The partition function does have a tractable form, for example, when the prior distribution is Gaussian or when X is a binary Ising MRF of infinite extent, but we are primarily interested in more general cases.

[4]It should be noted that a particular conditional density, $p_s(x_s|x_{\partial s})$, may not (typically does not) correspond to a consistent global density, $p(x)$. However, for the purposes of our analysis we will assume that such a consistent global density does exist. In practice, some corrections must be made experimentally to account for this fact, which are described latter.

quadratic form of the function makes solution of the surrogate MAP optimization problem straight forward using a wide range of standard optimization methods.

Without loss of generality, we can simply assume that $c = 0$ because the value of c does not affect the result of optimization. Alternatively, we can always assume that $u(x') = 0$ through appropriate renormalization of the partition function.

The values of the parameter vector d can be easily computed from the gradient of the conditional densities as

$$d_s = -\left.\frac{\partial}{\partial x_s}\log p_s(x_s|x_{\partial s})\right|_{x=x'} . \quad (4)$$

Appendix A provides a derivation of this result, but intuitively, the relationship is required by the fact that the functions $u(x;x')$ and $u(x)$ must be tangent at the point $x = x'$.

It only remains to choose the symmetric matrix B sufficiently large so that $u(x;x')$ upper bounds the true energy function. However, if B is chosen to be too large, then convergence of the algorithm will be slow; so it is best to select a B which represents as tight an upper bound as possible. Therefore, our approach is to first find strong necessary conditions that B must satisfy, and then present a method to compute a matrix B that satisfies these conditions for our specific choice of the conditional distribution. Once this is done, we can then scale the magnitude of B or its diagonal as is necessary to insure an upper bound.

The following three conditions must hold for any matrix B which satisfies the equations of (1) and (2). (See proofs in Appendix B.)

Condition 1: The symmetric matrix B must be positive definite.

Condition 2: It must be the case that $B \geq H$, i.e. $B-H$ must be a positive semi-definite matrix, where H is the Hessian of the energy function $u(x)$ at $x = x'$. Moreover, the elements of H are given by

$$H_{s,r} = -\left.\frac{\partial^2}{\partial x_s \partial x_r}\log p_s(x_s|x_{\partial s})\right|_{x=x'} . \quad (5)$$

Condition 3: It must be the case that $B_{s,s} \geq D_{s,s}$ where D is a diagonal matrix with entries

$$D_{s,s} = 2\sup_{x_s \neq x'_s}\left\{\frac{-\log p_s(x_s|x'_{\partial s}) + \log p_s(x'_s|x'_{\partial s}) - \Delta_s d_s}{\Delta_s^2}\right\} \quad (6)$$

where $\Delta_s = x_s - x'_s$ and d_s is from equation (4). Furthermore, it is the case that $D_{s,s} \geq H_{s,s}$.

For our particular example, we will use a homogeneous conditional distribution with the form of a mixture distribution

$$p(x_s|x_{\partial s}) = \sum_k \frac{\gamma_k}{\sqrt{2\pi}\sigma_k}\exp\left\{-\frac{1}{2\sigma_k^2}(x_s - A_k x_{\partial s} - \beta_k)^2\right\},$$

where A_k is a row vector, β_k and σ_k are constants, and $\gamma_k = p(k|x_{\partial s})$. So by computing the partial derivative with respect to x_s we obtain

$$-\frac{\partial}{\partial x_s}\log p(x_s|x_{\partial s}) = \sum_k \frac{1}{\sigma_k^2}(x_s - A_k x_{\partial s} - \beta_k)p(k|x_s, x_{\partial s})$$

where

$$p(k|x_s, x_{\partial s}) = \frac{\frac{\gamma_k}{\sqrt{2\pi}\sigma_k}\exp\left\{-\frac{1}{2\sigma_k^2}(x_s - A_k x_{\partial s} - \beta_k)^2\right\}}{\sum_j \frac{\gamma_j}{\sqrt{2\pi}\sigma_j}\exp\left\{-\frac{1}{2\sigma_j^2}(x_s - A_j x_{\partial s} - \beta_j)^2\right\}} .$$

From this we can compute,

$$d_s = \sum_k \frac{1}{\sigma_k^2}(x'_s - A_k x'_{\partial s} - \beta_k)p(k|x'_s, x'_{\partial s}) \quad (7)$$

$$H_{s,r} \cong \tilde{H}_{s,r} = \sum_k \frac{1}{\sigma_k^2}(\delta_{s=r} - \delta_{s\neq r}A_{k,r})p(k|x'_s, x'_{\partial s}) , \quad (8)$$

where the Hession is efficiently computed using the approximation that $\frac{\partial}{\partial x_r}p(k|x_s, x_{\partial s})$ is small. We also note that it may be the case that the computed value of $\tilde{H}$ is not symmetric, so we impose symmetry by computing $\tilde{H} \leftarrow (\tilde{H} + \tilde{H}^t)/2$.[5]

Using the 1D case of the lemma below (proved in Appendix C), we can see that $\tilde{H}_{s,s} \geq D_{s,s}$; so Condition 3 is met. It is important that B be positive definite (i.e. that Condition 1 holds) in order to insure that the inner loop is convergent. If this is not the case, then we can enforce Condition 1 by selecting B so that

$$B = \tilde{H} + \alpha\,\mathrm{diag}\{\tilde{H}\} , \quad (9)$$

where $\mathrm{diag}\{\tilde{H}\}$ is the positive-definite matrix formed by the diagonal of $\tilde{H}$, and $\alpha \geq 0$ is a positive constant. Notice that from the form of (9) and with the approximation of (8), then Condition 2 is met.

Lemma: Surrogate functions for exponential mixtures
Let $f : \mathbb{R}^N \rightarrow \mathbb{R}$ be a function which takes the form,

$$f(x) = \sum_k w_k \exp\{-u_k(x)\} \quad (10)$$

where $w_k \in \mathbb{R}^+$, $\sum_k w_k > 0$, and $u_k : \mathbb{R}^N \rightarrow \mathbb{R}$. Furthermore $\forall (x, x') \in \mathbb{R}^N \times \mathbb{R}^N$ define the function

$$Q(x;x') \triangleq -\log f(x') + \sum_k \tilde{\pi}_k(u_k(x) - u_k(x'))$$

where $\tilde{\pi}_k = \frac{w_k \exp\{-u_k(x')\}}{\sum_j w_j \exp\{-u_j(x')\}}$. Then $Q(x;x')$ is a surrogate function for $-\log f(x)$, and $\forall (x, x') \in \mathbb{R}^N \times \mathbb{R}^N$,

$$\begin{aligned} Q(x';x') &= -\log f(x') \\ Q(x;x') &\geq -\log f(x) \end{aligned}$$

3. ESTIMATION OF MRF CONDITIONAL PROBABILITY

The advantage of the implicit prior approach is the ability to build a prior model based on accurate estimates of the MRF conditional probability, $p(x_s|x_{\partial s})$. This can be done in many different ways, and the machine learning field includes many examples of techniques for density estimation [18]. In the paper, we will use a conditional probability model that is based on the use of Gaussian mixture distributions. This model has been successfully used over many years in applications such as image interpolation [19], noise reduction and medical image enhancement [20].

The model uses a conditional distribution with the form

$$p(x_s|x_{\partial s}) = \sum_{k=1}^{M} p(x_s|x_{\partial s}, k)p_{k|z}(k|z) \quad (11)$$

[5] Theoretically, the matrix H should be symmetric. However, in practice the estimated conditional distribution may not correspond to a consistent MRF model, so it may be the case that for the Hession computed using equation (5), $H_{s,r} \neq H_{r,s}$.

where $z = f(x_{\partial s})$ is a feature vector extracted from the neighborhood $x_{\partial s}$, and

$$p(x_s|x_{\partial s}, k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{1}{2\sigma_k^2}\left\{x_s - (A_k x_{\partial s} + \beta_k)\right\}^2\right\} \tag{12}$$

and

$$p_{k|z}(k|z) = \frac{\exp\left(-\frac{1}{2\sigma_c^2}||z - \bar{z}_k||^2\right)\pi_k}{\sum_{l=1}^{M} \exp\left(-\frac{1}{2\sigma_c^2}||z - \bar{z}_l||^2\right)\pi_l} \ . \tag{13}$$

The feature vector, z, is computed by first forming an 8-dimensional column vector, z', from the 3×3 window of neighboring pixels to x_s, and subtracting the mean value from the vector. The vector z' is then rescaled using a factor $p = 0.50$.

$$z = \begin{cases} z'\|z'\|^{p-1} & \text{if } z' \neq 0 \\ 0 & \text{else} \end{cases} \tag{14}$$

Notice that the full set of parameters for the model is then given by $\theta = \left(\{\sigma_k^2, A_k, \beta_k, \bar{z}_k, \pi_k\}_{k=1}^M, \sigma_c^2\right)$.

Intuitively, the conditional probability models each pixel as a Gaussian mixture with M components. For each component, the conditional distribution of x_s given $x_{\partial s}$ is assumed to be Gaussian with fixed variance σ_k^2 and a mean which is a linear function of the neighboring pixels given by $(A_k x_{\partial s} + \beta_k)$. The feature vector, z, is modeled as a Gaussian mixture with fixed covariance $\sigma_c^2 I$ and component means given by $\bar{z}_k$. The references [19, 20] describe the details of the models construction and derive the EM algorithm for efficient parameter estimation of θ.

In order to control the level of regularization, we added a tunable parameter, λ, to the conditional distribution model so that

$$p(x) = \frac{1}{z} \exp\left\{-\frac{1}{\lambda^2} u(x)\right\} \ , \tag{15}$$

which is simply implemented by scaling the implicit prior by $1/\lambda^2$ at each step.

4. EXPERIMENTAL RESULTS

(a) Ground truth image x (b) Noisy image y

Fig. 3. (a) Ground truth image, (b) image with additive white noise standard deviation $\sigma_w = 20$

In order to better understand the implicit prior method, we performed simulations for the very simple inverse problem of removing additive white Gaussian noise from an image. More specifically, we generated a noisy observed image, y, from the "ground truth" image x by

$$y = x + w \ ,$$

where w is i.i.d. Gaussian noise with distribution $N(0, \sigma_w^2)$ with $\sigma_w = 20$. Figure 3 shows the ground truth image x, noisy image y. The 25 grayscale training images were taken from a set of natural scene photos, and the ground truth image used in testing was not contained in the set of training images. The images were captured by a Nikon D90 camera, the RGB values were converted to the luma values, and then the images were filtered and subsampled down to approximately 363×288 resolution, so as to be most suitable for illustrating results in this publication. Parameters of the MRF conditional probability density, $p(x_s|x_{\partial s})$, were estimated as described in Section 3 using $M = 32$ and $p = 0.50$.

In addition to using the implicit prior, we also ran comparisons with a range of different parameter values for the generalized Gaussian MRF (GGMRF) [21], and with the more general qGGMRF [11] which represents the current state-of-the-art in MRF priors for inverse problems such as tomography. The energy function of the form is

$$u(x) = \frac{1}{P\sigma^P} \sum_{\{i,j\}} b_{i-j} \rho(\Delta) \tag{16}$$

$$\rho(\Delta) = \frac{|\Delta|^P}{1 + |\frac{\Delta}{C}|^{P-Q}} \tag{17}$$

where $\Delta = x_i - x_j$, and the parameter constraints are $1 \leq Q \leq P \leq 2$ and C is a positive threshold. The sum is over all pairs $\{i, j\}$ such that i and j are 8-point neighbors and the coefficients b_i sum to 1, and the ratio of $b_i/b_j = \sqrt{2}$ when i is a 4-point neighbor and j is an 8-point neighbor. We used $P = 2$ and the two values of $Q = 1.0$ and $Q = 1.2$ to illustrate a typical non-Gaussian MRF prior. If we set $P = Q$, then this is the form of the generalized Gaussian MRF (GGMRF). In all cases, we adjusted the scale parameter, σ, to minimize the root mean squared error (RMSE) between the reconstruction and the ground-truth image. For constancy, we also adjusted the value of λ from equation (15) for the implicit prior in order to achieve minimum RMSE.

Figure 4 shows the root mean squared error (RMSE) of the implicit prior method as a function of the number of iterations. The blue line shows the RMSE between the restored image and the ground-truth image; and the black lines shows the RMSE between the restored image and the converged result of the algorithm. The plot indicates that the MAP estimate converges after about 20 iterations. In practice, smaller values of α tend to result in faster convergence; however, if α is chosen to be too small then the convergence may not be robust. In this case, $\alpha = 0.45$ was larger than necessary, but we found this value consistently produced robust convergence in a wide array of examples.

Figure 5 attempts to graphically illustrate the values in the matrix $B_{s,r}$ after 20 iterations. For a pixel s, the color was set to green $= 255 * B_{s,s+(0,-1)}/B_{s,s}$ and red = blue $= 255 * B_{s,s+(-1,0)}/B_{s,s}$, where $r = s + (0, -1)$ is the pixel immediately to the left of s, and $r = s + (-1, 0)$ is the pixel immediately above s. This image shows how the local weights in the surrogate energy function adapt to the local edge structure in the image.

Figure 6 shows the comparison of MAP image reconstructions using the implicit prior and the qGGMRF prior with ($P = 2.0$, $Q = 1.2$) in Figure 6(c), and ($P = 2.0$, $Q = 1.0$)(d) in Figure 6(d). In each case, the threshold C and the regularization σ were chosen to achieve the minimum RMSE. The implicit prior result is slightly sharper with slightly better detail than both the qGGMRF cases. This conclusion is supported by the objective measures of RMSE presented in Figure 7 where the implicit prior techinique has the smallest value among the three techniques. The RMSE of the qGGMRF prior technique is plotted as a function of the treshold C

(a) Ground truth

(b) Implicit prior

(c) qGGMRF ($P = 2$, $Q = 1.2$, $C = 0$)

(d) qGGMRF ($P = 2$, $Q = 1.0$, $C = 1.5$)

Fig. 6. The image restoration comparison between the implicit prior technique and qGGMRF prior technique ($P = 2$, $Q = 1.2$, $C = 0$) and ($P = 2$, $Q = 1.0$, $C = 1.5$). In each case, the threshold C and regularization parameter were chosen to achieve the minimum RMSE.

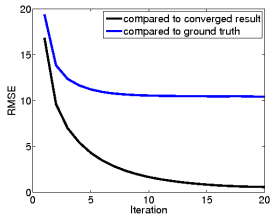


Fig. 4. Convergence of RMSE

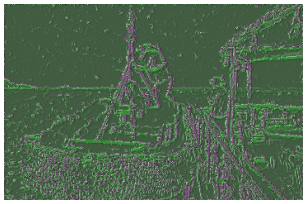
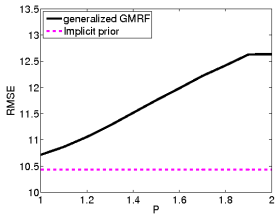
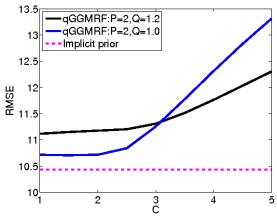


Fig. 5. The matrix $B_{s,r}$ entry.



(a) Comparison with GGMRF



(b) Comparison with qGGMRF

Fig. 7. These plots compare the RMSE of the implicit prior with the RMSE of the GGMRF and qGGMRF priors, respectively. Each plot varies a parameter of the model. In each case, the regularization parameter was chosen to achieve the minimum RMSE.

while the RMSE of GGMRF prior technique is plotted as a function of P values. Notice that the result of $C = 0$ for the qGGMRF with P and Q is equivalent to the result of $P = Q$ for GGMRF because in both these cases the prior corresponds to the prior term $\rho(\Delta) = |\Delta|^P$, within a multiplicative constant. Interestingly, the total-variation prior does not produce the minimum RMSE reconstruction, but it does achieve a value near the minimum .

5. CONCLUSIONS

We introduce a new method of MAP inversion which allows for the use of a Gibbs distribution which is only implicitly specified through the conditional probabilities of an MRF. The advantage of this approach is that it allows for the use of a much wider range of MRF models, which can in turn allow for more accurate modeling of data. The key to our approach is a method for explicitly computing local approximation to the energy function of the Gibbs distribution. These local approximations serve as a series of surrogate energy functions in the computation of the MAP inversion.

We provide a simple example of image denoising, but the method is generally applicable to any continuously valued MRF prior model, and could be combined with more sophisticated methods for estimation of the MRF's conditional distribution. Further-

more, the method has the potential for continuous improvement as methods for estimation of the MRF's conditional densities improve.

APPENDIX

A. CALCULATION OF D PARAMETER VECTOR

Assuming that both $u(x;x')$ and $u(x)$ are continuously differentiable functions of x, then the conditions of of (1) and (2) imply that the functions $u(x;x')$ and $u(x)$ must be tangent at $x = x'$. Therefore, we have that

$$\nabla_x u(x;x')\big|_{x=x'} = \nabla_x u(x)|_{x=x'} \ .$$

Using the form of (3), it is easily shown that

$$\nabla_x u(x;x')\big|_{x=x'} = d \ .$$

Now the partial derivative of $u(x)$ with respect to x_s can be evaluated by using the fact that

$$\frac{1}{z}\exp\{-u(x)\} = p(x) = p_s(x_s|x_{\partial s})p(x_r, r \neq s).$$

It is given by

$$\begin{aligned}
\frac{\partial u(x)}{\partial x_s} &= \frac{\partial}{\partial x_s}\{-\log p(x) - \log z\} \\
&= \frac{\partial}{\partial x_s}\{-\log\left(p_s(x_s|x_{\partial s})p(x_r\, r \neq s)\right)\} \\
&= -\frac{\partial}{\partial x_s}\log p_s(x_s|x_{\partial s}) - \frac{\partial}{\partial x_s}\log p(x_r\, r \neq s) \\
&= -\frac{\partial}{\partial x_s}\log p_s(x_s|x_{\partial s})
\end{aligned}$$

So equating the two expressions, we obtain the result.

$$d_s = -\left.\frac{\partial}{\partial x_s}\log p_s(x_s|x_{\partial s})\right|_{x=x'} \ .$$

B. NECESSARY CONDITIONS ON B

Proof of Condition 1

There must exist μ and c such that

$$u(x;x') = \frac{1}{2}(x-\mu)^t B(x-\mu) + c \ .$$

So therefore we know that

$$\begin{aligned}
u(x) &\leq u(x;x') \\
\frac{1}{z}\exp\{-u(x)\} &\geq \frac{1}{z}\exp\{-u(x;x')\} \\
\int p(x)dx = \int \frac{1}{z}\exp\{-u(x)\}dx &\geq \int \frac{1}{z}\exp\{-u(x;x')\}dx
\end{aligned}$$

So therefore we know that if B is not positive definite, then

$$\begin{aligned}
\int p(x)dx &\geq \int \frac{1}{z}\exp\{-\frac{1}{2}(x-\mu)^t B(x-\mu)\}dx + \frac{c}{z} \\
&\geq \int \frac{1}{z}\exp\{-\frac{1}{2}x^t Bx\}dx = \infty \ ,
\end{aligned}$$

which is a contradiction.

Proof of Condition 2

First we evaluate that the Hessian of the energy function. We can do this by using the fact that

$$\frac{1}{z}\exp\{-u(x)\} = p(x) = p_s(x_s|x_{\partial s})p(x_r, r \neq s) \ .$$

Then the partial derivatives of $u(x)$ with respect to x_s and x_r are given by

$$\begin{aligned}
H_{s,r} &= \frac{\partial^2 u(x)}{\partial x_s \partial x_r} \\
&= \frac{\partial}{\partial x_s \partial x_r}\{-\log p(x) - \log z\} \\
&= \frac{\partial}{\partial x_s \partial x_r}\{-\log\left(p_s(x_s|x_{\partial s})p(x_r\, r \neq s)\right)\} \\
&= -\frac{\partial}{\partial x_s \partial x_r}\log p_s(x_s|x_{\partial s}) - \frac{\partial}{\partial x_r}\left\{\frac{\partial}{\partial x_s}\log p(x_r\, r \neq s)\right\} \\
&= -\frac{\partial}{\partial x_s \partial x_r}\log p_s(x_s|x_{\partial s}) \ .
\end{aligned}$$

We next prove that when equations of (1) and (2) hold, then $B \geq H$. So it is enough to show that if the matrix $E = B - H$ is *not* positive semi-definite then equations of (1) and (2) are violated. Define the function $g(x) = u(x;x') - u(x)$. Then there exists a vector, v, such that $v^t E v < 0$, which implies that

$$\left.\frac{\partial^2 g(x' + \alpha v)}{\partial \alpha^2}\right|_{\alpha=0} = v^t E v < 0 \ .$$

Now we break the proof into two cases. First consider the case when $\frac{\partial g(x'+\alpha v)}{\partial \alpha} \neq 0$. In this case, there exists an ϵ such that $g(x'+\epsilon) < 0$, and the equations of (1) and (2) are violated. In the second case, $\frac{\partial g(x'+\alpha v)}{\partial \alpha} = 0$. For this case, since we know that $\frac{\partial^2 g(x'+\alpha v)}{\partial \alpha^2} < 0$, then we know that for some $\epsilon > 0$, $g(x'+\epsilon) < 0$. This also violates the equations of (1) and (2), so the result is proved.

Proof of Condition 3

We know that when equations of (1) and (2) hold, then in particular $\forall x_s$ and for $x_r = x'_r$ for $r \neq s$ $u(x;x')$ must also upper bound $u(x)$. This means that $\forall x_s$, we know that

$$\frac{1}{2}\Delta_s^2 B_{s,s} + \Delta_s d_s \geq -\log p_s(x_s|x'_{\partial s}) + \log p_s(x'_s|x'_{\partial s})$$

where $\Delta_s = x_s - x'_s$ and d_s must be chosen according to equation (4). Solving for $B_{s,s}$ yields

$$B_{s,s} \geq 2\frac{-\log p_s(x_s|x'_{\partial s}) + \log p_s(x'_s|x'_{\partial s}) - \Delta_s d_s}{\Delta_s^2} \ .$$

Since we know that this inequality must hold $\forall x_s$, then we have that

$$B_{s,s} \geq D_{s,s} = 2\sup_{x_s \neq x'_s}\left\{\frac{-\log p_s(x_s|x'_{\partial s}) + \log p_s(x'_s|x'_{\partial s}) - \Delta_s d_s}{\Delta_s^2}\right\}$$

Now observe that

$$\begin{aligned}
H_{s,s} &= -\frac{\partial^2}{\partial x_s^2}\log p_s(x_s|x_{\partial s}) \\
&= \lim_{x_s \to x'_s} 2\left\{\frac{-\log p_s(x_s|x'_{\partial s}) + \log p_s(x'_s|x'_{\partial s}) - \Delta_s d_s}{\Delta_s^2}\right\}
\end{aligned}$$

Which implies that $B_{s,s} \geq D_{s,s} \geq H_{s,s}$.

C. SURROGATE FUNCTIONS FOR EXPONENTIAL MIXTURES

Proof of Lemma

$$\begin{aligned}
\log f(x) &= \log f(x') + \log\left(\frac{f(x)}{f(x')}\right) \\
&= \log f(x') + \log\left(\sum_k \frac{w_k}{f(x')}\exp\{-u_k(x)\}\right) \\
&= \log f(x') + \log\left(\sum_k \left(\frac{w_k \exp\{-u_k(x')\}}{\sum_{k'} w_{k'}\exp\{-u_{k'}(x')\}}\right)\right. \\
&\quad \left.\times \exp\left\{-u_k(x) + u_k(x')\right\}\right) \\
&= \log f(x') + \log\left(\sum_k \tilde{\pi}_k \exp\{-u_k(x) + u_k(x')\}\right) \\
&\geq \log f(x') + \sum_k \tilde{\pi}_k\{-u_k(x) + u_k(x')\}
\end{aligned}$$

The last inequality results from Jensen's inequality. Taking the negative of the final expression results in

$$-\log f(x) \leq -\log f(x') + \sum_k \tilde{\pi}_k\{u_k(x') - u_k(x)\} = Q(x;x')\,.$$

and evaluating this result for $x = x'$ results in $-\log f(x') = Q(x';x')$.

REFERENCES

[1] S. Geman and D. McClure, "Bayesian images analysis: An application to single photon emission tomography," in *Proc. Statist. Comput. sect. Amer. Stat. Assoc.*, Washington, DC, 1985, pp. 12–18.

[2] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Trans. on Medical Imaging*, vol. 8, no. 2, pp. 194–202, June 1989.

[3] C. A. Bouman and K. Sauer, "A unified approach to statistical tomography using coordinate descent optimization," *IEEE Trans. on Image Processing*, vol. 5, no. 3, pp. 480–492, March 1996.

[4] Mario A. T. Figueiredo, Jose M. Bioucas-Dias, and Robert D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. on Image Processing*, vol. 16, no. 12, pp. 2980–2991, 2007.

[5] Peyman Milanfar, "A tour of modern image filtering," *IEEE Signal Proc. Magazine*, Invited feature article in review.

[6] A. Buades, B. Coll, and J. M. Morel, "Image denoising by non-local averaging," *Proc. of IEEE Int'l Conf. on Acoust., Speech and Sig. Proc.*, vol. 2, pp. 25–28, March 2005.

[7] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE Trans. on Image Processing*, vol. 16, no. 8, pp. 2080 – 2095, August 2007.

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning*, 2001, p. 282289.

[9] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. on Image Processing*, vol. 12, no. 11, pp. 1338 – 1351, November 2003.

[10] Michael Elad and Michal Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. on Image Processing*, vol. 15, no. 12, pp. 3736 – 3745, November 2006.

[11] Jean-Baptiste Thibault, Ken Sauer, Charles Bouman, and Jiang Hsieh, "A three-dimensional statistical approach to improved image quality for multi-slice helical ct," *Medical Physics*, vol. 34, no. 11, pp. 4526–4544, November 2007.

[12] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society B*, vol. 36, no. 2, pp. 192–236, 1974.

[13] L. Kantorovich and G. Akilov, *Functional Analysis in Normed Spaces*, Fitzmatgiz, Moscow, 1959.

[14] J. M. Ortega and W. C. Rheinoboldt, *Iterative Solutions of Nonlinear Equations in Several Variables*, pp. 253–255, Academic, New York, 1970.

[15] P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.

[16] A. R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," *IEEE Trans. Med. Imaging*, vol. 14, no. 1, pp. 132–137, 1995.

[17] Chuan-Sheng Foo, Chuong B. Do, and Andrew Y. Ng, "A majorization-minimization algorithm for (multiple) hyperparameter learning," in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, June 2009.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, chapter 6, Springer, New York, 2001.

[19] C. B. Atkins, C. A. Bouman, and J. P. Allebach, "Optimal image scaling using pixel classification," in *Proc. of IEEE Int'l Conf. on Image Proc.*, 2001, vol. 3, pp. 864–867.

[20] Hasib Siddiqui and Charles A. Bouman, "Training-based descreening," *IEEE Trans. on Image Processing*, vol. 16, no. 3, pp. 789–802, March 2007.

[21] C. A. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. on Image Processing*, vol. 2, no. 3, pp. 296–310, July 1993.