Information-Theoretic Image Reconstruction and Segmentation from Noisy Projections

Gerhard Visser, David L. Dowe, and Imants D. Svalbe

Monash University, VIC 3800, Melbourne, Australia
gerhardus.visser@infotech.monash.edu.au

Abstract. The minimum message length (MML) principle for inductive inference has been successfully applied to image segmentation where the images are modelled by Markov random fields (MRF). We have extended this work to be capable of simultaneously reconstructing and segmenting images that have been observed only through noisy projections. The noise added to each projection depends on the classes of the pixels (material) that it passes through. The intended application is in low-dose (low-flux) X-ray computed tomography (CT) where irregular projections are used.

1 Introduction

When the members of an observed set of data points each belong to a class from a finite set of classes, the task of inferring those classes is known as classification. When the number of the classes and their properties are not known, this is called clustering (or mixture modelling). If the data points have spatial components, as with images and tomograms, the word segmentation is often used.

This article focuses on segmentation where the pixels are located on a lattice (as with discrete images and tomograms) and the class of each pixel is positively correlated with the classes of its closest neighbours. The pixels are not observed directly but rather noisy projections (summations over subsets of pixels) have been observed.

An example of such a problem is low dose X-ray computed tomography. The pixel intensities are the absorption values at different locations, the classes are the materials (or like regions) and the projections are the sums of the pixel intensities along beams following some path through the object plus noise.

In section 4 we describe what future extensions must be done before a comparison with existing computed tomography (CT) methods can be made.

MML inductive inference has been successfully applied to clustering spatially-correlated data (including image segmentation) in [17], where the images are modelled by Markov random fields (MRF). That work has been extended in [14] to select between different MRF models. Our aim is to extend that work to the image reconstruction problem described above.

One of the advantages of MML inference in image segmentation is that it can infer the number of classes present and the parameters defining those classes. While these two features have not yet been implemented in this work we will

A. Nicholson and X. Li (Eds.): AI 2009, LNAI 5866, pp. 170–179, 2009.

describe how they can be implemented in section 4. For more details on the advantages of MML in classification and clustering see [5, 18, 20, 22].

Our models have been designed for problems where there is a lot of noise present in the observations (as with low-flux X-ray tomography). We will discuss how this work can be extended to infer the nature and degree of the noise from the observed data. For the case where there is little noise, non-probabilistic approaches (such as [13] and [6]) are preferable.

1.1 Minimum Message Length

Minimum message length (MML) [18, 19] is a Bayesian inference method with an information-theoretic interpretation. The minimum message length principle states that the best hypothesis is the one that gives the shortest explanation of the observed data using an optimal two-part encoding scheme.

By minimising the length of a two-part message of Hypothesis (H) followed by Data given Hypothesis ($D|H$), we seek a quantitative trade-off between the desiderata of model simplicity and goodness-of-fit to the data. This can be thought of as a quantitative version of Ockham's razor [10] and is compared to Kolmogorov complexity and algorithmic complexity [2, 9, 12] in [21]. For further discussions of these issues and for contrast with the much later minimum description length (MDL) principle [11], see other articles in that 1999 special issue of the *Computer Journal,* [3, sec. 11.4] and [18, chap. 10].

For our problem the hypothesis H would be some inferred value of x (the class assignments) and y (the pixel intensities) while the detail D is the observed noisy projection values s.

An earlier application of MML to image reconstruction from noisy projections was presented in [4]. While our methods are somewhat different, our aims are similar.

2 Model and Solution

2.1 Image Reconstruction Model

This subsection describes the probabilistic model used. Let $y = (y_1, y_2, ...y_N)$ be a set of pixel intensities, indexed by N locations, with $x_i \in \{1, 2, ..., C\}$ being the class of pixel $y_i \in \{0, 1, ..., B\}$. Here $B + 1$ is the number of intensity values that a pixel can have and C is the number of classes that a pixel can be assigned to. The sequence of projections is $q = (q_1, q_2, ..., q_M)$ where each projection $q_i \subseteq \{1, 2, ..., N\}$ represents a group of locations. For X-ray tomography these groups would be paths followed by the X-ray beams.

The members of x are arranged on a lattice and the a priori distribution over x ($P(x)$) forms a Markov random field (MRF) for a neighbourhood structure defined over the members of x. The neighbourhood structure is defined by a set of neighbours $n_i \subset \{1, 2, .., N\}$ for each location $i \in \{1, 2, .., N\}$. If $i \in n_k$ then $k \in n_i$. For $P(x)$ to be a MRF it is required that $\forall x, P(x) > 0$ and that,

$$P(x_i|x_{\forall j \neq i}) = P(x_i|x_{\forall j \in n_i}) \; . \tag{1}$$

The second condition (expressed in equation 1) states that the class of a point is conditionally independent of all other points given the classes of its neighbours. We assume that the parameters defining the right hand side of equation (1) are known.

A variety of MRF models that can be used exist in the image segmentation literature, following [17] we will use the auto-logistic model on a toroidal square lattice (see section 2.2).

For now we assume that it is known a priori what the class distributions (defined by $P(y_i|x_i)$) are. For all work presented in this article the classes are defined as Poisson distributions, $P(y_i|x_i) = e^{-\lambda_{x_i}} \lambda_{x_i}^{y_i} / y_i!$. Here x_i is the class that location i is assigned to, so λ_{x_i} is the mean associated with class $x_i \in \{1, 2, ..., C\}$.

The values of the data points y_i are given an upper limit and the probabilities for all values greater than B are added to $P(y_i = B)$. The vector of class parameters λ is also assumed to be known.

Our task is to infer values for x and y given a set $s = \{s_1, s_2, ..., s_M\}$ of noisy summations over the projections q.

$$s_j = \sum_{r \in q_j} (y_r + z_{j,r}) \tag{2}$$

Here s_j is a summation over q_j and $z_{j,r}$ is the noise in s_j contributed by location r. The noise $z_{j,r}$ is assumed to be normally distributed with zero mean and a standard deviation that depends on the class x_r of location r, denoted by σ_{x_r}.

$$P(z_{j,r}|x_r) = \frac{1}{\sigma_{x_r}\sqrt{2\pi}} \exp\{-\frac{z_{j,r}^2}{2\sigma_{x_r}^2}\} \tag{3}$$

The standard deviation σ_k of each class $k \in \{1, 2, ..., C\}$ is for now assumed known. Let z_j be the sum of noise terms for q_j,

$$z_j = \sum_{r \in q_j} z_{j,r} \ . \tag{4}$$

2.2 Auto-logistic Model

We have used in our tests the auto-logistic model to express the spatial relations between members of x (the class assignments). The auto-logistic model reduces to the well-known Ising model when the number of classes is equal to two ($C = 2$). When expressed as a Gibbs random field, the prior over x takes the form,

$$P(x) = \frac{1}{U} \exp[\sum_{i=1}^{N} \log \alpha_{x_i} - w\beta] \tag{5}$$

where w is the number of neighbour pairs (over the entire lattice) that do not have the same class values. Remember that each location has a set of neighbouring locations. For our tests the locations are arranged on a toroidal square-lattice so that each location has four neighbours (left, right, above and below).

The parameters $\alpha = (\alpha_1, \alpha_2, ..., \alpha_C)$ and β are assumed to be known. The vector α is analogous to mixing proportions of the C classes while β determines the degree to which neighbouring classes agree. For example if $x_i = 2$ then $\alpha_{x_i} = \alpha_2$ is the value associated with class 2. Note that the values α_i are equivalent to mixing proportions only when $\beta = 0$. Higher values of β lead to neighbouring class assignments being more likely to have the same value while $\beta = 0$ makes the members of x independent of each other. With w_i defined as the number of neighbours of location i that do not have the same class value as x_i it can be shown that,

$$P(x_i|x_{\forall j \neq i}) = P(x_i|x_{\forall j \in n_i}) \propto \exp\left[\log \alpha_{x_i} - w_i \beta\right] . \tag{6}$$

2.3 The Message Length

This section describes our MML solution to the problem described in section 2.1. All message lengths in this article are measured in nits where 1 bit is equal to $\log_e(2)$ nits.

Given the set of noisy summations s, estimates x (for the class assignments) and y (for the data points) are to be inferred. The optimal estimate is the one that leads to the shortest message length as described in section 1.1. For a given point estimate (a pair $(\hat{x}, \hat{y})$) this code is,

1a. an encoding of $\hat{x}$
1b. an encoding of $\hat{y}$ given $\hat{x}$
2. an encoding of the observed s given $\hat{x}$ and $\hat{y}$

Here the encoding of parts 1a and 1b is known as the assertion and part 2 is known as the detail. The objective function which we try to minimise is the message length, which is equal to the length of assertion plus the length of the detail.

We approximate this message length L as follows,

$$L = T_1 + T_2 + T_3 - T_4 \tag{7}$$

where,

1. T_1 is the length for encoding $\hat{x}$ precisely
2. T_2 is the length for encoding $\hat{y}$ precisely given $\hat{x}$
3. T_3 is the length for encoding s given $\hat{y}$ and $\hat{x}$
4. T_4 is the entropy of the pair $(\hat{y}, \hat{x})$ given s

The first term is equal to the negative log likelihood of $\hat{x}$,

$$T_1 = -\log P(\hat{x}) . \tag{8}$$

We describe in section 2.6 how this can be approximated following [17]. The second term is,

$$T_2 = -\sum_{i=1}^{N} \log P(\hat{y}_i|\hat{x}_i) . \tag{9}$$

To encode s_j given $\hat{y}$ and $\hat{x}$ we need only specify the noise term z_j (see subsection 2.1). These noise terms are normally distributed with mean zero and standard deviation σ_j. The standard deviations depend on $\hat{x}$ and are,

$$\sigma_j^2 = \sum_{i \in q_j} \sigma_{\hat{x}_i}^2 \ . \tag{10}$$

T_3 is then the sum of the negative log likelihoods of the noise terms z_j, using the above values for σ_j.

Finally, in an optimal code, $\hat{y}$ and $\hat{x}$ do not need to be stated precisely. The number of alternatives that could have been used giving a similar value for $T_1 + T_2 + T_3$ can be approximated by e^{-T_4}. This means that if the pair $(\hat{y}, \hat{x})$ was stated imprecisely it would have cost approximately T_4 nits less. This bits-back approach to calculating the message length was introduced in [15]. The next subsection describes how T_4 can be approximated.

2.4 The Precision of $\hat{y}$ and $\hat{x}$

The entropy of (y, x) given s is defined as,

$$T_4 = -\sum_{\forall (y,x)} P(y, x|s) \log P(y, x|s) \tag{11}$$

Performing the summation over all possible values of y and x is impractical so a numerical approximation for T_4 is used. To explain this approximation we must first express the distribution $P(y, x|s)$ as a Gibbs random field (GRF). First note that applying Bayes's rule,

$$P(y, x|s) \propto P(s|x, y)P(x, y) = P(s|x, y)P(y|x)P(x) \ . \tag{12}$$

Define the energy V of (y, x) given s as,

$$\begin{aligned} V(y, x|s) &= -\log\left[P(s|x, y)P(y|x)P(x)U\right] \\ &= -\log P(x)U - \textstyle\sum_{i=1}^{N} \log P(y_i|x) - \sum_{j=1}^{M} \log P(s_j|y, x) \\ &= -\log P(x)U - \textstyle\sum_{i=1}^{N} \log P(y_i|x_i) - \sum_{j=1}^{M} \log P(z_j|x) \\ &= -[\textstyle\sum_{i=1}^{N} \log \alpha_{x_i} - w\beta] - \sum_{i=1}^{N} \log P(y_i|x_i) - \sum_{j=1}^{M} \log P(z_j|x) \ . \end{aligned} \tag{13}$$

Note that given y, x and s this energy $V(y, x|s)$ can be easily calculated. We can now rewrite $P(y, x|s)$ as a GRF,

$$P(y, x|s) = e^{\log P(y,x|s)} = \frac{1}{Z} e^{-V(y,x|s)} \tag{14}$$

where Z is called the partition function and is independent of x and y. Next define,

$$P_T(y, x|s) = \frac{1}{Z_T} e^{-V(y,x|s)/T} \tag{15}$$

as the distribution (over x and y given s) at temperature T. As T increases this distribution reaches its maximum possible entropy. Note that at $T = 1$ this distribution is equivalent to the original distribution $P_1(y, x|s) = P(y, x|s)$. The entropy of this distribution at temperature T is,

$$H_T(y, x|s) = -\sum_{\forall(x,y)} P_T(x, y|s) \log P_T(x, y|s) \ . \tag{16}$$

The expected energy at temperature T is,

$$Q_T = \sum_{\forall(x,y)} P_T(x, y|s) V(y, x|s) \ . \tag{17}$$

It can be shown that $\frac{dH_T}{dT} = \frac{dQ_T}{dT}/T$ (hence $dH_T = dQ_T/T$). Gibbs sampling can be used to sample random states of (y, x) given T and s, and hence Q_T can be approximated at any temperature by averaging the energies of those samples.

At $T = \infty$ the entropy H_T attains its maximum value, which is $N \log(CB)$. The entropy of the distribution at temperature $T = 1$ can be calculated as follows. Starting at $T = 1$ and slowly incrementing T up to some value high enough to give a distribution similar to that attained at $T = \infty$, calculate Q_T at each temperature increment. By subtracting the term dQ_T/T at each increment from the maximum entropy, we end with a good estimate of $H_1 = T_4$.

Note that using Gibbs samples from the distribution at each temperature is computationally expensive and to get a good estimate requires that small increments be used [17, Sec. 5.6]. The Gibbs sampling process is discussed in the following subsection.

Q_∞ can be approximated by sampling from the maximum entropy distribution over x and y. It is simple to show then that the error (in calculating H_1) caused by terminating at temperature $T = t$ instead of $T = \infty$ is no greater than $(Q_\infty - Q_t)/t$. This can be used to determine when to terminate the algorithm.

2.5 Estimating $\hat{y}$ and $\hat{x}$ to Optimise the Message Length

For high-dimensional vectors of discrete parameters (such as the one defined by a pair (x, y)) a random selection from the posterior distribution $P(y, x|s)$ can be used as the MML estimate. This type of estimate is discussed in [16] and is also used in [17, sec. 5.1] and [14].

To create such samples we use the Gibbs sampler. This works by repeatedly choosing a random member of (x, y) and changing its value according to its probability distribution given all other values. For example, if x_i is selected then it is re-assigned according to $P(x_i|y, s, x_{\forall k \neq i})$. If this process is repeated for long enough the resulting pair (x, y) can be considered a pseudo-random sample from $P(y, x|s)$.

The same process can be used to sample from $P_T(y, x|s)$ (equation 15) to calculate the approximation for T_4 (equation 11) described in subsection 2.4.

When there is very little noise in the observations s, sampling at temperatures close to $T = 1$ can be difficult due to there being many local minima for the

Gibbs sampler to fall into. This problem can be addressed by using a variation of simulated annealing. In fact, by using simulated annealing the task of finding an estimate $(\hat{x}, \hat{y})$, and the calculation of T_4 (section 2.4), can be done in one step. This is achieved by starting sampling at a high temperature (as described in the last paragraph of section 2.4) and gradually lowering the temperature to $T = 1$. The changes in Q_T are recorded at each decrement and used to calculate T_4 while the final state of (x, y) is used as our MML estimate.

Note that the estimates can be obtained without calculating the message length. There are two uses for calculating the message lengths in our problem. The first is that in some cases multiple runs of the estimation algorithm described above will settle in separate local minima. The message length is a measure of the explanatory value of a hypothesis and can select between these. The second use is for determining the number of classes that can be justified by the data (section 4) [17, 18, 20, 22].

2.6 The Length for Encoding $\hat{x}$

As in section 2.4 equation (15), we define $P_T(x)$ as the distribution over x at temperature T. Since $P(x)$ is a Markov random field (MRF) it can be restated as a Gibbs random field (GRF). This is guaranteed by the Hammersley-Clifford theorem, proven in [1, 7, 8].

This allows us to approximate $H_1(x)$ (the entropy of x at temperature $T = 1$) using the method described in section 2.4. The negative log likelihood of our estimate for $\hat{x}$ is then calculated using,

$$-\log P(\hat{x}) = H_1(x) + V(\hat{x}) - Q(x) \tag{18}$$

where $V(\hat{x})$ is the energy of our estimate $\hat{x}$ and $Q(x)$ is the expected energy for the GRF over x. This type of calculation has also been used to calculate the message lengths for other image models in [17] and [14] and is discussed there in more detail.

3 Test on Artificial Data

For this test there are three classes $C = 3$. The auto-logistic model is used to define the prior $P(x)$ with parameters $\alpha = (1/3, 1/3, 1/3)$ and $\beta = 0.7$.

Similarly the vector of class parameters (class means) is $\lambda = (\lambda_1, \lambda_2, \lambda_3) = (5, 20, 35)$ and the noise parameters (standard deviations) for the three classes are $\sigma = (\sigma_1, \sigma_2, \sigma_3) = (1, 2, 3)$.

From this model, instances of x, y and s were generated with $N = 225$ locations arranged on a 15×15 toroidal square-lattice. The use of a toroidal square-lattice is simply for programming reasons and is not required by our method. The number of projections is $M = 225$ each containing 15 locations. The algorithm was run given s to infer estimates $\hat{y}$ and $\hat{x}$. The true x and inferred $\hat{x}$ class assignments are shown in figure 1. The true y and inferred $\hat{y}$ data point values are also shown in figure 1.

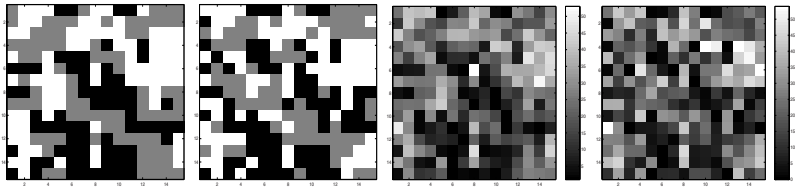

Fig. 1. Far left is the true class assignment vector x with class 1 ($\lambda_2 = 5$) as black, class 2 ($\lambda_2 = 20$) as grey and class 3 ($\lambda_3 = 35$) as white. Centre left is the inferred set of class assignments $\hat{x}$. Centre right is the true value of y and on the far right is the inferred estimate $\hat{y}$. The intensities range from white $y_i = 60$ to black $y_i = 0$ as shown by the bars to the right of the two rightmost images.

The message length calculated as $L = T_1 + T_2 + T_3 - T_4$ was $L = 1740$ with the individual terms being $T_1 = 228$, $T_2 = 664$ $T_3 = 1387$ and $T_4 = 540$. The value of T_4 tells us that there are roughly e^{540} different solutions for the pair (x, y) that are reasonable estimates and gives us some measure of the amount of noise present.

For comparisons to other work to be meaningful our work will have to be developed further. This paper is intended to show how the MML approach to intrinsic classification of spatially correlated data introduced by Wallace [17] can be applied to image reconstruction. The next section discusses what extensions are necessary and how they can be implemented.

4 Further Work

The first problem is with computational expensiveness. Our current implementation is in Java (not a performance language) and little effort was made to make it fast. This Gibbs sampling algorithm is highly parallelisable and specialised hardware is often used in image processing. Before we optimise our implementation we wish to first improve it in other respects.

The earliest applications of the minimum message length principle is in mixture modelling (clustering) [17, 19, 20, 22]. A strong point of MML in this area is the ability to estimate both the class parameters and the number of classes. The next step for our work would be to add those abilities. It should be possible to achieved this using the EM algorithm,

1. initialise all parameters
2. re-estimate x and y using the Gibbs sampler
3. re-estimate the parameters defining the class distributions λ
4. re-estimate the parameters defining $P(x)$
5. re-estimate the projection noise parameters σ
6. if the estimate is stable then stop, else return to step 2

This algorithm should gradually move towards a local minimum in the message length as each individual step reduces it.

To estimate the number of classes, the algorithm above is run several times assuming a different number of classes each time. The number of classes that leads to the shortest message length is preferred.

5 Conclusion

We have shown how Minimum Message Length (MML) can be used to reconstruct and classify (or segment) data sets (images/tomograms) that have been observed only through noisy projections. As a quantitative version of Ockham's razor [10], MML separates noise and pattern information using prior (domain specific) knowledge and it is capable of performing well on noisy data, while being resistant to overfitting. For this reason, applications of MML to low-dose computed tomography are worth exploring.

We have demonstrated how the classification, reconstruction and message length calculations can be done following the approach of [17]. The next step will be to add the ability to infer the class parameters, the noise parameters and the number of classes.

References

1. Besag, J.E.: Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society B36(2), 192–236 (1974)
2. Chaitin, G.J.: On the length of programs for computing finite binary sequences. Journal of the Association of Computing Machinery 13, 547–569 (1966)
3. Comley, J.W., Dowe, D.L.: Minimum message length and generalized Bayesian nets with asymmetric languages. In: Grünwald, P., Pitt, M.A., Myung, I.J. (eds.) Advances in Minimum Description Length: Theory and Applications, pp. 265–294. MIT Press, Cambridge (2005)
4. Dalgleish, A.P., Dowe, D.L., Svalbe, I.D.: Tomographic reconstruction of images from noisy projections - a preliminary study. In: Orgun, M.A., Thornton, J. (eds.) AI 2007. LNCS (LNAI), vol. 4830, pp. 539–548. Springer, Heidelberg (2007)
5. Dowe, D.L.: Foreword re C. S. Wallace. Computer Journal 51(5), 523–560 (2008)
6. Fayad, H., Guedon, J.P., Svalbe, I.D., Bizais, Y., Normand, N.: Applying mojette discrete radon transforms to classical tomographic data. In: Medical Imaging 2008: Physics of Medical Imaging. Proceedings of the SPIE, April 2008, vol. 6913, p. 69132S (2008)
7. Geman, S., Geman, D.: Stochastic relaxations, Gibbs distributions and the Bayesian restoration of images. IEEE Tran. on PAMI PAMI-6, 721–741 (1984)
8. Grimmett, G.R.: A theorem about random fields. Bull. London Math. Soc. 5, 81–84 (1973)
9. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems of Information Transmission 1, 1–17 (1965)
10. Needham, S.L., Dowe, D.L.: Message length as an effective Ockham's razor in decision tree induction. In: Proc. 8th Int. Workshop of Artificial Intelligence and Statistics (AISTATS 2001), Key West, FL, pp. 253–260 (2001)
11. Rissanen, J.: Modeling by the shortest data description. Automatica 14, 465–471 (1978)

12. Solomonoff, R.J.: A formal theory of inductive inference. Information and Control 7, 1–22, 224–254 (1964)
13. Svalbe, I., van der Speck, D.: Reconstruction of tomographic images using analog projections and the digital radon transform. Linear Algebra and its Applications 339, 125–145 (2001)
14. Visser, G., Dowe, D.L.: Minimum message length clustering of spatially-correlated data with varying inter-class penalties. In: Proc. 6th IEEE International Conference on Computer and Information Science (ICIS 2007), Melbourne, Australia, July 2007, pp. 17–22 (2007)
15. Wallace, C.S.: An improved program for classification. In: Proceedings of the Nineteenth Australian Computer Science Conference (ACSC-9), Monash University, Australia, vol. 8, pp. 357–366 (1986)
16. Wallace, C.S.: False Oracles and SMML Estimators. In: Proc. Information, Statistics and Induction in Science conference (ISIS 1996), Was Tech Rept TR 89/128, Monash University, Australia, pp. 304–316. World Scientific, Singapore (1996)
17. Wallace, C.S.: Intrinsic classification of spatially correlated data. Computer Journal 41(8), 602–611 (1998)
18. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. Springer, Heidelberg (2005)
19. Wallace, C.S., Boulton, D.M.: An information measure for classification. Computer Journal 11, 185–194 (1968)
20. Wallace, C.S., Dowe, D.L.: Intrinsic classification by MML - the Snob program. In: Proc. 7th Australian Joint Conf. on Artificial Intelligence, pp. 37–44. World Scientific, Singapore (1994)
21. Wallace, C.S., Dowe, D.L.: Minimum message length and Kolmogorov complexity. Computer Journal 42(4), 270–283 (1999)
22. Wallace, C.S., Dowe, D.L.: MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. Statistics and Computing 10, 73–83 (2000)