Space-Variant Model Fitting and Selection for Image Information Extraction

Matteo Soccorsi*, Marco Quartulli† and Mihai Datcu*

*German Aerospace Center (DLR),
Remote Sensing Institute, Image Analyzing (IMF-BW)
Oberpfaffenhofen D-82234, Wessling, Germany
†Advanced Computer System,
Via della Bufalotta 378, 00139 Rome, Italy

Abstract. With the growing importance of model-based signal analysis methods, the dependence of their performance on the choice of the models needs to be addressed. Bayesian theory incorporates model selection in a natural and direct way: we apply it to the space-variant choice of the best model in a given reference class in the framework of parameter estimation from complex data. In particular, we introduce an algorithm for image information extraction that is based on a two-level model, it estimates local texture Gauss-Markov Random Field (GMRF) parameters and local GMRF model order for incomplete data. Model selection is based on an approximate numerical computation of the evidence integral. Results are presented on Synthetic Aperture Radar (SAR) images.

Key Words: GMRF, Model Selection, Model Fitting, Parameter Estimation, Evidence.

INTRODUCTION

Image modeling and information extraction

Model based filtering and information extraction belong to the literature of image processing since a number of years. The seminal works by Besag [2][4] were instrumental in introducing the ideas of stochastic Markov modeling properties to the field of image processing. Geman and Geman's article [7] introduced the techniques of Gibbs modelling and sampling to the field. Their ideas were applied in providing solutions to a number of problems from noisy image data filtering [3] to modeling feedback in human-computer interaction studies [14]. Bayesian analysis is often characterized by the fundamental role played in it by a priori distributions. The usage of subjective ones has often been the ground for objections and controversies. Jeffreys [9] and Jaynes [6] layed the ground for the development of techniques that can be applied to generate objective a priori descriptions starting from a set of objective constraints to the problem and from the principle of Maximum Entropy. When a whole class of a priori descriptions of the phenomenon under analysis is available instead, a principled choice of the most probable model according to the data can be made by the second level of Bayesian inference, by selecting the one that maximizes a second level Maximum A Posteriori estimation criterion [5]. A number of the principles and techniques of hierarchical Bayesian

modelling and two-level Bayesian inference for the modelling and estimation of noisy, non-stationary 2D signals are summarized in [11] and [12]. These works introduce the general problem of estimation theory in a Bayesian framework centering on the properties of 2D Markov random fields and their role in estimation. The focus of [13] centers instead on the extraction of reliable estimates of the parameters of these models from noisy, non-stationary observations in a two-levels Bayesian modelling approach. Gauss-Markov random fields are used to describe textured clean backscatter images corrupted by noise. The described system performs an estimation of the texture parameters of the clean image. The order of the model that is used as a priori description of the data is not an object of the estimation, though, and is considered a fixed input parameter instead.

HIERARCHICAL BAYESIAN MODELLING AND INFERENCE

Bayesian inference and MAP estimation

In Bayesian probability theory, logical links are expressed by means of conditional probability distributions

$$p(y|x) = \frac{p(x,y)}{p(x)}. \tag{1}$$

It expresses the degree of belief that an event y takes place given the occurrence of an event x. An immediate consequence of the definition of conditional probability is the so-called Bayes' law

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}, \tag{2}$$

which enables the reversal of probabilistic links and threfore it allows a direct model based inference. The law can be seen as a rule for updating an existing description, the Prior $p(y)$, of a phenomenon y, based on new information-new data or a new description of the phenomenon x.
The direct link from the old to the new description is modelled by the likelihood $p(x|y)$. Furthermore, the evidence normalization term, $p(x)$, describes the distribution of the data and it can be computed by marginalization:

$$p(x) = \int p(x|y)p(y)dy. \tag{3}$$

The posterior description of phenomenon y is often summarized in terms of the position of its maximum, by means of the Maximum a Posteriori (MAP) estimator

$$\hat{y}_{\mathrm{MAP}} = \arg\max_{y} p(y|x). \tag{4}$$

We observe that, in classical estimation theory, using a cost function is nothing else but describing a type of Prior information. The expression for the posterior encapsulates

the deterministic Prior knowledge represented by the forward model. In addition, the knowledge about the observation noise and the a priori information about the desired parameter are also included. We conclude that MAP is a complete frame for model-based approaches in information extraction. This can be demonstrated [15] to be equivalent to a Minimum Description Length (MDL) estimate obtained by considering that the best model of a phenomenon is the one that produces the most compact encoding of it. A very similar approach, also considering two terms, a data one requiring the maximization of a likelihood and a penalty term considering the complexity of the model, is the Akaike Information Criterion (AIC) [8].

Hierarchical models

As noted by O'Hagan in [1], the $p(y|x)$ posterior statistical model and $p(y)$ Prior model together form an ordered structure in which the distribution of the data x is written conditionally on parameter y as $p(x|y)$. The Prior distribution of y can be conditioned by an hyper-parameter z as $p(y|z)$ and completed by the distribution of z, $p(z)$. Since we can go further and write the conditionally distribution of z on an *hyper-hyper* parameter t as $p(z|t)$, and we can continue this process as long as necessary, it is generated a hierarchical model. The distribution of the parameter at any level of the hierarchy depends, by conditioning, on the parameter at lower level and it is independent from the parameters at all levels below it. For instance, if we model the distribution of y in terms of $p(y|z)$ and $p(z)$, the likelihood $p(x|y)$ will be formally the distribution of x given y and z. If we write $p(x|y)$, it means that if we know y then knowing z will not add any information about x. This is reasonable because z has been introduced only as a way of formulating $p(y)$. The reason for making this interpretation of $p(x|y)$ is that otherwise the distributions of $p(x|y)$, $p(y|z)$ and $p(z)$ together do not completely specify the joint distribution of x, y and z. Thus, this extra assumption allows us to write:

$$p(x, y, z) = p(x|y)p(y|z)p(z). \tag{5}$$

A hierarchical model specifies always the full joint distribution of all quantities in the previous way.

Principle of Inference

We consider that each model H_i has a vector of parameters θ. A model is defined by its functional form and two probability distributions: the Prior distribution $p(\theta|H_i)$ which states what values the model's parameters might plausibly take; and the prediction $p(D|\theta; H_i)$ that the model makes about the data D when its parameter θ has a particular value. Note that models with the same parameterisation but different Prior over the parameters are defined as different models. At the first level of inference, we assume that one model H_i is true, and we infer the value of the parameter θ given the data D.

Using Bayes' rule in eq. 2, the posterior probability of the parameters θ is:

$$p(\theta|D;H_i) = \frac{p(D|\theta;H_i)p(\theta|H_i)}{p(D|H_i)}. \tag{6}$$

The normalization constant $p(D|H_i)$ is commonly ignored, since it is irrelevant to the first level of inference, i.e., the estimation of θ. It is important in the second level of inference, and we name it the evidence for H_i .

Occam razor and Occam factor

As noted by [5], model comparison is a difficult task because it is not possible simply to choose the model that fits the data best since more complex models can always fit the data better. Then the maximum likelihood model choice leads us inevitably to implausible over-parameterized models which generalize poorly. *Occam's razor* is the principle that states that unnecessarily complex models should not be preferred to simpler ones. Since Bayesian method automatically and quantitatively embodies *Occam's razor* [16][9], without the introduction of any penalty terms, complex models are automatically self-penalized under Bayes' rule.
This is useful at the second level of inference where we wish to infer which model is most plausible given the data. The posterior probability of the model is:

$$p(H_i|D) \propto p(D|H_i)p(H_i), \tag{7}$$

where the data-dependent term $p(D|H_i)$ is the evidence for H_i. It appears as the normalizing constant in eq. 6. The second term, $p(H_i)$, is a *subjective* Prior over the hypothesis space. It is kept constant when there is no reason to assign strongly differing prior $p(H_i)$ to the alternative models. In order to assign a preference to alternative models H_i, the evidence has to be evaluated, since it embodies the *Occam's razor* as shown below. The evaluation of the marginalization integral of eq. 3

$$p(D|H_i) = \int p(D|\theta,H_i)p(\theta|H_i)d\theta \tag{8}$$

can be conducted by approximating the posterior as a Gaussian around its MAP peak, using Laplace's method:

$$p(D|H_i) \approx p(D|\hat{\theta}_{\mathrm{MAP}},H_i)\cdot p(\hat{\theta}_{\mathrm{MAP}}|H_i)\det{}^{-\frac{1}{2}}(A/2\pi) \tag{9}$$

where the Hessian $A = \nabla\nabla \log p(\theta|D;H_i)$ appears in the last two terms which account for the *Occam factor*:

$$\Omega \equiv p(\hat{\theta}_{\mathrm{MAP}}|H_i)\det{}^{-\frac{1}{2}}(A/2\pi) < 1. \tag{10}$$

It penalizes H_i for depending on the parameter θ. Thus, the evidence is obtained by multiplying the best fit likelihood by the Occam factor, which favors in the set of models the less complicated ones. The maximization of the evidence is therefore regarded as a criterion for the choice of a suitable model to explain a given dataset.

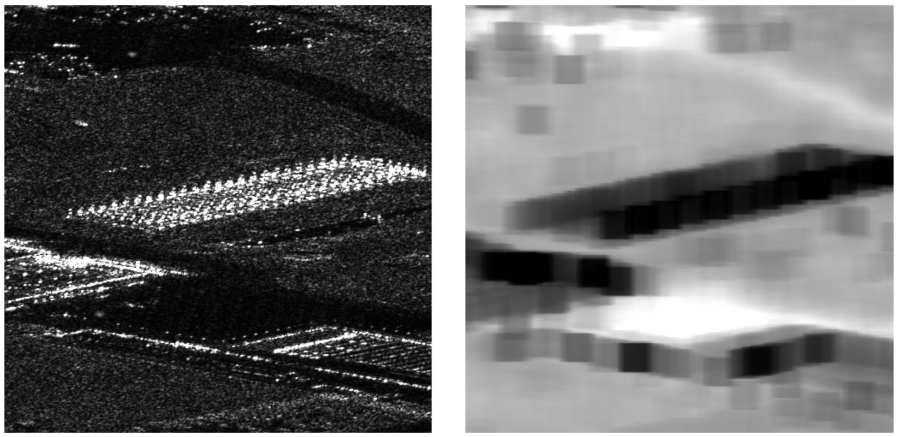

FIGURE 1. Tile of 500x500 pixels from E-SAR SLC image of Oberpfaffenhofen, Germany. On the left, the original image which shows strong scatterers from a large metallic structure (range resolution 1.99 m, azimuth resolution 0.72 m). On the right, the evidence of the image for the GMRF model: it is possible to distinguish three main areas coded by black, grey and white.

IMAGE MODELS

We investigate the analysis of image data acquired by coherent systems, such as echography, sonar, SAR and computer tomography. We consider the full information contained in both amplitude and phase instead of consider only the description in terms of the local image intensities. This because the intensity images are affected by coherent superposition of many scatterers responses which populate the resolution cell. It appears as a well-known kind of strong multiplicative noise called *speckle* [10], which is not visible in the fully description given by amplitude and phase. The Prior model we consider for our image data is the Gauss-Markov Random Field (GMRF) [11][12]:

$$p(X_s|X_r, r \in \mathcal{N}, \sigma, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[- \frac{(X_s - \sum_{r \in \mathcal{N}} \theta_{r_r} X_{s+r})^2}{2\sigma^2} \right] \qquad (11)$$

specified by σ and by the parameter vector $\boldsymbol{\theta} = (\theta_{r_0}, \ldots \theta_{r_m})$ defined on the neighborhood of cliques $\mathcal{N}$ centered on the generic pixel X_s such that the scalar parameters are symmetric around the central element. The main strength of the Gauss-Markov model lies in the ability to model structure in a wide set of locally stationary textured images while still allowing analytical tractability. The data model we consider is adapted to model real and imaginary part of the complex signal, i.e. the two orthogonal channel of the coherent system. The likelihood therefore employed in the Bayes equation is a

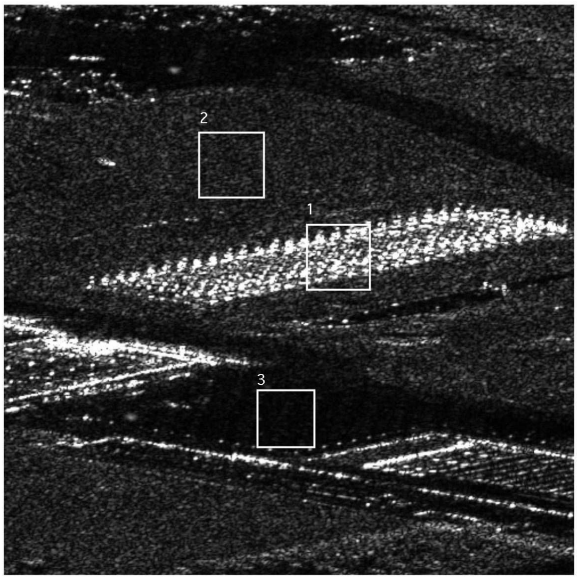

FIGURE 2. From the original image we selected three different areas: the first one from the large metallic structure, the second from grass landscape and the third from asphalt pavement. We expected to have different model order in the space variant model selection. The areas appear different both in the real image and in the evidence map.

space-variant circular complex Gaussian distribution with zero mean:

$$p(x_i, y_i) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{x_i^2 + y_i^2}{2\sigma^2}\right] \tag{12}$$

where x_i and y_i are the two orthogonal channel for the real and imaginary part of the complex signal.

MODEL ORDER SELECTION

Different neighborhood sizes of the Gauss-Markov model can be used for both information extraction from image data and for image filtering. The selection of the most probable model order is performed within the Bayesian framework. We maximize $p(y|\theta, M_{order})$ as a function of θ for a given model order M_{order} , i.e. we perform a model parameter estimation for different neighborhood sizes and we choose the one with highest evidence. Since computing time increases with the number of models to test, we usually use a fixed model order for practical application. We remind that a full model order

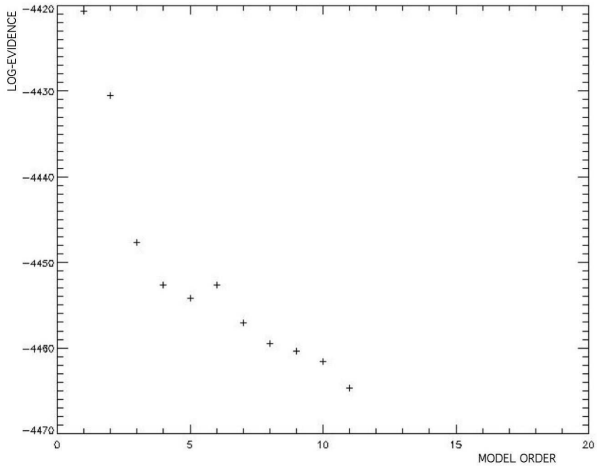

FIGURE 3. Plot of the evidence vs model order for the first area (metallic structure). The local maximum represents the best model explaining the space diversity of the data. It corresponds to model order 6.

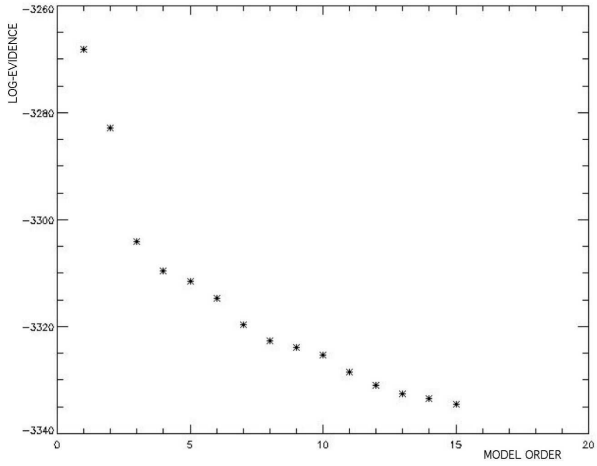

FIGURE 4. Plot of the evidence vs model order for the second area (grass landscape). The evidence does not show any local maxima. In this case a simple model has to be preferred from a more complex one, so a good choice is the model of order 4 for which the evidence starts to become stable.

selection independent of the estimate $\hat{\theta}$ requires an additional Bayesian layer becoming a really complicated task. If a MAP estimation system is used for the space-variant model order selection, the results in the evidence and model order map tend to discriminate the different phenomena.

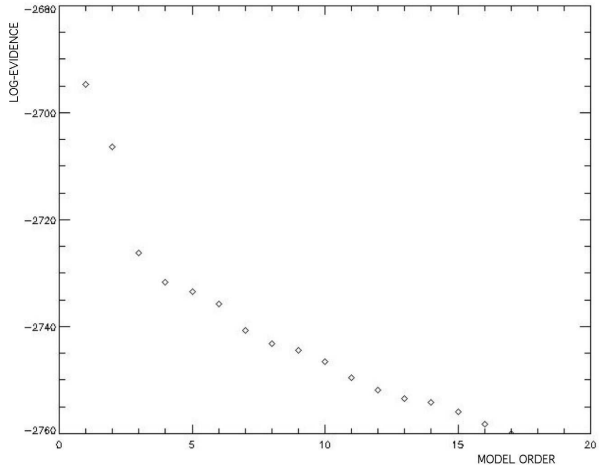

FIGURE 5. Plot of the evidence vs model order for the third area (asphalt pavement). As for the previous area the evidence does not show any local maxima. Then even in this this case a simple model has to be preferred from a more complex one. The model of order 4 is a reasonable choice.

CASE STUDIES: MODEL ORDER SELECTION FOR SPACE-VARIANT INFORMATION EXTRACTION FROM SAR IMAGE

We give example of information extraction and model order selection on SAR imagery which is a coherent imaging system that operates in the microwave domain, ranging from 30 meters to sub-meter resolution. We consider an E-SAR SLC image of Oberpfaffenhofen, Germany (figure 1, left). In the image is clearly visible a large geometrical metallic structure characterized by a set of strong scatterers. It is the skeleton of a building under construction surrounded by a grass land. Figure 1, on the right, shows the evidence of the original image for GMRF model of fixed order. Since the computing time increases with the complexity of the model, we selected from the original image three areas showing different textures. Figure 2, shows the selected areas: the first one from the strong scatterers structure, the second from grass landscape and the third from asphalt pavement. It is also possible to distinguish the three different areas in the evidence map of figure 1, in which they are coded in grey-scales. We performed the model selection computing the evidence for model order from one, for the first area up to eleven, for the second area up to fifteen and for the third area up to seventeen. The results are presented in figures 3, 4 and 5, where in the abscissa is the model order and in the ordinate is the correspondent value of the log-evidence. Figure 3 shows the the evidence behavior for the first area. From a maximum value for the lowest model order the evidence decrease and then it has a local maxima for model order 6, which is the selected model best explaining the data. The results in figures 4 and 5 for the second and third areas do not show any local maxima, then a reasonable choice for the data is model order 4 where the evidence becomes stable.

CONCLUSIONS

For SAR SLC data the evidence shows a different behavior from the theoretical one. It decreases increasing the model order complexity. This behavior is due to the noise on the real data. Using the evidence is possible to select the optimal model order which is order six for the metallic structure and model order four for grass landscape and asphalt pavement. Bayesian model selection by evidence maximization can be applied to the space-variant choice of the best model in a given class. We introduced a model based algorithm for texture parameter estimation and space-variant model order selection. Example of feature extraction and classification on local model selection is given on SLC HR SAR image.

REFERENCES

1. A. O'Hagan, Kendall's Advanced Theory of Statistics, volume 2B Bayesian Inference. Arnold, London, 1994.
2. J. Besag, Spatial Interaction and the Statistical Analysis of Lattice Systems. J.Royal Statistical Society B, 36: 192-236, 1974.
3. J. Besag, On the Statistical Analysis of Dirty Pictures. J.Royal Statistical Society B, 48: 259-302, 1986.
4. J. Besag, P. Green, D. Higdon, and K. Mengersen, Bayesian Computation and Stochastic Systems (with discussion). Statistical Science, 10: 3-66, 1995.
5. D.J.C. MacKay, Bayesian Interpolation. Neural Computation, 4(3): 415-447, 1992.
6. E.T. Jaynes, Bayesian methods: general background. In J.H.Justice, editor, Maximum Entropy and Bayesian Methods in Applied Statistics. CUP, 1986.
7. S. Geman and D. Geman, Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(6): 721-741, 1984.
8. H. Akaike, A New Look at the Statistical Model Identification. IEEE Transactions on Automation and Control, 19(6): 716-723, 1974.
9. H. Jeffreys, Theory of Probability. Oxford Univ. Press, 1939.
10. J.W. Goodman, Statistical Properties of Laser Speckle Patterns. In J.C.Dainty, editor, Laser Speckle and Related Phenomena. SpringerÐVerlag, Berlin, 1975.
11. M. Datcu, K. Seidel, and M. Walessa, Spatial Information Retrieval from Remote Sensing Images - Part I: Information Theoretical Perspective. IEEE Trans. Geosci. and R.S., 36(5): 1431-1445, 1998.
12. M. Schröeder, H. Rehrauer, K. Seidel, and M. Datcu. Spatial Information Retrieval from Remote Sensing Images - Part II: Gibbs Markov Fields. IEEE Trans. Geosci. and R.S., 36(5): 1446-1455, 1998.
13. M. Walessa and M. Datcu, Model-based Despeckling and Information Extraction from SAR Images. IEEE Trans. Geosci. and R.S., 38(5): 2258-2269, 2000.
14. V. Pavlovic, R. Sharma, and T. S. Huang, Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7): 677-695, 1997.
15. J. Rissanen, Minimum-Description-Length Principle. In Encyclopedia of Statistical Sciences, volume 5, pages 523-527. Wiley, New York, 1985.
16. S.F. Gull, Bayesian Inductive Inference and Maximum Entropy. In G.J.Erickson and C.R.Smith, editor, Maximum Entropy and Bayesian Methods in Science and Engeneering, volume 1: Foundations. Kluwier, 1988.