The g Factor: Relating Distributions on Features to Distributions on Images

James M. Coughlan and A. L. Yuille
Smith-Kettlewell Eye Research Institute,
2318 Fillmore Street,
San Francisco, CA 94115, USA.
Tel. (415) 345-2146/2144. Fax. (415) 345-8455.
Email coughlan@ski.org, yuille@ski.org

Abstract

We describe the g-factor which relates probability distributions on image features to distributions on the images themselves. The g-factor *depends only on our choice of features and lattice quantization* and is independent of the training image data. We illustrate the importance of the g-factor by analyzing Minimax Entropy Learning (MEL) [8] (which learns image distributions in terms of clique potentials corresponding to feature statistics). We first use our analysis of the g-factor to determine when the MEL clique potentials decouple for different features. Secondly, we show that MEL clique potentials can be computed analytically by approximating the g-factor. We support our analysis by computer simulations.

1 Introduction

There has recently been a lot of interest in learning probability models for vision. The most common approach is to learn histograms of filter responses or, equivalently, to learn *probability distributions on features*. See, for example, [6], [5], [4]. (In this paper the features we are considering will be extracted from the image by filters – hence we almost always use the terms "features" and "filters" synonymously.)

An alternative approach, however, is to learn probability distributions *on the images themselves*. The Minimax Entropy Learning (MEL) theory [8] uses the maximum entropy principle to learn Gibbs distributions in terms of clique potentials determined by the feature statistics. When applied to texture modeling it gives a way to unify the filter based approaches (which are often very effective) with the Gibbs distribution approaches (which are theoretically attractive).

As we describe in this paper, distributions on images and on features can be related by a g-factor (such factors arise in statistical physics, see [3]). Understanding the

g-factor *helps explain why the clique potentials learnt by MEL take the form that they do* as functions of the feature statistics. Moreover, *the MEL clique potentials for different features often seem to be decoupled and the g-factor can explain why, and when, this occurs.* (I.e. the two clique potentials corresponding to two features A and B are identical whether we learn them jointly or independently).

The g-factor is determined only by the form of the features chosen and *the spatial lattice and quantization of the image grey-levels.* It is completely independent of the training image data. It should be stressed that the choice of image lattice and grey-level quantization can make a big difference to the g-factor and hence to the probability distributions which are the output of MEL. Approximations to the g-factor are often best when the quantization is fine.

In Section (2), we briefly review Minimax Entropy Learning. Section (3) introduces the g-factor and determines conditions for when clique potentials are decoupled. In Section (4) we describe a simple approximation which enables us to learn the clique potentials analytically.

2 Minimax Entropy Learning

Suppose we have training image data which we assume has been generated by an (unknown) probability distribution $P_T(\vec{x})$ where $\vec{x}$ represents an image. Minimax Entropy Learning (MEL) [8] approximates $P_T(\vec{x})$ by selecting the distribution with maximum entropy constrained by observed feature statistics $\vec{\phi}(\vec{x}) = \vec{\psi}_{obs}$. This gives $P(\vec{x}|\vec{\lambda}) = \frac{e^{\vec{\lambda}\cdot\vec{\phi}(\vec{x})}}{Z[\vec{\lambda}]}$, where $\vec{\lambda}$ is a parameter chosen such that $\sum_{\mathbf{x}} P(\vec{x}|\lambda)\phi(\vec{x}) = \vec{\psi}_{obs}$. Or equivalently, so that $\frac{\partial \log Z[\vec{\lambda}]}{\partial \vec{\lambda}} = \vec{\psi}_{obs}$.

We will treat the special case where the statistics $\vec{\phi}$ are the histogram of a shift-invariant filter $\{f_i(\vec{x}) : i = 1, ..., N\}$. So $\psi_a = \phi_a(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{a,f_i(\vec{x})}$ where $a = 1, ..., Q$ indicates the (quantized) filter response values. The potentials become $\vec{\lambda}\cdot\vec{\phi}(\vec{x}) = \frac{1}{N}\sum_{a=1}^{Q}\sum_{i=1}^{N}\lambda(a)\delta_{a,f_i(\vec{x})} = \frac{1}{N}\sum_{i=1}^{N}\lambda(f_i(\vec{x}))$. Hence $P(\vec{x}|\vec{\lambda})$ becomes a Gibbs distribution with clique potentials given by $\lambda(f_i(\vec{x}))$. This determines a Markov random field with the clique structure given by the filters $\{f_i\}$.

MEL also has a feature selection stage based on Minimum Entropy to determine which features to use in the Maximum Entropy Principle. The features are evaluated by computing the entropy $-\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}) \log P(\vec{x}|\vec{\lambda})$ for each choice of features (with small entropies being preferred). A *filter pursuit* procedure was described to determine which filters/features should be considered (our approximations work for this also).

3 The g-Factor

This section defines the g-factor in subsection (3.1) and starts investigating its properties in subsection (3.2). In particular, when, and why, do clique potentials decouple? More precisely, when do the potentials for filters A and B learned simultaneously differ from the potentials for the two filters when they are learnt independently?

3.1 Basic Properties of the g-Factor

We now address these issues by introducing the g-factor $g(\vec{\psi})$ and the associated distribution $\hat{P}_0(\vec{\psi})$:

$$g(\vec{\psi}) = \sum_{\vec{x}} \delta_{\vec{\phi}(\vec{x}),\vec{\psi}}, \quad \hat{P}_0(\vec{\psi}) = \frac{1}{L^N} g(\vec{\psi}). \tag{1}$$

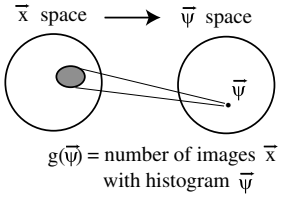

Figure 1: The g-factor $g(\vec{\psi})$ counts the number of images $\vec{x}$ that have statistics $\vec{\psi}$. Note that the g-factor depends only on the choice of filters and is independent of the training image data.

The g-factor is essentially a combinational factor which counts the number of ways that one can obtain statistics $\vec{\psi}$, see figure (1). Equivalently, $\hat{P}_0$ is the default distribution on $\vec{\psi}$ if the images are generated by white noise (ie. completely random images).

We can use the g-factor to compute the induced distribution $\hat{P}(\vec{\psi}|\vec{\lambda})$ on the statistics determined by MEL:

$$\hat{P}(\vec{\psi}|\vec{\lambda}) = \sum_{\vec{x}} \delta_{\vec{\psi},\vec{\phi}(\vec{x})} P(\vec{x}|\vec{\lambda}) = \frac{g(\vec{\psi}) e^{\vec{\lambda}\cdot\vec{\psi}}}{Z[\vec{\lambda}]}, \quad Z[\vec{\lambda}] = \sum_{\vec{\psi}} g(\vec{\psi}) e^{\vec{\lambda}\cdot\vec{\psi}}. \tag{2}$$

Observe that both $\hat{P}(\vec{\psi}|\vec{\lambda})$ and $\log Z[\vec{\lambda}]$ are sufficient for computing the parameters $\vec{\lambda}$. The $\vec{\lambda}$ can be found by solving either of the following two (equivalent) equations:

$$\sum_{\vec{\psi}} \hat{P}(\vec{\psi}|\vec{\lambda})\vec{\psi} = \vec{\psi}_{obs}, \quad or \quad \frac{\partial \log Z[\vec{\lambda}]}{\partial \vec{\lambda}} = \vec{\psi}_{obs}, \tag{3}$$

which shows that *knowledge of the g-factor and* $e^{\vec{\lambda}\cdot\vec{\psi}}$ *are all that is required to do MEL.*

Observe from equation (2) that we have $\hat{P}(\vec{\psi}|\vec{\lambda} = 0) = P_0(\vec{\psi})$. In other words, setting $\vec{\lambda} = 0$ corresponds to a uniform distribution on the images $\vec{x}$.

3.2 Decoupling Filters

We now derive an important property of the minimax entropy approach. As mentioned earlier, it often seems that the potentials for filters A and B decouple. In other words, if one applies MEL to two filters A, B simultaneously by letting

$\vec{\psi} = (\vec{\psi}^A, \vec{\psi}^B)$, $\vec{\lambda} = (\vec{\lambda}^A, \vec{\lambda}^B)$, and $\vec{\psi}_{obs} = (\vec{\psi}^A_{obs}, \vec{\psi}^B_{obs})$, then the solutions $\vec{\lambda}^A, \vec{\lambda}^B$ to the equations:

$$\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}^A, \vec{\lambda}^B)(\vec{\phi}^A(\vec{x}), \vec{\phi}^B(\vec{x})) = (\vec{\psi}^A_{obs}, \vec{\psi}^B_{obs}), \tag{4}$$

are the same (approximately) as the solutions to the equations $\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}^A)\vec{\phi}^A(\vec{x}) = \vec{\psi}^A_{obs}$ and $\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}^B)\vec{\phi}^B(\vec{x}) = \vec{\psi}^B_{obs}$, see figure (2) for a real world example.

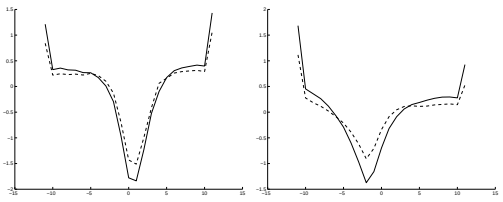

Figure 2: Evidence for decoupling of features. The left and right panels show the clique potentials learnt for the features $\partial/\partial x$ and $\partial/\partial y$ respectively. The solid lines give the potentials when they are learnt individually. The dashed lines show the potentials when they are learnt simultaneously. Figure courtesy of Prof. Xiuwen Liu, USF.

We now show how this decoupling property arises naturally if the g-factor for the two filters factorizes. This factorization, of course, is a property only of the form of the statistics and is *completely independent of whether the statistics of the two filters are dependent for the training data.*

Property I: *Suppose we have two sufficient statistics $\vec{\phi}^A(\vec{x}), \vec{\phi}^B(\vec{x})$ which are independent on the lattice in the sense that $g(\vec{\psi}^A, \vec{\psi}^B) = g^A(\vec{\psi}^A)g^B(\vec{\psi}^B)$, then:*

$$\log Z[\vec{\lambda}^A, \vec{\lambda}^B] = \log Z^A[\vec{\lambda}^A] + \log Z^B[\vec{\lambda}^B], \quad \hat{P}(\vec{\psi}^A, \vec{\psi}^B) = \hat{P}^A(\vec{\psi}^A)\hat{P}^B(\vec{\psi}^B), \tag{5}$$

which implies that the parameters $\vec{\lambda}^A, \vec{\lambda}^B$ can be solved from the independent equations:

$$\frac{\partial \log Z^A[\vec{\lambda}^A]}{\partial \vec{\lambda}^A} = \vec{\psi}^A_{obs}, \quad \frac{\partial \log Z^B[\vec{\lambda}^B]}{\partial \vec{\lambda}^B} = \vec{\psi}^B_{obs} \;\; or \;\; \sum_{\vec{\psi}^A} \hat{P}^A(\vec{\psi}^A)\vec{\psi}^A = \vec{\psi}^A_{obs}, \; \sum_{\vec{\psi}^B} \hat{P}^B(\vec{\psi}^B)\vec{\psi}^B = \vec{\psi}^B_{obs}. \tag{6}$$

Moreover, the resulting distribution $P(\vec{x})$ can be obtained by multiplying the distributions $(1/Z^A)e^{\vec{\lambda}^A \cdot \vec{\psi}^A(\vec{x})}$ and $(1/Z^B)e^{\vec{\lambda}^B \cdot \vec{\psi}^B(\vec{x})}$ together.

The point here is that the potential terms for the two statistics $\vec{\psi}^A, \vec{\psi}^B$ decouple if the phase factor $g(\vec{\psi}^A, \vec{\psi}^B)$ can be factorized. *We conjecture that this is effectively the case for many linear filters used in vision processing.* For example, it is plausible that the g-factor for features $\partial/\partial x$ and $\partial/\partial y$ factorizes – and figure (2) shows that their clique potentials do decouple (approximately). Clearly, if factorization between filters occurs then it gives great simplification to the system.

It may, however, be questioned whether this decoupling is desirable. Recall that this "factorization" is purely a property of the filters and the lattice (plus quantization) and is *completely independent* of the training image data. If the g-factor

factorizes then MEL (using the feature marginals) will imply that $\hat{P}(\vec{\psi}^A, \vec{\psi}^B) = \hat{P}^A(\vec{\psi}^A)\hat{P}^B(\vec{\psi}^B)$ and so will *predict* that the joint histograms $\vec{\psi}^A_{obs}, \vec{\psi}^B_{obs}$ are statistically independent and uncorrelated.

4 Approximating the g-factor for a Single Histogram

We now consider the case where the statistic is a single histogram. Our aim is to understand why features whose histograms are of stereotypical shape give rise to potentials of the form given by figure (2).

Our results, of course, can be directly extended to multiple histograms if the filters decouple, see subsection (3.2). We first describe the approximation and then discuss its relevance for filter pursuit.

We rescale the $\vec{\lambda}$ variables by N so that we have:

$$P(\vec{x}|\lambda) = \frac{e^{N\vec{\lambda}\cdot\vec{\phi}(\vec{x})}}{Z[\vec{\lambda}]}, \quad \hat{P}(\vec{\psi}|\lambda) = g(\vec{\psi})\frac{e^{N\vec{\lambda}\cdot\vec{\psi}}}{Z[\vec{\lambda}]}, \tag{7}$$

We now consider the approximation that the filter responses $\{f_i\}$ are *independent of each other when the images are uniformly distributed*. This is the *multinomial approximation*. (We attempted a related approximation [1] which was less successful). It implies that we can express the phase factor as being proportional to a multinomial distribution:

$$g(\vec{\psi}) = L^N \frac{N!}{(N\psi_1)!...(N\psi_Q)!}\alpha_1^{N\psi_1}...\alpha_Q^{N\psi_Q}, \quad \hat{P}_0(\vec{\psi}) = \frac{N!}{(N\psi_1)!...(N\psi_Q)!}\alpha_1^{N\psi_1}...\alpha_Q^{N\psi_Q}, \tag{8}$$

where $\sum_{a=1}^{Q}\psi_a = 1$ (by definition) and the $\{\alpha_a\}$ are the means of the components $\{\psi_a\}$ with respect to the distribution $\hat{P}_0(\vec{\psi})$. As we will describe later, the $\{\alpha_a\}$ will be determined by the filters $\{f_i\}$. See technical report [2] for details of how to compute the $\{\alpha_a\}$.

This approximation enables us to calculate MEL *analytically*.

Theorem *With the multinomial approximation the log partition function is:*

$$\log Z[\vec{\lambda}] = N\log L + N\log\{\sum_{a=1}^{Q} e^{\lambda_a + \log\alpha_a}\}, \tag{9}$$

and the "potentials" $\{\lambda_a\}$ can be solved in terms of the observed data $\{\psi_{obs,a}\}$ to be:

$$\lambda_a = \log\frac{\psi_{obs,a}}{\alpha_a}, \quad a = 1, ..., Q. \tag{10}$$

We note that there is an ambiguity $\lambda_a \mapsto \lambda_a + K$ where K is an arbitrary number (recall that $\sum_{a=1}^{Q}\psi(a) = 1$). We fix this ambiguity by setting $\vec{\lambda} = 0$ if $\vec{\alpha} = \vec{\psi}_{obs}$.

Proof. *Direct calculation.*

Our simulation results show that this simple approximation gives the typical potential forms generated by Markov Chain Monte Carlo (MCMC) algorithms for

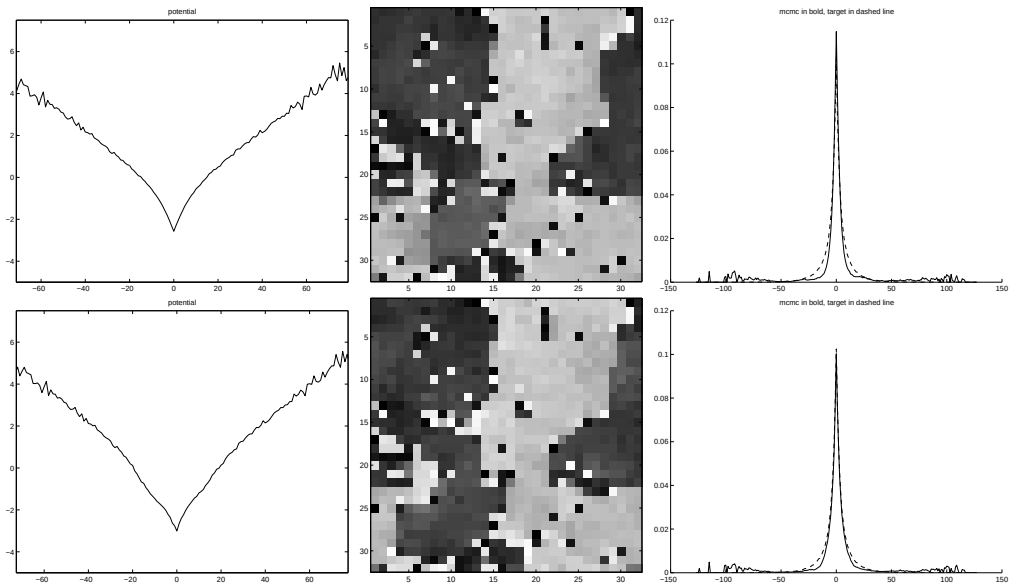

Figure 3: (Top) the multinomial approximation. (Bottom) full MEL. (Left Panel) the potentials, (Centre Panel) synthesized images, and (Right Panel) the difference between the observed histogram (dashed line) and the histogram of the synthesized images (bold line). Filters were d/dx and d/dy.

Minimax Entropy Learning. Compare the multinomial approximation results with those obtained from a full implementation of MEL, see figure (3)

Filter pursuit is required to determine which filters carry most information. MEL [8] prefers filters (statistics) which give rise to low entropy distributions (this is the "Min" part of Minimax). The entropy is given by $H(P) = -\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}) \log P(\vec{x}|\vec{\lambda}) = \log Z[\vec{\lambda}] - \sum_{a=1}^{Q} \lambda_a \psi_a$. For the multinomial approximation this can be computed to be $N \log L - N \sum_{a=1}^{Q} \psi_a \log \frac{\psi_a}{\alpha_a}$.

This gives an intuitive interpretation of feature pursuit: we should prefer filters whose statistical response to the image training data is *as large as possible* from their responses to uniformly distributed images. This is measured by the Kullback-Leibler divergence $\sum_{a=1}^{Q} \psi_a \log \frac{\psi_a}{\alpha_a}$. Recall that if the multinomial approximation is used for multiple filters then we should simply add together the entropies of different filters.

5 Discussion

This paper describes the g-factor which depends on the lattice and quantization and is independent of the training image data. Alternatively it can be thought of as being proportional to the feature responses when the input images are uniformly distributed.

We showed that the g-factor can be used to relate probability distributions on features to distributions on images. In particular, we described approximations

which, when valid, enable MEL to be computed analytically. In addition, we can determine when the clique potentials for features decouple. These approximations throw light on MEL and give guidelines to determine whether marginal histograms should be used as input to MEL (or joint distributions are needed).

Our approach also emphasizes the importance of understanding the feature properties *independent of the dataset* and, in particular, to determine what the feature histograms are when the input images are uniformly distributed. This depends strongly on the quantization procedure used to describe the images. We also point out that the problem of estimating clique potentials may get simpler for fine quantization (because the approximations become more accurate).

Acknowledgements

We thank Anand Rangarajan, Xiuwen Liu, and Song Chun Zhu for helpful conversations. Sabino Ferreira gave useful feedback on the manuscript. This work was supported by the National Institute of Health (NEI) with grant number RO1-EY 12691-01.

References

[1] J.M. Coughlan and A.L. Yuille. "A Phase Space Approach to Minimax Entropy Learning; The Minutemax approximation". In *Proceedings NIPS'98*. 1998.

[2] J.M. Coughlan and A.L. Yuille. "The g Factor: Relating Distributions on Features to Distributions on Images". Technical Report. Smith-Kettlewell Eye Research Institute. San Francisco, CA 94115. 2001.

[3] C. Domb and M.S. Green (Eds). **Phase Transitions and Critical Phenomena**. Vol. 2. Academic Press. London. 1972.

[4] S. M. Konishi, A.L. Yuille, J.M. Coughlan and Song Chun Zhu. "Fundamental Bounds on Edge Detection: An Information Theoretic Evaluation of Different Edge Cues." In *Proceedings Computer Vision and Pattern Recognition CVPR'99*. Fort Collins, Colorado. 1999.

[5] A.B. Lee, D.B. Mumford, and J. Huang. "Occlusion Models of Natural Images: A Statistical Study of a Scale-Invariant Dead Leaf Model". *International Jouranl of Computer Vision*. Vol. 41, No.s 1/2. January/February. 2001.

[6] J. Portilla and E. P. Simoncelli. "Parametric Texture Model based on Joint Statistics of Complex Wavelet Coefficients". *International Journal of Computer Vision*. October, 2000.

[7] Y. Wu, S.C. Zhu, and X. Liu. "Equivalence of Julesz texture ensembles and FRAME models", *International Journal of Computer Vision*, 38(3), 247-265. 2000.

[8] S.C. Zhu, Y. Wu, and D. Mumford. "Minimax Entropy Principle and Its Application to Texture Modeling". Neural Computation. Vol. 9. no. 8. Nov. 1997.