

Research Statement

Bruno Jedynak

1. OVERVIEW

Pattern Theory was initiated by Ulf Grenander about thirty years ago. The aim is to analyze patterns from a statistical point of view in all “signals” generated by the world, whether they be visual, acoustical, textual, molecular (e.g., DNA strings), neural, etc. Patterns are described using hidden variables, together with their probability distributions, whereas signals, or relevant functions of the signals, are modeled conditionally on the hidden variables. In principle, the detection of patterns in noisy and ambiguous samples can then be achieved by the use of Bayes’s rule. An overview of pattern theory as a mathematical theory of perception was presented during the International Congress of Mathematics in 2002; see [13].

There are enormous difficulties in realizing the pattern theory program. Initially, I was inspired by problems arising in computer vision where the signal is a still image and the pattern is an object. This research is described in Section 2. Recently, I have diversified my research to include applications in natural language modeling and bio-informatics. This work was motivated by questions concerning statistical modeling in the small sample situation and in particular the technique of maximum entropy on the mean. This work is described in section 3.

2. BAYESIAN MODELING AND INFORMATION THEORY IN VISION

Vision offers an overwhelmingly rich source of challenging questions. How can a system identify an object in an image in the presence of clutter and occlusion? Locally, the existence of the object is always ambiguous, whereas globally it is unambiguous. Humans identify faces and skin very efficiently from still images. A trained radiologist can precisely identify brain structures from magnetic resonance images. Can one design efficient computer vision systems that replicate these capabilities ?

In each situation, one starts with data and seeks concepts in the sense of interpretations. It is then necessary to use a quantitative measure of information that can be applied to a large class of objects. Shannon information theory provides some of the theoretical foundations. However, the classical goals of information theory – coding and compression – are not exactly the same as in computer vision. Also, there are many interesting connections between statistics and information theory. One of particular interest and simplicity is related to the problem of Bayesian classification.

In Bayesian classification, there is a finite set of objects or interpretations of interest, denoted \mathcal{Y} . The data, often multidimensional, lives in a set \mathcal{X} . Assuming a suitable probability structure and a binary cost function, the best guess for the object, having observed a data point $x \in \mathcal{X}$, is $\hat{Y}(x)$, the mode of the posterior probability $P(Y|X = x)$. Now the expected error of this classifier, denoted e^* , is very closely related to the conditional entropy $H(Y|X)$ (say in base 2) which measures the expected amount of information that X provides about Y . On one hand, elementary calculus provides $e^* \leq \ln(2)H(Y|X)$; and, on the other, Fano’s inequality yields a reciprocal bound: $H(Y|X) \leq H(e^*) + e^* \log(|\mathcal{Y}| - 1)$. The situation is similar for regression.

In the applications outlined previously, the objects are roads, faces, anatomical landmarks and so forth, and the data consist of images. The Bayesian classifier is in general not available. First, the joint distribution of (X, Y) is not known, and, even if it were, the computation of the posterior would be extremely challenging or impossible. However, the equalities above suggest a line of research consisting of building classifiers that serve to iteratively reduce the conditional entropy $H(Y|X)$ or a suitable estimate thereof. Notice that this approach does not necessarily require one to model the full dependency structure of (X, Y) , but rather to successively concentrate on important components, addressing at the same time the problems of statistical modeling and computational efficiency.

My early research along these lines includes work on decision and classification trees, [2]. With my collaborators, we have extended the methodology to settings in which it was originally not applicable, e.g., road tracking and face detection. In a more theoretical paper, we analyzed the tradeoffs between global and greedy procedures for conditional entropy reduction.

2.1. Road Tracking. This work was done while I was a PhD student under the supervision of Donald Geman.



FIGURE 1. An algorithm for tracking roads

Road tracking consists of identifying a road in a remotely sensed image, starting with a pixel on the road in the image and a direction, both manually selected. \mathcal{Y} is a finite but very large ($\approx 3^{100}$) discretization of the set of roads organized as a ternary tree. \mathcal{X} is a product space. Each component is indexed by a location and an orientation in the image, whose value is obtained by applying a given real-valued filter at that location. The number of components is so large that the likelihood, even without the normalization constant, cannot be evaluated. We proceed by selecting components, one at a time, in a sequential and adaptive manner, in order to reduce as much possible the conditional entropy $H(Y|X)$ given the results of the filters already evaluated.

Whereas sequential reduction of entropy is shared with the methodology in decision and regression trees, here, in sharp contrast, the computation is on-line versus off-line; as a consequence, a single branch is computed, the one corresponding to the branch followed by the image at hand; see [3].

2.2. Outlier Detection and Asymptotic Properties of the Road Tracking Algorithm. This work was initiated during my PhD Thesis and continued at JHU in collaboration with Damianos Karakos.

Our motivation for this paper originates in the work on road tracking described above. Below a certain clutter level, that algorithm could track a road accurately, but suddenly, with increased clutter, tracking would become impossible.

We consider the problem of detecting a target in the presence of background clutter. We study an ultra-simplified model, introduced in [14], where a phenomenon of phase transition is observed: there are $M + 1$ sequences of independent discrete random variables, each sequence being of length N , and all sequences have components with the same probability mass function p_0 except for one sequence, the target, whose elements have probability mass function p_1 . We focus on asymptotic bounds of performance, and we show that the error of the maximum likelihood estimator for the target converges to 0 or to 1, depending on the behavior of the fundamental quantity $M2^{-N D(p_1, p_0)}$, where $D(., .)$ is the Kulback-Leibler divergence. Moreover, we describe a target detector for the case where p_0 and p_1 are unknown, and we prove that it has the same phase transition behavior as in the case of known distributions. See [10].

2.3. Face Detection. This work was done during my post-doc at the University of Chicago in collaboration with Yali Amit.

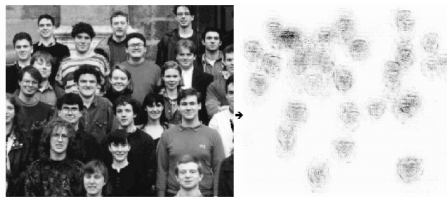


FIGURE 2. Face detection. The amount of processing as a function of the location in the image

Face detection consists in identifying the locations of the faces, if any, in an image. It is a necessary step for performing face recognition from unconstrained images. Here the class variable takes only two values corresponding to the presence or absence of a face in a sub-image. This apparent simplicity hides a complex mixture of situations when a face is present, corresponding to instances of pose, identity and lighting, not to mention the enormous variations in the nature of the cluttered background. It is actually surprising that any statistically meaningful performance could be achieved. Detection is done in two stages: (i) “focusing”, during which a relatively small number of regions-of-interest are identified, minimizing computation and false negatives at the temporary expense of false positives; and (2) “intensive classification”, during which a selected region-of-interest is labeled face or background based on multiple decision trees and normalized data. In contrast to most detection algorithms, the processing is then very highly concentrated in the regions near faces and near false positives, as can be seen in Figure 2.3. See [1]. Unfortunately, such a computational design does not emerge naturally from greedy entropy. We studied this phenomenon in a more general context as described in the following subsection.

2.4. Global vs Greedy Procedures for Entropy Reduction. This work was done while I was an assistant professor at USTL, Lille, France, in collaboration with Donald Geman.

The construction of classification trees is nearly always top-down, locally optimal, and data-driven. Such recursive designs are often globally inefficient, for instance, in terms of

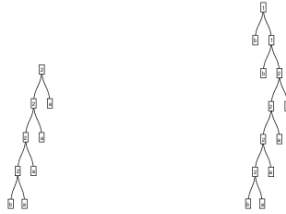


FIGURE 3. Left: Locally optimal tree. Right: Globally optimal tree. The error rates are the same but the *mean* depth of the global tree is smaller.

the mean depth necessary to reach a given classification rate. We consider statistical models for which exact global optimization is feasible, using dynamic programming, and thereby demonstrate that recursive and global procedure may result in very different tree graphs and overall performance. Here is a toy example that was motivated by the work on face detection. There are two classes. One noted a is “object” and the other noted b is “background”, with prior probabilities $p(a) = 10^{-4}$ and $p(b) = 1 - 10^{-4}$. There are two types of tests, X_1 and X_2 . X_1 has a 0 false positive rate, i.e., keeps all the background together, but has false negative rate 0.5. X_2 has a 0 false negative rate, i.e., loses no objects, but 0.5 false positive rate. These tests are assumed to be repeatable, the sequence of outcomes being independent conditional on the class. Figure 2.4 shows the tree obtained by the greedy entropy reduction as well as a globally optimal tree with maximum depth 6. The error rate for these trees are approximatively the same but the mean depth is about 4 for the greedy one, and about 2.5 for the optimal one. At the same mean depth, the optimal strategy may have an error rate ten times smaller than the greedy strategy. See [4].

2.5. Automatic Landmark Detection from Brain MRI. This work is being conducted at JHU in collaboration with Camille Izard, a PhD Student under my supervision from USTL.

An anatomical landmark in the brain is a well-defined point of the anatomy of the brain. Locating a landmark in a magnetic resonance brain image, or “landmarking,” consists of selecting a particular voxel in the image, corresponding to the anatomical landmark in the imaged brain. This voxel, like an anchor, is a precious piece of information for measuring and registering brain structures. Landmarking can be a tedious manual procedure, expensive and time consuming. It might be error prone, difficult to assess, and dependent on the scanner and on the landmarker. We have developed a generic algorithm that permits one to partially automate the landmarking process. The algorithm has two components. One is an off-line procedure, the other is on-line. The former is a system that estimates the parameters of a probabilistic model from a training set of landmarked images using the Estimation Maximization (EM) algorithm. The later inputs an image as well as the parameters previously estimated and outputs a tentative location for the landmark as well as a covariance metric that assesses the remaining uncertainty. The selected location can then be validated or corrected manually. The probabilistic model has two components corresponding to photometry and the geometry. The former is a mixture of Gaussian distributions whereas the later is a probabilistic model over sets of deformations. We have considered various classes of affine deformations and currently are experimenting with small nonlinear deformations using kernels.

An instantiation of the method for detecting the apex of the Head of the Hippocampus (HoH) is shown in the figure. See [9][7][8]

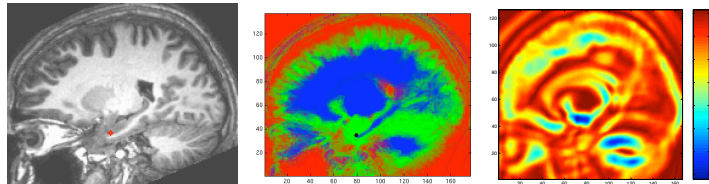


FIGURE 4. **Left:** A sagittal slice of the brain. The Apex of the head of the Hippocampus (HoH) is shown in red. **Center:** Probabilistic model predicting the probability for matter type given the location of the HoH. Red channel : cerebrospinal fluid, Green channel : grey matter, Blue channel : white matter. **Right :** Expected variance reduction in localizing the HoH according to the learned probabilistic model. Most informative voxels are in blue, least informative voxels are in red

3. MAXIMUM ENTROPY MODELING AND SMALL SAMPLE STATISTICS

Maximum Entropy Modeling is a statistical modeling methodology aimed at selecting a probability distribution given a data set. It is a two-step procedure. In the first step, one chooses a subset of probability distribution that is consistent with the data. Typically, one constrains the mean of certain functions of the data, also called features, to agree with the empirical mean derived from the data at hand. In the second step, one chooses a *reference* probability distribution or positive measure. Then, the “closest” distribution to the reference, within the subset defined in the first step, is selected. For example, if the reference is the Uniform measure and the Kulback-Leibler distance is used to define “closest”, than this amounts to selecting the distribution with maximum entropy.

The whole procedure might be viewed as an alternative to Bayesian modeling since one is not obliged to choose a whole prior over a set of distributions but rather a single element, the reference, together with a set of features.

This method was pioneered in statistical mechanics where the object of study is a very large set of interacting particles. The “microstate” is defined as the collection of the states of the particles. It is to be modeled. However, the set of microstates is so large that it cannot be directly modeled from observing a few instances. Alternatively, one has access to “macrostates”. These are quantities that are averaged over the set of particles. Choosing the maximum entropy model among those which replicate the observed macrostate values leads to the important class of Gibbs models, or Markov Random Fields.

This approach was shown to be of great practical importance in low level imaging in [5]. More recently, the use of large sets of natural images has led, using MEM, to the construction of models for textures [16].

3.1. Models for the Texture of Skin. This work was done in collaboration with Mohamed Daoudi and Huicheng Zheng at USTL, while Mohamed and I were co-supervising Huicheng’s PhD thesis.



FIGURE 5. Three models for the classification of skin pixels

In order to classify pixels as “skin” or “non skin”, we have experimented with Maximum Entropy Modeling with tree approximations to Markov Random Fields. Indeed, when the underlying graph is a tree, the optimization procedure required to estimate the parameters of the model can be tackled by an efficient procedure, already used in natural language processing, known as iterative scaling. Moreover, classification of pixels as *skin* or *not skin* is achieved through an efficient combinatorial optimization procedure, closely related to dynamic programming and known as belief propagation. We build a sequence of three models by adding features one at a time. The observed statistics come from a collection of hand-segmented images. The first model imposes constraints on one-pixel color histograms given “skin” and given “non-skin”. The solution is a baseline model in which pixels are considered independent. This model is well-known among practitioners. The baseline model is certainly too weak and does not take into account the fact that skin zones are made of large regions with regular shapes. Hence, in the second model, we add constraints on the distribution of neighboring labels (skin or not-skin) in order to smooth the solution. Finally, a color gradient is included in building the third model. Figure 3.1 depicts examples of the resulting segmentations. The color is proportional to the posterior probability for skin. State-of-the-art performance is reported. See[15] [12].

3.2. MEM in the Small Sample Setting and Language Modeling. This work was done at Johns Hopkins University, in collaboration with Sanjeev Khudanpur and Ali Yazgan from the Center for Language and speech Processing.

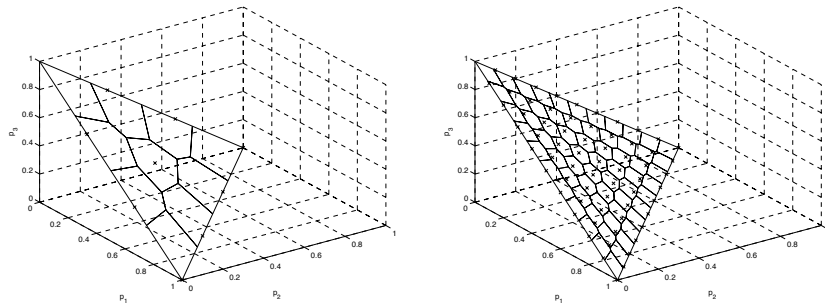


FIGURE 6. Maximum Likelihood Sets for $k = 3$ outcomes. Left: $n = 3$ observations. Right: $n = 10$ observations.

There are challenging applications in statistics where the number of samples is small compared to the dimensionality of the data. If one wants to adapt the MEM approach to these situations, one has to take into account the natural variability of the empirical mean of the features around their expectation in order to define a set of distributions consistent with the data. How can this be done in a systematic way? An example arises in natural language modeling where one needs to define the probability for the next word in a sequence. Even,

more basically, one needs to estimate the probability of appearance of a word, independently of the past words. Assume there are $k = 50,000$ English words in the dictionary and a corpus of size $n = 1,000,000$ from the Wall Street Journal. Typically, 13,000 words are not present in the corpus and 13,000 are seen only once. This is a small sample situation. When estimating conditional distributions, the small sample effect is even more severe. The simplest features in this context are the indicator functions for a presence of a word. There are k such features. However, the set of distributions over words that replicate the observed frequencies for each word is reduced to a single distribution – the empirical measure, or *type* which put zero mass on about 1/4 of the vocabulary. We propose a parameter-free method to release the hard constraint on the word frequencies. We choose the set of distributions on words that make the observed frequencies more likely than any other with the same sample size. This defines a closed convex polyhedron in the space of distributions that we call the Maximum Likelihood Set; see figure 3.2. We then choose the one in this set closest in Kulback-Leibler divergence to the Zipf distribution, the natural prior in this context. The obtained estimator is shown to be competitive with state-of-the art methods, see [11].

3.3. Ongoing work in Bioinformatics. This work is conduct at JHU in collaboration with Joel Bader and Haliang Huang from the Institute of Computational Medicine.

Despite progress in high-throughput protein interaction screens, the number of protein-protein interactions in human and model organisms remains uncertain. Coverage estimates rely primarily on the overlap between disparate data sets. For direct protein-protein interactions identified by two-hybrid screens, we present a novel computational method for generating an intrinsic estimate of the network complexity including the degree distribution and the number of interaction partners per bait. The mathematical model is as follows: for each bait, there is an unknown list of size k of interacting protein . An experiment consists in sampling n times *with replacement* from this list. The goal is to infer k . The main difficulty come from the fact that for most experiments n is small. There is another complication: the identity of the protein is sometimes not recorded properly. We show that a sufficient statistic for k is the vector v made of the counts of the counts: the number of interacting protein that appeared once, that appears twice, etc... We use a bayesian approach, assuming a prior distribution for k with parameters that need to be estimated. We are able to write the conditional distribution of v given k and use the EM algorithm across the set of experiments to estimate the parameters of the prior. The results are surprisingly good as demonstrated with simulated dataset. Finally, for each experiment, k is estimated using the posterior mean. See [6].

REFERENCES

- [1] Y. Amit, D. Geman, and B. Jedynek. *Face Recognition: From Theory to Applications*, chapter Efficient focusing and face detection. H. Wechsler and J. Phillips, editors, NATO ASI Series F. Springer-Verlag, Berlin, 1998.
- [2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.
- [3] D. Geman and B. Jedynek. An active testing model for tracking roads in satellite images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(1):1–14, 1996.
- [4] D. Geman and B. Jedynek. Model-based classification trees. *IEEE Transactions on Information Theory*, 47(3):1075–1082, 2001.

- [5] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. on PAMI*, 6:721–741, 1984.
- [6] H. Huang, B. Jedynek, and J.S. Bader. Where have all the interactions gone ? estimating the coverage of two-hybrid protein interaction maps. Submitted to PLoS Computational Biology, 2006.
- [7] C. Izard and B. Jedynek. Bayesian registration for anatomical landmark detection. In *IEEE International Symposium on Biomedical Imaging*, Arlington, VA, April 2006. Working paper.
- [8] C. Izard and B. Jedynek. Spline-based probabilistic model for anatomical landmark detection. In *MIC-CAI*, Copenhagen, October 2006.
- [9] C. Izard, B. Jedynek, and C. Stark. Automatic landmarking of magnetic resonance brain images. In Joseph M. Reinhardt; Eds J. Michael Fitzpatrick, editor, *Proc. SPIE Medical Imaging 2005: Image Processing*, volume 5747, pages 1329–1340, 2005.
- [10] B. Jedynek and D. Karakos. Finding a needle in a haystack: Conditons for reliable detection in the presence of clutter. In revision for "Statistics and Probability Letters", January 2006.
- [11] B. Jedynek and S. Khudanpur. Maximum likelihood set for estimating a probability mass function. *Neural Computation*, 17(7):1508 – 1530, July 2005.
- [12] B. Jedynek, H. Zheng, and M. Daoudi. Skin detection using pairwise models. *Image and Vision Computing*, 23(13):1122–1130, November 2005.
- [13] D. Mumford. Pattern theory: the mathematics of perception. In *ICM*, 2002. <http://www.dam.brown.edu/people/mumford/Papers/ICM02proceedings.pdf>.
- [14] A. L. Yuille and J. M. Coughlan. Fundamental limits of bayesian inference: Order parameters and phase transitions for road tracking. *IEEE Transactions on PAMI*, 22(2):160–173, February 2000.
- [15] H. Zheng, M. Daoudi, and B. Jedynek. Blocking adult images based on statistical skin detection. *Electronic Letters on Computer Vision and Image Understanding*, 4(2):1–14, 2004.
- [16] S.C. Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (frame): towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.