MODELING NATURAL MICROIMAGE STATISTICS

A Dissertation Presented

by

ALEXEY A. KOLOYDENKO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2000

Department of Mathematics and Statistics

MODELING NATURAL MICROIMAGE STATISTICS

A Dissertation Presented

by

ALEXEY A. KOLOYDENKO

Approved as to style and content by:

Donald Geman, Chair

Joseph Horowitz, Member

Christopher Raphael, Member

Edward Riseman, Member

Donald St. Mary, Department Head
Mathematics and Statistics

ACKNOWLEDGMENTS

This work concludes my rather long process of formal education. My family, and especially my parents, have greatly contributed to the development of my motivation to study natural sciences.

I am grateful to my family and my friends for their support in all my endeavors and will especially remember the help I received from them at the most trying times of my dissertation work.

My gratitude extends to all of my prior teachers for kindly sharing with me their knowledge and experience. I remember them dearly and at times regret not always following their advice.

It has been a privilege to learn from and to work with Professor Donald Geman, a great teacher and my thesis advisor. His courses on Probability Theory and Mathematical Image Analysis helped me realize my own research interests. I am grateful to Professor Geman for helping me find applications of my scientific aspirations in the fields of probabilistic modeling and applied statistics. Working with him has been one of the most enjoyable parts of my learning experience in this Department.

I sincerely thank my committee members, Professors Joseph Horowitz, Christopher Raphael, and Edward Riseman for their interest in and evaluation of my work and for their efforts to help me improve my thesis. Explanations of mathematical and statistical concepts as well as general comments and suggestions that I received from Professors John Buonaccorsi, Eduardo Cattani, John Fogarty, Melvin Janowitz, Ivan Mirkovic, Professors Emeritus David Foulis and Berthold Schweizer at various stages of my dissertation work have profoundly influenced my way of thinking about this work and are very much appreciated.

I also learned very much from courses taught by Professors George Avrunin, Eduardo Cattani, David Foulis, Joseph Horowitz, Markos Katsoulakis, Ivan Mirkovic, Walter Rosenkrantz, and Bruce Turkington.

I have enjoyed and benefited from discussions of my work with Gilda Garibotti and Javier Fernandez, for which I thank them sincerely. My special thanks are to Volker Ecke for his enormous help with computer-related, mathematical, and other issues of my work. I am also grateful to Chunming Li, Dmitry Rumynin, Mark Stowell, and Kenneth Wilder for their help with various aspects of my work.

I also appreciate scientific communication in the form of seminars and informal discussions with members of the Computer Vision, Robotics, and Intelligent Information Retrieval Laboratories of the University of Massachusetts at Amherst and with our colleagues from Brown University.

Many faculty and staff members of this Department as well as my fellow graduate students have made my work and learning process enjoyable and productive. Professor and Department Head Donald St. Mary, Professors and Graduate Program Directors Eduardo Cattani and James Humphreys, as well as Professors Richard Ellis, George Knightly, and Nathaniel Whitaker deserve high praise for their dedication to the Department and its graduate students.

I am grateful to the Department for the opportunity to teach calculus courses, especially that chaired by Professor Emeritus Samuel Holland.

ABSTRACT

MODELING NATURAL MICROIMAGE STATISTICS

SEPTEMBER 2000

ALEXEY A. KOLOYDENKO, B.S., NORWICH UNIVERSITY
AND VORONEZH UNIVERSITY
M.S., VORONEZH UNIVERSITY AND UNIVERSITY OF
MASSACHUSETTS AMHERST
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST
Directed by: Professor Donald Geman

A large collection of digital images of natural scenes provides a database for analyzing and modeling small scene patches (e.g., 2×2) referred to as natural microimages. A pivotal finding is the stability of the empirical microimage distribution across scene samples and with respect to scaling. With a view toward potential applications (e.g. classification, clutter modeling, segmentation), we present a hierarchy of microimage probability models which capture essential local image statistics. Tools from information theory, algebraic geometry and of course statistical hypothesis testing are employed to assess the "match" between candidate models and the empirical distribution. Geometric symmetries play a key role in the model selection process.

One central result is that the microimage distribution exhibits reflection and rotation symmetry and is well-represented by a Gibbs law with only pairwise interactions. However, the acceptance of the up-down reflection symmetry hypothesis is borderline and intensity inversion symmetry is rejected. Finally, possible extensions to larger patches via entropy maximization and to patch classification via vector quantization are briefly discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOTATION

Σ_{im} - Natural image population

Σ - Natural microimage population

$(\Omega_{im}, \mathbb{P}_{im})$ - Natural image probability space

$\mathbb{P}_{im}(I)$ - Probability of an image $I \in \Omega_{im}$

(Ω, p) - Probability space of natural microimages

$p(\omega)$ - Probability of a microimage $\omega \in \Omega$

$M_{m\times n}(\mathcal{C})$ - $m \times n$ - Matrices over some set $\mathcal{C}$

Θ - Finite dimensional probability simplex

Q^+ - Positive quadrant of a real vector space

$\Theta^+ = \Theta \cap Q^+$ - Set of all strictly positive probability vectors

Θ_0 - A model-dependent subset of Θ

δ_{ij} - Kronecker's symbol, i.e. $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise

$H(q)$ - (Shannon's) entropy of a discrete distribution q

$H(X)$ - Entropy of a discrete random variate X

$H(X|B)$ - Conditional entropy of X given an event B

$H(X|Y = y)$ - Entropy of X conditioned on the event $\{Y = y\}$

$\mathbb{E}_q X$ - Expected value of X under some probability distribution q

$var(X) = \sigma_X^2$ - Variance of X

$H(X|Y)$ - Entropy of X conditioned on Y: $\mathbb{E}\phi(Y)$ $(\phi(y) = H(X|Y = y))$

$D(q_1, q_2)$ - Kullback-Leibler divergence *from* q_1 *to* q_2

$|A|$ - Cardinality of set A

$\mathbb{I}_A$ - Indicator function of set A

G - A group

P_L^l - The number of permutations of L elements taken l at a time.

$\langle g_1, \ldots, g_n | r_1(g_1, \ldots, g_n), \ldots, r_m(g_1, \ldots, g_n)\rangle$ - A group generated by $g_1, \ldots, g_n$ with relations $r_1, \ldots, r_m$

$\mathcal{S} = W/\sim$ - Quotient set of W under the equivalence relation $\sim$

$\mathcal{O} \in \mathcal{S}$ - An individual equivalence class; an *orbit* in the *group action* context

$J \subset \mathcal{R}$ - An ideal of a ring

$\mathbb{R}[x_1, \ldots, x_n]$ - Set (*ring* and *algebra*) of polynomials in n indeterminates with real coefficients

$\mathbb{R}[x_1, \ldots, x_n]^G$ - Set (*ring* and *algebra*) of polynomials in n indeterminates with real coefficients, that are invariant under a group G action

CHAPTER 1

INTRODUCTION

With the emergence and fast development of computer technologies, the word "image" has nowadays acquired yet another common meaning. Thus, an image, or more precisely, a digital image refers to a representation of visual information as a discrete array of numbers. Digital imagery deeply penetrates modern human activities, ranging from medicine to warfare. To one degree or another, many of us are involved in digital image processing, transmission, and storage on almost a daily basis; think, for example, of sending electronically a recent photograph to friends or enhancing and storing a family photo-album on a PC.

Although behind each digitally stored photographic image there stands a numerical matrix, one does not perceive images directly through numbers. In fact, what we see with our eyes is a result of physical processing (via an electronic display, for example) of the numerical representation. This representation also becomes indispensable if one wants to automate a certain procedure to transform one image into another. If, for instance, the goal were to brighten a photograph, this could typically be achieved by shifting the numerical values of all or some of the image *pixels* (a common term for the image matrix entries) to the "brighter" side of the intensity range. Among the central goals of *Image Analysis* and *Image Processing* is the search for and application of mathematical operations on the numerical image representations in order to achieve certain desirable effects, e.g. perceptual quality enhancement and visibility of object boundaries.

Particular studies may focus on images of specific origin or from a specific domain. Thus, whereas digitized brain scans obtained by means of magnetic resonance (*MRI*) are of central interest to researchers automating the process of brain tumor identification and measurement, forensic image analysts often deal with images representing human fingerprints. Recognition of cluttered military targets might involve *infrared images* of particular geographic locations, whereas industrial quality control labs often analyze laser images of integrated circuit boards.

Independently of the domain and image origin, some features are common to almost all digital imagery. In a sense, one can talk about "generic images". A simple example of such a feature is the enormous size of a typical representation space. This leads to rather specific mathematical challenges in many fields of image analysis. For instance, at the abstract level mathematics provides a very clean and well-developed theory of operators (here, mappings from the image space to itself) but this theory encounters significant difficulties when in practice the priority of the research shifts from mathematical elegance to computational feasibility.

"Natural images", such as digital representations of visible light photographs of 3D scenes, correspond to a very sparse set of visually meaningful matrices. Such matrices are extremely "rare" among all possible ones, i.e. they occupy only a tiny portion of the appropriate representation space. So, if a numerical matrix were selected "at random" and displayed on the computer screen, one would almost surely not see any meaningful "picture" corresponding to that matrix. To maintain such compactness of their family, natural images must possess a great deal of internal *structure*. In particular, micro parts of images typically appear in a relatively small number of possible arrangements (described by mathematical relations), by far smaller than the total number possible in an arbitrary sub-matrix of the image space. Whereas the number and precise characterizations of such relations may be domain dependent, one may argue that an analysis of the generic situation is still sensible.

In this context, the work presented is an attempt to analyze, model, and explain microscopic portions of a large class of digital images. The approach is largely based on probabilistic reasoning. Indeed, studying the relative frequency of occurrence of images and image functions is now widely recognized as a natural and powerful approach to most problems in image analysis and modeling.

1.1 Related Research and Motivation

The availability of large data sets of digital images [30],[35] and direct sampling from the world wide web makes it possible to collect various image statistics relevant to large classes of imagery. This has significantly contributed to the flourishing of the field of "natural image statistics", pioneered in [19] and [59] and other work. Our work is partially motivated by the key findings in this area, perhaps the most important of which is the scale-invariance of many natural image statistics [9],[23],[34],[35],[50],[58]. Besides its theoretical influence on image modeling (best documented, perhaps, in [50]), statistical scale-invariance has other, somewhat more concrete, applications. For instance, it explains the success of *fractal* or *wavelet* image compression based on multiresolution block coding in the domains of natural imagery [60]. We also verify some modes of scale invariance in our image data, though compression is not among our ultimate goals.

Among image statistics commonly reported to exhibit strong scale invariance are responses of mostly local linear image filters, e.g., directional derivatives, coefficients of Haar wavelets and Gabor filters. Based on very large data sets, recent studies of natural images with fine intensity ranges have, in addition to scale-invariance, also revealed such important properties of natural images as various symmetries of two pixel differencing operator (discrete derivative) and certain departures from normality in multivariate Haar coefficient statistics [34],[35]. Our work, while similar in spirit, shifts the emphasis to the analysis and modeling of the multivariate distribution of raw intensities (i.e. empirical distributions on microscopic image patches) as opposed to the statistics of linear filter responses. This is performed under some restrictions (necessitated by relatively small sample sizes), strongest among which is coarsening of the originally fine intensity range (typically 256-level) to eight levels and sometimes even four levels. Although the coarsening, or *quantization*, can also be viewed as filtering, it is essentially different from the types of filtering referred to above: Coarsening does not alter the patch geometry. Moreover, we will generally use the "most unbiased" such operator, i.e. separate, uniform quantization of the intensity range of each variable. At any rate, we never degenerate to the binary case, and indeed one of our goals has been to extend some ideas from binary image analysis.

In that regard, the original motivation for this work was quite different from the pure phenomenology of natural image statistics. It derives from using basic information-theoretic principles to code 5×5 patches of binary images of handwritten digits [2],[66]. A highly non-uniform empirical patch distribution was induced from a large sample of binary patches extracted from training data. The authors then took advantage of high redundancy in the pixel information under the empirical patch distribution and efficiently vector-quantized (using decision trees) the set of 2^{25} binary matrices by sequential application of the "divide-and-conquer" (maximization of information gain) strategy. Their method used single pixel indicators as underlying primitive operations, or *elementary tests*, corresponding to the tree nodes. The paths of the resulting tree then served as local features for recognition. In this context, our present work has been driven by the idea of finding stable, elementary features, but now for grey scale images. Work on face detection [1] has served as an example of potential applications for such features. Chapter 2 documents some of our early experiments in this direction.

The principle of *Maximum Entropy Extension* (MEE) or, simply, *Entropy Maximization* will often appear in this work. Borrowed from the statistical physics, MEE has now been extensively applied to probabilistic modeling. This principle is also central to statistical learning and has been successfully used in vision-related learning in particular. Some early applications (e.g. [47]) have gradually developed into a powerful learning paradigm - *Minimax Entropy* learning theory [68],[69],[70].

1.2 Objectives and Main Contributions

The essence of this work is a "direct" analysis and modeling of a large class of multivariate probability distributions carried by tiny (e.g. 2×2, 3×3) image blocks, or *microimages*. The analysis is direct because it is not mediated by special purpose filters. Despite its seemingly overwhelming complexity and variability, the microworld of digital images of "natural scenes" does in fact appear to induce rather universal probability distributions. This is by no means obvious a priori and to our knowledge little has been reported in this venue. Imagine for example two distinct domains of digital imagery: vast landscapes and densely populated sites. Macroscopic image attributes (say, those involving objects and measurements of the same order of magnitude as the image size) immediately captivate our attention and are therefore responsible for the formation of a semantic interpretation of the image in our brain: In the first case we might see “fields”, “trees” , “lakes”, “animals”, and so on, and in the second case our eyes register “human faces”, “streets”, “buildings”, etc. The more dramatic is the semantic difference between the two, the less obvious it is that any similarities can exist between them. Yet, the human vision system switches between the two contexts incredibly fast. Must not there then be generic, reusable tools and resources at the lowest levels of visual processing in order to maintain such efficiency? If so, then a representation of basic image constituents must exist which would unify all images of natural scenes regardless of their domain and origin. It has long been known that the natural processing of 2D signals starts with intensive sampling of small image areas. Thus, it is natural to think of the basic image constituents as elements of a virtually infinite population of small blocks or matrices of intensity values. This population of microimages is at the core of our work. The natural framework here would necessarily involve probabilistic concepts. Thus we aim at discovering probabilistic similarities in the micro configurations of images representing diverse domains.

That such similarities indeed exist can be directly verified when, armed with our “magnifying glass”, we zoom in to view millions of tiny arrangements of image pixels. This was essentially the prelude to the present work, which has been driven by our quest for “structure” in the vagaries of miniature image patches. Among the first and easy findings of such an excavation is the fully expected one that, at the microscopic level, all natural images are non-uniformly distributed and, in particular, are dominated by “flat regions”, or “background”. Perhaps less obvious is to find that configurations conforming with our intuitive notion of “edge” are next on the list of the most frequent ones. However, before proposing distributions that would replicate these and other characteristics observed in experiments, one needs to be sure that those characteristics are in fact stable, for example, across various scenes.

We restrict our universe to the totality of digital imagery stored on the world wide web. This defines (still rather vaguely) what we mean by images of "natural scenes". Of course, we must cope with the vagueness of the very notion of "natural scene". To preempt (at least, partially) potential criticism of our broad use of the term, we state as an assumption (supported by our observations) that "unnatural scenes", i.e. meaningful only to the eyes of a very specific category of viewers, are very rare. Hence, in this work we will think of the degree of being "natural" as proportional to the frequency of occurrence of the scene (more formally, the exact numerical array representing a particular view of the scene) in the public domain of the web. Note that this viewpoint is in a perfect agreement with the one that "natural images" are digitized visible light photographs of 3D scenes: The latter do presently dominate the web.

Next, this work delivers a positive answer to the following basic question: "Does the population of digital microimages of natural scenes have a sensible probability distribution?" In the course of our studies we explain what we mean by "sensible" and perform numerous statistical experiments to support our claim, as well as to illustrate typical difficulties one should expect in further exploration of the subject.

Since its early days and until now, image analysis has naturally been dominated by vector space-based approaches and statistical image analysis is no exception in this regard ([5],[30],[43],[44],[48],[51],[60],[68],[69],[70] to name a few). This status-quo may lead to skepticism about the utility of alternative approaches, in particular, those not involving linear filtering. We hope to refute the perception that models built directly on the space of raw image intensities are of limited value and especially sample-dependent due to the immense variability and complexity of the grey scale microimages. What is indeed unrealistic due to variability (when the sample size is fixed) is accurate estimation of rare point-masses under fine quantizations. However, we also show that the microimage probability laws exhibit some remarkable properties (in the form of mathematical relations) that can already be captured with relatively small samples. Naturally, the sample size determines an appropriate degree of aggregation (quantization) at which these properties exhibit sufficient statistical significance. We find aggregations appropriate for our sample sizes, which allows us to model these properties adequately.

Concerned primarily with the trade-off between model complexity (i.e. dimension of the corresponding space of distributions) and model generality, we gradually explore a hierarchy of distributions that possess properties observed empirically. This is essentially a model selection phase and, as such, it relates to the *Bias-Variance Dilemma*: more constrained models promise more stable parameter estimation at the expense of fine structure of the unknown target distribution.

Our model-building tools include elementary Algebraic Geometry, Entropy Maximization Principle and some basic concepts of Information Theory. We adopt a common approach to modeling discrete multivariate probability laws by imposing Internal or External Constraints on the set of distributions on a given state space. Depending on the type of estimation used, these two categories cover log-linear, linear and even more general models. For example, some of our simpler models are based on constraining probabilities of a few "heavy" states (or even aggregates of states) to equal the respective values observed in the data; these are typical examples of internal constraints. (Shannon's) Entropy Maximization provides one way to estimate parameters, and we discuss some other estimation methods as well. We also study several models based on geometric and photometric symmetries of the state space. Here, hypotheses are put forward that the microimage distribution respects prescribed symmetries. Equations induced by such symmetries are examples of external constraints and we perform a series of statistical tests to quantify the significance of our evidence (data) in regard to these hypotheses.

We also discuss the duality between internal and external constraint problems in order to enrich our supply of computational methods for model estimation. Naturally, the main tools here are basic Linear Algebra and Information Theory.

Closed-form analytic expressions for some of our model distributions are also discussed. The purpose is to induce models for microimage quantizations which are finer than those dictated by the samples available here. Another natural extension would be to larger microimages, for example 5×5, a typical size for the support of local filters. A combination of basic Invariant Theory and Markov Random Fields provides a natural framework for such generalizations. This is also our point of contact with minimax learning. Of course, presently we focus mainly on microscopic phenomena of image statistics, as opposed to the far more ambitious aim of minimax learning - specifying distributions on larger image lattices. We stay "local" because we believe that existing knowledge of the image microworld is still rather incomplete, thus allowing a more systematic analysis than previously attempted prior to focusing on macroscopic problems. For example, microimage distributions emerging from our studies can also be naturally represented as maximum entropy extensions relative to certain local image filters. However, we do not start with a large supply of generic filters and then solve a massive optimization task in order to select the most relevant ones, as in minimax learning and texture modeling. In contrast, the origin of our filters is more algebraic (as opposed to purely analytic): they are merely bases for the microimage functions with specified symmetries. Due to the relatively small sizes of the mathematical spaces involved, we can hope to better understand the consequences of using one basis or another from our hierarchy by testing a series of statistical hypotheses.

Finally, it has not been among our goals to develop models dedicated to a specific imaging task. Nor do we suggest a single universal model, suitable for all practical situations. Instead, we examine a variety of models, each potentially useful for a particular category of applications.

1.3 Organization of the Thesis

Chapter 2 presents an account of our preliminary investigations of the natural image microworld. This chapter is also intended to introduce a concrete framework which embeds the central part of this work as an important chain.

The proper work begins by defining probability spaces appropriate for our analysis of microimage distributions. This is done in the first part of Chapter 3. The central theme of Chapter 3 is a statistical analysis of several types of distributional stability of natural microimages. The necessary hypothesis testing framework is briefly introduced in the same chapter, but the technical details of the particular statistical tests are deferred mainly to Appendix B. Related issues such as microimage sampling and alternate microimage distributions are supplemented in Appendix C. The conclusion of Chapter 3 lays the foundation for the rest of the work: We determine an appropriate, rather coarse (but yet non-trivial) level of quantization at which distributional stability of natural microimages is apparent. For example, respective estimators of the microimage distribution vary insignificantly from sample to sample from scale to scale. Therefore, modeling these distributions becomes meaningful.

Chapter 4 introduces a hierarchy of models for the microimage distribution. Model generation is based on our subjective experience accumulated in preliminary studies. Some models are inspired by *Gibbs-Markov Random Fields*. Technical details of several models are postponed until Chapter 6 and Appendix D in order to sooner enter the model testing phase.

Chapter 5 is then devoted to statistical testing of symmetry and other hypotheses associated with our models. There, our intuition confronts the facts: some seemingly reasonable models are "rejected", whereas others (including the one based on geometric symmetries) are "confirmed"; verification of yet other models proves somewhat more problematic. In the end, we obtain a detailed quantitative assessment of a variety of models.

In Chapter 6, we discuss the modeling methodologies followed in Chapter 4. Most of the issues are computational in nature. Some extend beyond stochastic image modeling, which also justifies placing them late in the presentation. Computations postponed from Chapter 4 are now properly explained. With the aim of extending the symmetry models to microimage spaces with finer quantization and/or larger supports, the second half of Chapter 6 is devoted to analytic representations for microimage distributions; Appendix D supplies further technical details.

In Chapter 7 we reiterate our key findings about the microworld of natural images and indicate directions for further work. These include extensions of our symmetric models to translation invariant Markov Fields (Gibbs Distributions) on larger lattices and applications of these models to tree-based microimage vector-quantization, with an eye towards defining efficient elementary features for a range of imaging tasks. (One possible specific scenario is outlined in §2.2.)

Finally, some statistical and information theoretic background information is briefly reviewed in Appendix A.

CHAPTER 2

EARLY WORK

2.1 Tags

Our earliest experiments with microimage statistics involved large samples of 6×6 patches extracted initially from hundreds of leaf images (representing different biological species) and later from a large photographic image. The goal was to analyze the quality of several hand-picked binary local image features based on *edge-detectors*, or *"tags"* [1]. Specifically, we wanted to measure the sensitivity of the detectors to changes in resolution. We were also interested in the degree of coherence among the positive responses of a particular detector at multiple resolutions and independent of the image class (i.e. species). Designed to be further aggregated into more complex and discriminating features, the tags were merely conjunctions of elementary tests of two types. One type of elementary operation was thresholding the signed difference between ω_s and ω_t, the intensities of two neighboring pixels: $X_{s,t,\delta} = \mathbb{I}_{\{\omega_s - \omega_t \geq \delta\}}$. The other type involved a comparison of unsigned differences between two pairs of neighboring pixels with one pixel in common: $Y_{s,t,u} = \mathbb{I}_{\{|\omega_s - \omega_t| \geq |\omega_t - \omega_u|\}}$. For example, the six constraints for having a "non-polar, diagonal tag" are listed in (2.1); the corresponding pixels are depicted in Figure 1, left.

$$\begin{aligned}
&|\omega_{s3}-\omega_{s6}| > |\omega_{s1}-\omega_{s3}|, \qquad &&|\omega_{s3}-\omega_{s6}| > |\omega_{s2}-\omega_{s3}|, \\
&|\omega_{s3}-\omega_{s6}| > |\omega_{s4}-\omega_{s3}|, \qquad &&|\omega_{s3}-\omega_{s6}| > |\omega_{s6}-\omega_{s5}|, \\
&|\omega_{s3}-\omega_{s6}| > |\omega_{s6}-\omega_{s7}|, \qquad &&|\omega_{s3}-\omega_{s6}| > |\omega_{s6}-\omega_{s8}|,
\end{aligned} \tag{2.1}$$

"Vertical" and "horizontal" tags were defined similarly. Thus in general, a (nonpolar) tag is present (say, at $s3$) if the intensity contrast between $s3$ and $s6$, the two "central sites" is larger than the contrasts in the six adjacent pairs. The tag polarity was introduced through $X_{s3,s6,\delta}$, an additional constraining test function. Thus, a polar tag was a conjunction of seven elementary tests.

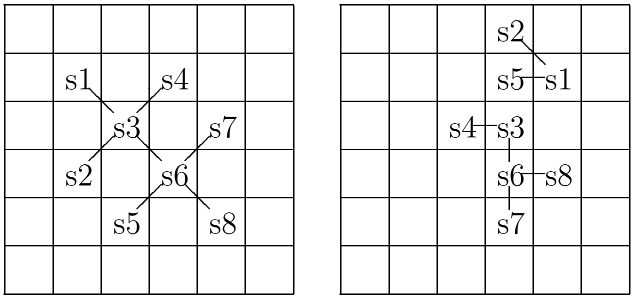


Figure 1: Examples of original (left) and new (right) tag configurations

We attempted to induce "better" conjunctions, measuring quality by the conjunction size (i.e. number of elementary tests). Specifically, probabilities of the tag presence were estimated from patch samples in order to specify the quality criterion: We wanted to construct test conjunctions of size six and larger whose positive responses would be at least as high as those of the tags. (Additionally, constraints on the fractions of the two types of elementary tests were also considered.) We then also wanted to check if candidate conjunctions that were successful at one scale would still be reasonably frequent at other scales. Simulations were conducted to grow conjunctions by randomly ordering the elementary tests and stopping once the frequency of the current conjunction fell below that of tags. For example, a new conjunction defined by (2.2) (see Figure 1, right) was found successful at one scale but was too rare at other scales.

$$\begin{aligned} \omega_{s1} - \omega_{s2} &\geq 8, \qquad \omega_{s4} - \omega_{s3} \geq 8, \\ \omega_{s6} - \omega_{s5} &\geq 8, \qquad |\omega_{s3} - \omega_{s1}| > |\omega_{s3} - \omega_{s2}|, \\ |\omega_{s1} - \omega_{s2}| > |\omega_{s1} - \omega_{s7}|&, \qquad |\omega_{s3} - \omega_{s7}| > |\omega_{s3} - \omega_{s2}| \end{aligned} \tag{2.2}$$

In summary, it proved generally difficult to find alternatives that would be better than the original tags at several scales simultaneously. This and the somewhat irregular spatial form of the new conjunctions confirmed the link between the tag regularity (expressed by multiple symmetries) and stability of their response statistics. It also hinted at the types of microimage symmetries that might be respected by a natural microimage distribution.

2.2 Microimage Coding and Vector Quantization

Before we present our next attempt to explore statistics of microimages (§2.3), we discuss a larger, computational framework that has motivated most of our statistical analysis of natural microimages. The main goal in developing this framework is to define computationally efficient, local image features using tree structured *vector quantization* (VQ) of microimages.

There are three main components of this approach. First, we assume the existence of *"universal" microimage codes* that partition the appropriate microimage space into *equivalence classes* and correspond to perceptually meaningful microimage *groupings*. "Universality" refers to certain types of distributional stability of

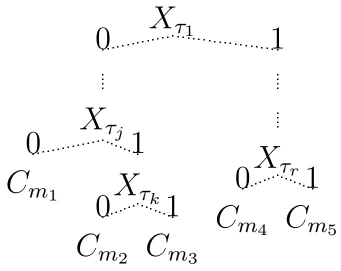


Figure 2: A fragment of a binary classification tree

microimages and the next section (§2.3) begins to explicate the subject and provides some examples of stable codes. Given such a coding scheme, each image pixel can, in principle, be transformed to a unique code value C absorbing "essential" information about its immediate vicinity. In practice, however, computing C exactly may be unnecessary and approximations are then entertained balancing the amount of detail requested against the amount of computations available. This leads to the other two components: $\mathcal{T} = \{X_\tau\}_\tau$, *elementary tests* that comprise the computations, and a decision criterion that defines the test selection strategy. Naturally, the three ingredients assemble into a microimage classification tree whose terminal nodes are the microimage codes $C_1, C_2, \ldots$, and whose intermediate nodes are the elementary tests X_τ. Figure 2 depicts a fragment of a binary classification tree based on some (binary) tests X_τ. Note that any partial (hence approximate) computation of the original codes C defines another, coarser microimage coding.

In summary, the idea is to effectively replace the *unsupervised* "divide-and-conquer" (§1.1 and [2]) approach to feature generation by *supervised* microimage classification. The proposed features are still conjunctions of elementary tests along tree paths. However, the tree induction mechanism that was previously based on the information theoretic concepts only, is now more flexible due to the introduction of the perceptual component through the microimage classes C.

We will return to the general discussion of microimage coding in the next section (§2.3) while now we are going to present, in compressed form, one concrete scenario to illustrate these ideas in the context of grey level microimages. To keep the exposition concrete, we fix the microimage size at 3×3; generalization to other sizes will be obvious. In [22] we defined "microimage quantization maps" and proposed a particular map, F_a, referred to as "alternate quantization", in order to "meaningfully" partition the microimage space $\Omega = M_{3\times 3}(\{0, \ldots, 255\})$ into "patterns". The number of patterns is considerably smaller than the number of original patches; also, the patterns respect only *relative* brightness and contain no direct information about absolute intensities. Thus, the corresponding microimage coding enjoys a degree of *photometric invariance*. Also, when the patterns were ranked based on relative frequencies (§2.3) estimated from a large microimage sample, the ranks were more stable than in similar experiments on uniform intensity quantization. The alternate, nonlinear quantization also yields a clean perceptual interpretation of the relative frequencies: The "Background" pattern comes first, then "edge"-like patterns, followed by "T-junctions", etc.

In this section, we focus on the issue of efficient computation of these patterns, and coarser codes based on them. Just as in the tag experiments (§2.1), the computations are based on $\mathcal{T}$, a collection of *elementary tests* of the form $X_{s,t,\delta} = \mathbb{I}_{\{\omega_s - \omega_t \geq \delta\}}$, where s, t, and δ are two pixel locations and a fixed threshold, respectively. In the vector quantization framework based on these tests, the number of tests entertained along a path of the resulting tree serves as a measure of computational complexity. These tests also reveal another advantage of the alternate quantization relative to the uniform one: The binary functions defined by these tests are sufficient to distinguish the F_a-patterns, whereas it can also be shown that patterns based on uniform quantizations cannot be computed exactly with these tests. In other words, the *maximal* partition attainable by these tests is at least as fine as the one induced by F_a. *Thus, the elementary tests determine the patterns.* We first reproduce the necessary definitions from [22] including that of F_a, and then prove the last statement.

As before, let $\mathcal{C}_L$ and $\mathcal{C}_{L_0}$ be two intensity ranges with $L_0 \leq L$. Any map F from the microimage set Ω_L to the pattern set Ω_{L_0} is called a *quantization* if it preserves *relative brightness*. Thus, if $\omega \in \Omega_L$, then

$$\omega_s \leq \omega_t \Rightarrow (F\omega)_s \leq (F\omega)_t,$$

The *alternate quantization* was defined to allow filtering out gradual intensity changes. Namely, assign the darkest pixel(s) the value 0, then assign the next darkest pixel(s) the label 0 if the difference is less than L/L_0 and the label 1 otherwise, and so forth. More precisely, suppose $L = 256$ and $L_0 = 8$; rank the pixels $s_1, s_2, \ldots, s_9$ by their intensities: $\omega_{s_1} \leq \omega_{s_2} \leq \cdots \leq \omega_{s_9}$. Put $(F_a\omega)_{s_1} = 0$. For $i = 2, \ldots, 9$, set $(F_a\omega)_{s_i} = (F_a\omega)_{s_{i-1}}$ if $\omega_{s_i} - \omega_{s_{i-1}} \leq 32$ and set $(F_a\omega)_{s_i} = (F_a\omega)_{s_{i-1}} + 1$ otherwise. Unlike uniform quantization, if $(F_a\omega)_s \neq (F_a\omega)_t$ then $|\omega_s - \omega_t| > L/L_0$.

For the sake of concreteness, suppose that $\mathcal{T}$ is ordered from 1 to 72 in some way so that $X(\omega) \in \mathcal{X} = \{0,1\}^{|\mathcal{T}|}$ returns the appropriately ordered bit-string of corresponding test values.

Theorem 2.1 Let X be the 72-dimensional binary vector whose components are precisely the tests in $\mathcal{T}$ with $\delta = L/L_0$ as above. Then $X(\omega)$ uniquely determines $F_a(\omega)$.

Proof. Fix a pixel enumeration independent of ω: $\{\omega_1, \ldots, \omega_9\}$.

Step 1: Let $\omega \in \Omega_L$ be fixed. We first define an equivalence relation on $\{\omega_s, s = 1, \ldots, 9\}$, whose equivalence classes $[\omega_s]_\omega$ are in one-to-one correspondence with the set of all labels assigned to ω_s's under F_a:

Definition 2.2 $\omega_s \sim_\omega \omega_t \iff$ either $\mid \omega_s - \omega_t \mid < \delta$, or $\exists i_1,..,i_K = 1,..,9$, such that $\mid \omega_{i_k} - \omega_{i_{k+1}} \mid < \delta$, where $k = 0,..,K+1$, and $i_0 = s$, $i_{K+1} = t$.

Clearly, the map $\omega \mapsto \sim_\omega$ is computable by the tests X, i.e., $\sim_\omega = \sim_{X(\omega)}$. In order to see this, start, for instance, with $A_1^{\omega_1} \overset{\text{def}}{=} \{\omega_1\}$ and evaluate the appropriate tests $X_{1,s,\delta}$ and $X_{s,1,\delta}$, $s = 2,\ldots,9$. This finds $A_2^{\omega_1}$, the subset of $[\omega_1]_\omega$ consisting of all ω_s such that $\mid \omega_1 - \omega_s \mid < \delta$. Repeat the same procedure by evaluating the appropriate tests for every $\omega_s \in A_n^{\omega_1} \setminus A_{n-1}^{\omega_1}$, producing $A_{n+1}^{\omega_1}$ until the expansion terminates: $A_{n+1}^{\omega_1} = A_n^{\omega_1} = [\omega_1]_\omega$.

Step 2: Let $\mathcal{E}^\omega = \{\omega_s, s = 1,\ldots,9\}/\sim_\omega = \{C_0^\omega,\ldots,C_{z(\omega)}^\omega\}$ be the set of corresponding equivalence classes written, for instance, in accordance with the order in $\{\omega_1,\ldots,\omega_9\}$ (i.e., $C_0^\omega = [\omega_1]_\omega$, $C_1^\omega = [\omega_s]_\omega$, where ω_s is the next pixel after ω_1 such that $\omega_s \notin C_0^\omega$, and so on). Furthermore, $\mathcal{E}^\omega$ can now be well-ordered by "$\prec$" defined as follows:

Definition 2.3 For C_1^ω, $C_2^\omega \in \mathcal{E}$, $C_1^\omega \prec C_2^\omega \iff \omega_2 - \omega_1 \geq \delta$, for some (and therefore, for all) $\omega_1 \in C_1^\omega$, $\omega_2 \in C_2^\omega$.

In order to see that the above relation is well-defined and is indeed a total-order relation on $\mathcal{E}^\omega$, it suffices to note that for any two C_1^ω, $C_2^\omega \in \mathcal{E}^\omega$ and any $\omega_1 \in C_1^\omega$, $\omega_2 \in C_2^\omega$, either $\omega_2 - \omega_1 \geq \delta$ or $\omega_1 - \omega_2 \geq \delta$ holds. Hence, if $m_i = \min C_i^\omega$, and $M_i = \max C_i^\omega$, where $i = 1,2$ and min and max are taken under the usual "$<$", then either $m_1 - M_2 \geq \delta$ or $m_2 - M_1 \geq \delta$. Thus, in fact, either $\omega_2 - \omega_1 \geq \delta$ or $\omega_1 - \omega_2 \geq \delta$ for all pairs $\omega_1 \in C_1^\omega$, $\omega_2 \in C_2^\omega$.

Step 3: Let $h(\cdot;\omega) : \mathcal{E}^\omega \to \{0,\ldots,8\}$ be the natural enumeration of $(\mathcal{E}^\omega, \prec)$. By definition of $\prec$, h depends on ω only through $X(\omega)$. Finally, noticing that $F_a(\omega)_s = h([\omega_s]_\omega, X(\omega))$ concludes the proof. $\diamond$

It also should be noted that $F_a(\omega)$ does *not* determine $X(\omega)$ as seen from the following example with three pixels: $\omega_1^1 = 0$, $\omega_2^1 = 20$, $\omega_3^1 = 40 \xrightarrow{F_a} 0, 0, 0$ with $\delta = 32$ and also $\omega_1^2 = 0$, $\omega_2^2 = 0$, $\omega_3^2 = 0 \xrightarrow{F_a} 0, 0, 0$, while clearly $X(\omega^1) \neq X(\omega^2)$.

2.3 In Search for Statistically Stable Microimage Codes

In the course of our search for statistically stable, primitive image codes [22] we entertained an *order-theoretic* approach to image analysis [27],[32],[36],[57]. Specifically, the idea was to relate perceptually meaningful *groupings* to partitions of the microimage space induced by special microimage operators. One intuitive choice is the *quantizations maps* used in [22], i.e. operators that at least preserve the order of pixel intensities when mapping microimages to patterns. The alternate quantization map F_a (§2.2) is our central example. Partially ordering such operators by "fineness" of the constancy classes provides some control over the amount of detail of the corresponding microimage codes. Probabilistically, these operators are random variables, and the partial order "coarser/finer" can be uniquely extended to a *total order* determined by *information content* or simply entropy. The following well-known elementary result proves useful for this task.

Proposition 2.4 Let F_1 and F_2 be random variables defined on $(\Omega, \mathbb{P})$, a common discrete probability space, in such a way that the partition $\mathcal{P}_2$ induced by F_2 is finer than or equal to the partition $\mathcal{P}_1$ induced by F_1. Then $H(F_2) \geq H(F_1)$.

Proof. That the partitions are nested is just another way of saying that F_2 determines F_1 (i.e., $F_1 = f(F_2)$ for some function f). Hence $H(F_2) = H(F_1, F_2)$. Since $H(F_1, F_2) = H(F_2|F_1) + H(F_1)$, it follows that $H(F_2) - H(F_1) = H(F_2|F_1) \geq 0$, as entropy is always nonnegative. $\diamond$

Thus, without the notion of entropy one can only judge about the amount of detail preserved by two microimage operators if one operator is a deterministic function of the other. Putting this into the information theoretic framework not only allows us to rank all such operators by the amount of information they convey about the underlying distribution, but it also ensures that such a ranking is consistent with the usual, deterministic order.

This naturally provides quantitative control over information loss due to aggregation. The qualitative, or perceptual, aspect, as well as the utility of the retained information for a particular imaging task, depends on the definition of the microimage operator.

In general, the order-theoretic framework for image analysis can be more flexible and natural than that of vector spaces. (A comparative discussion of the two, which predates the boom of *wavelet*-based techniques in image analysis, is presented in [57].) For instance, considering non-linear operators allows one to include more naturally the *photometric* dimension along with spatial geometry in the design and analysis of primitive image features with desired degrees and types of invariance. A common concern is the loss of computational efficiency enjoyed by linear filters. The vector quantization framework advocated in [22] and outlined in the previous section (§2.2) provides a sensible alternative to coding schemes based on linear filters. We have also attempted to address this issue in §6.6, discussing a framework for efficiently assembling banks of local non-linear, purely algebraic, filters. Such filters are designed to be invariant under basic symmetry transformations of 2×2 microimages, which are also respected by the empirical microimage distribution p. However, analogous constructions for larger microimages may be significantly less compact and require additional research.

We now turn to the issue of "universality" of microimage codes. Specifically, in order for the VQ-based microimage features ("supervised by" microimage codes) to be valuable in practice, it is highly desirable that the distribution of the underlying microimage codes change as little as possible across various image scenes and with respect to scaling and some global image intensity transformations. (This links the present work to the study of microimage statistics in [22].) We would then be able to focus on "the distribution" of natural microimage patterns. This goal, of course, imposes further conditions on the quantization maps. Intuitively, candidate quantizations must respect natural microimage coherence, or structure.

In [22], we analyzed two quantization maps, F_a, the alternate quantization (§2.2), and a simple modification of the standard uniform quantization. Although not comparable deterministically in the sense of "coarser/finer" relation (i.e. neither is a function of the other), the "uniform" map resulted in considerably larger loss of information than that incurred under F_a. This and other statistical experiments involving the two maps were based on a small (40 images) image data set (yet more diverse than that in the tag experiments in §2.1).

First signs of statistical stability of both types of our patterns were observed in the following ranking experiments. In short, 3 × 3 patterns were ranked from most frequent to most rare based on a large microimage sample. A "leading" hundred were identified as patterns present in all the 40 images. Furthermore, most of such patterns invariably retained their intra-image frequency-based ranks. (The effect was more pronounced in the case of F_a.) As explicit information about absolute brightness was eliminated through quantization, all microimages of "negligible" gradients were grouped into a single pattern (3 × 3 null matrix), naturally called "background". Under F_a, the background pattern clearly dominated the empirical distribution, whereas its closest contenders, should they be named at all, would undoubtedly be called "edges"; see Table 1 for fragments of a typical output ordered by the pattern frequency.

Pattern	0 0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0	0 0 1 0 0 0 0 0 0	1 1 1 1 1 1 1 1 0	1 1 1 0 0 0 0 0 0	1 1 1 1 1 1 0 0 0	0 0 1 0 0 1 0 0 0	1 1 1 1 1 1 0 0 1
Rank	1	2	3	4	9	10	32	33
Pr	0.68678	0.0064	0.0062	0.006	0.0047	0.0047	0.0012	0.0011

Table 1: Fragments of alternate pattern data ranked by frequency.

Next, similarity in the likelihood ranks of patterns that could be identified with each other by some basic symmetry transformation (e.g. rotation of the square) amplified our belief in the corresponding symmetry hypotheses. In fact, when symmetric patterns were aggregated into classes *invariant* under the corresponding symmetry transformations, we even observed significant stability of many point estimates for class probabilities across scenes and with respect to downscaling by block averaging. (An additional set of 40 images was used to assess these types of stability.)

We also defined and analyzed a measure of pattern complexity, and its empirical distribution also appeared very stable in our experiments. The results became even more striking when we replaced absolute probabilities by conditional ones given "non-background".

These results naturally led to connections with related work. First, the research on statistics of natural images [19],[59],[70] and on scale invariance in natural images [9],[23],[58] provided evidence and an explanation for universal scaling laws, in particular regarding the probability distribution of object sizes. We also compared our microimage statistics for natural images with those generated artificially from Poisson disk models, sampling the disk radii according to several different densities.

After adjusting for the amount of "background," the best statistical match was achieved in the case of inverse cube laws, as discussed in [9],[42], and [50].

Secondly, a parallel was drawn with work to characterize or even "derive edges" as "eigen-features" relative to various linear bases for decomposing small images (e.g., 12×12). In Independent Component Analysis (ICA), a natural generalization of Principal Component Analysis to higher moments, edges are characterized as "the most statistically independent" image features and comparisons are drawn with biological operations in the primary visual cortex ([4],[30],[43],[44],[51]). Our work suggested a somewhat different characterization of edges as "the most probable non-background" microimage configurations. The idea of efficient coding of natural images also relates our work to generalizations of ICA (e.g. [44]).

In the same spirit but in the context of image segmentation, *textons* are promoted as universal microimage patterns, obtained as prototypes or codes under vector quantization in high dimensional spaces of largely redundant filter responses [48]. The idea of generic subimage classification ("archetype classification") was also pursued in fractal image compression (e.g., [8]).

Encouraged by the results on statistical stability of our micro patterns, we conducted experiments to vector-quantize samples of natural 3×3 microimages along the lines of the VQ-framework of §2.2. The first results were satisfying in that the greedy strategy based on entropy reduction seemed reasonably efficient. A possible direction for future research is to consider globally optimal strategies [21]. Replacing on-line microimage classification by an off-line computation of the VQ tree based on the models presented in this work defines another dimension for the evolution of this research. Also, the induced tree structures further increased our confidence in the presence of several distinct features (e.g. symmetries) in the natural microimage distributions, and led directly to the work here - a more detailed analysis and modeling of these distributions.

CHAPTER 3

STABILITY ANALYSIS

We are interested in the population of all digital images of natural scenes publicly available from the world-wide web. This is, of course, different from studying images from a well-defined specific domain, e.g. *MRI*, urban scenes of a particular country, or wild life photographs of a particular geographic zone. In fact, we realize that our findings may be greatly refined and further substantiated by narrowing the scope of the investigation. Nonetheless, we hope to demonstrate the existence and practical value of some signatures common to the micro structure of all natural scenes.

Briefly, the goal of this Chapter is to build a foundation for probabilistic modeling of microimages of natural scenes (Chapters 4 and 6). First, we will develop a necessary statistical framework, providing definitions of underlying probability spaces and selecting appropriate sampling schemes. In this framework, we will then investigate the following three modes of statistical stability of natural microimages: *Inter-Scene Stability*, *Spatial Scale Invariance*, and *Photometric Scale Invariance.* Having a degree of stability as above will justify modeling the microimage distributions without worrying about the origin of the image database or scaling effects.

3.1 Underlying Probability Spaces

We assume that all images have a common gray scale, $\mathcal{C}_L = \{0, \ldots, L-1\}$. The (virtually infinite) population of all such images Σ_{im} defines the probability space $(\Omega_{im}, \mathbb{P}_{im})$ in the obvious way: $\mathbb{P}_{im}(I) = \frac{n(I)}{|\Sigma_{im}|}$, where $n(I)$ is the number of instances of image I in the population. Multiple copies of the same image are indeed existent on the web (i.e. $n(I) > 1$), albeit rarely observed. The image space here is $\Omega_{im} = \bigcup\limits_{m,n \geq 1} M_{m \times n}(\mathcal{C}_L)$, where $M_{m \times n}(\mathcal{C}_L)$ is the set of $m \times n$ matrices with coefficients from the set $\mathcal{C}_L$. Imposing an upper bound on the matrix dimensions in the image population renders Ω_{im} finite. Obviously, we declare all the subsets of Ω_{im} to be measurable and hence will leave aside reference to σ-fields.

The sizes of microimages that we study are 1 × 2, 2 × 1, 2 × 2, and 3 × 3. The corresponding probability spaces are of the form (Ω, p), where $\Omega = M_{m \times n}(\mathcal{C}_L)$ with m, n among the cases above.

Sometimes during the modeling phase, we will rescale $\mathcal{C}_L$ to fit in some standardized range (e.g. $[-0.5, 0.5]$), but even then it will always be enumerated by $\{0, \ldots, L-1\}$ in accordance with the usual order "<". If necessary, by default we will then use the following enumeration of Ω:

$$\begin{gathered} \Omega = \{\omega_1, \ldots, \omega_K\}, \text{ where } K = |\Omega| \text{ and} \\ k(\omega) = 1 + \omega_{m,1} + L\omega_{m,2} + \cdots + L^{n-1}\omega_{m,n} + \\ \cdots + L^{(m-1)(n-1)+1}\omega_{1,1} + \cdots + L^{mn-1}\omega_{1,n}, \end{gathered} \tag{3.1}$$

and refer to $k(\omega)$ as the index of ω.

The microimage measure p is naturally induced from Σ_{im} through the super population Σ of all the image blocks of the given size $m \times n$ extracted from every image I of the image population Σ_{im}. So, to each microimage ω, p simply assigns its relative frequency in population Σ:

$$p(\omega) \stackrel{\text{def}}{=} \frac{1}{|\Sigma|} \sum_{I \in \Sigma_{im}} n(\omega, I) = \frac{\sum\limits_{I \in \Omega_{im}} n(\omega, I) \mathbb{P}_{im}(I)}{\sum\limits_{I \in \Omega_{im}} S(I) \mathbb{P}_{im}(I)} = \frac{\mathbb{E}_{im} n(\omega, I)}{\mathbb{E}_{im} S(I)}, \tag{3.2}$$

where the random (through I) variable $n(\omega, I)$ counts the number of occurrences of ω in image I, and $S(I)$ is the total number of microimages of the given size in I. If I is of the size v×h, and ω is $m \times n$, then $S(I) = (\text{v}-m+1)(\text{h}-n+1)$.

We emphasize that with this set-up the *image size is a random quantity*, unlike in the more usual case of populations with fixed size images (e.g. [30], [34], [35].) This minor complication slightly affects our microimage sampling mechanism (§C.1.)

Also, one could alternatively think of images of variable sizes as samples from a common random field (supported either discretely or continuously). In such an approach, scaling would need to be modeled explicitly, hence requiring additional assumptions and leading to even more complications (e.g. microimages would need to be rescaled accordingly).

3.2 Image Data

Before we describe our data, we emphasize that achieving reasonable accuracy in estimating $\mathbb{P}_{im}$ or p from a sample typically available with our resources seems hopeless. Moreover, the former distribution is of little interest itself. Of course, with the rapid development of the web resources the situation could change soon, e.g., dynamic samplers may be installed directly connected to the major search engines. However, here we operate under much scarcer resources (both computer and human). Still, we hope to capture some interesting properties of p based on a relatively small sample of four hundred images. These images are diverse in origin, scale, and quality; there are landscapes, urban sights, people, animals and even a galaxy. Among them, there are forty randomly chosen 196×128 thumbnail pictures from the database used in [30]. Most of the other images come from public image resources and personal collections of the author's friends and colleagues. The largest single contributor is a personal artistic collection [39], from which more than a hundred images were sampled. We are going to use $\mathcal{I}_{im}$ for our image sample $\{I_1, \ldots, I_N\}$, $N = 400$, and $\mathcal{I}$ will refer to the set of all patches of a given size extracted from every image of $\mathcal{I}_{im}$.

Using standard image processing tools, full color images are converted to a scalar range fully spanning the 256 grey levels.

As early experiments revealed the usual problem of zero counts, we decided to coarsen the range to $L = 8$ for two pixel microimages, and to $L = 4$ in the 2×2 case. *Unless explicitly stated otherwise, these will be the default values for L throughout this work.* To ameliorate the resulting perceptual degradation, the uniform quantization was preceded by eight-bin intra-image *histogram equalization.* Figure 3 displays six images from the database before and after the intensity coarsening. To facilitate a potential comparison of our data with another set, we present some

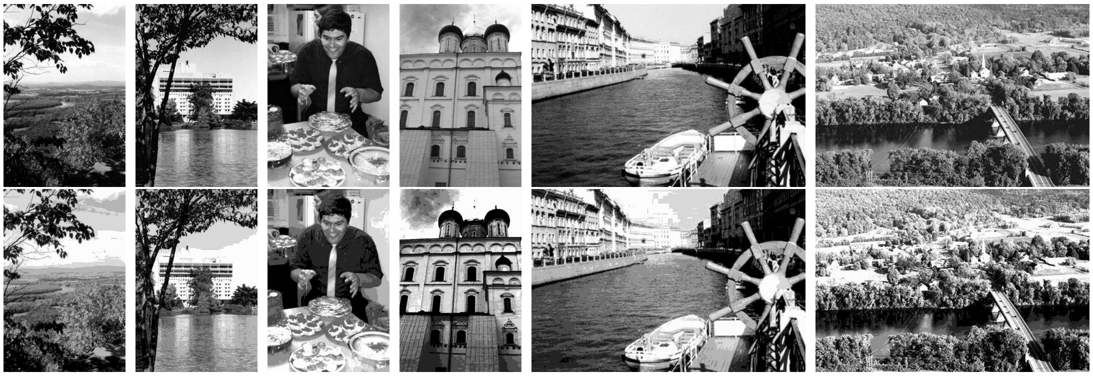


Figure 3: Image data. Original (top) and coarse (bottom) scales

crude statistics of image macro parameters. These are displayed in Figures 4. A

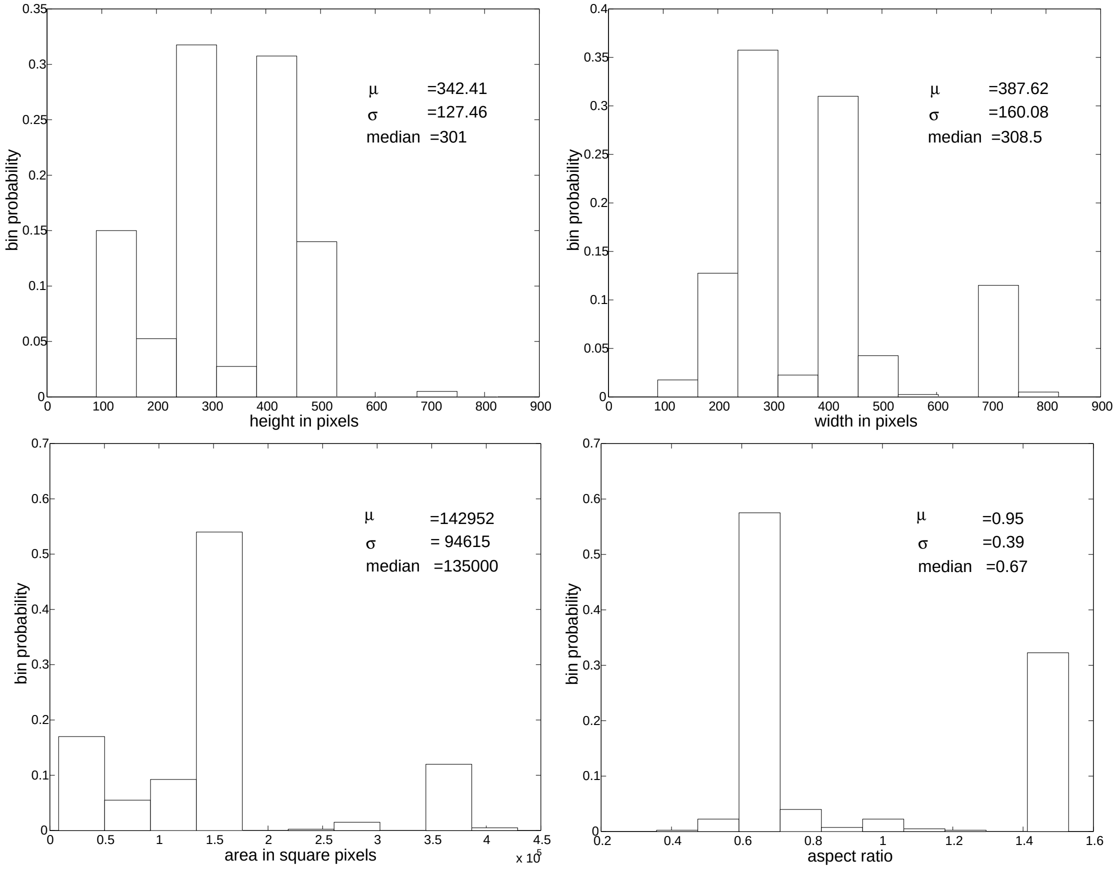


Figure 4: Image parameters. Top: Image height (left) and width (right). Bottom: Image area (left) and aspect (height/width) ratio (right)

small sample of 100 2 × 2 microimages extracted from one of the images in $\mathcal{I}_{im}$ is shown (eight gray levels were rescaled to maximize the contrast for better visualization) in Figure 5.

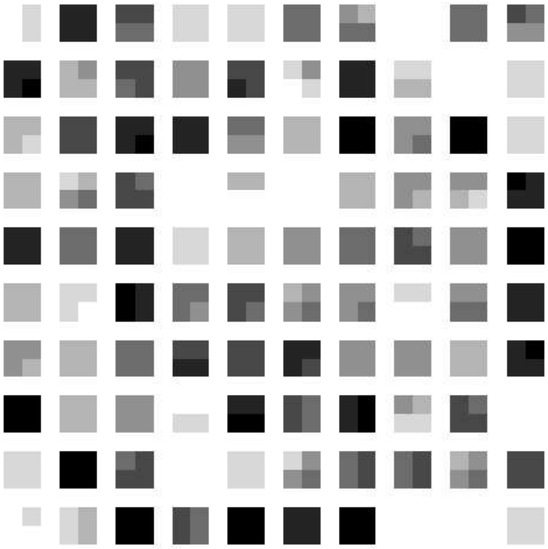


Figure 5: A subsample of 2×2 microimages from $\mathcal{I}$

3.3 Low Bin Counts and Aggregation of Rare Events

Obviously, our statistical analysis of natural microimages involves (sub)sampling microimages from $\mathcal{I}$ and the relevant discussion is presented in Appendix C. One of the related technical issues, however, needs to be mentioned right away. Namely, we will encounter the typical problem of low bin counts. For instance, in estimating p, sample independence and sufficient bin counts (e.g., five) are competing goals when N ($|\mathcal{I}_{im}|$) is fixed. Other resources being exhausted (in particular, $L = 4$ is the smallest intensity range we are willing to consider), we will collapse the rare events to a single cell, $D_\epsilon = \{\omega : \hat{p}(\omega) < \epsilon\}$. As long as D_ϵ^c has a non-degenerate composition (say, several dozens elements) of total mass close to 1 (e.g., 0.95), we will not be concerned with the effects of this aggregation and will assume we are still working with p. Indeed, for practical purposes rare configurations are often simply not important. Note however, that for a real application, a more intelligent and elaborate way to aggregate the rare events should be used; e.g. in the spirit of bottom-up *Huffman Coding* ([11],[22]) several, and not just one, aggregate states might be defined. Nonetheless, whenever we are merely concerned with the behavior of the principal masses of p, we will use the simple aggregation proposed above.

Finally, *we assume that we have a sufficiently large, random sample from p.* The corresponding empirical distribution (i.e. the *Maximum Likelihood Estimator* for p) is denoted by $\hat{p}^d$. Here, the superscript refers to the *sampling rate*, which constitutes one of the technical issues of generating microimage samples discussed in Appendix C (§C.2). However, in the ensuing presentation of the main work, the dependence of $\hat{p}^d$ on d will often be suppressed.

3.4 Symmetric Aggregation

The central theme of this chapter is the stability analysis of p and we are going to test a series of statistical hypotheses representing key stability properties. In this context, we are going to formulate such properties not only for p as a distribution on Ω, but also for the distribution induced by p on $\mathcal{S} \overset{\text{def}}{=} \Omega/\sim$, a quotient space of the 2×2 microimages ($L = 4$) under some symmetry equivalence relation $\sim$. Thus, in this case we analyze p integrated over elements of the partition $\mathcal{S}$. Still, we will continue to use "p", understanding by $p(\mathcal{O})$ the aggregate mass of $\mathcal{O} \in \mathcal{S}$, i.e. $p(\mathcal{O}) \overset{\text{def}}{=} \sum\limits_{\omega \in \mathcal{O}} p(\omega)$. When convenient, we will also refer to classes by a class representative in square brackets: Thus, if $\omega \in \mathcal{O}$, $[\omega]$ stands for $\mathcal{O}$.

Despite its technical similarity with partitions associated with quantization maps in §2.2, the ideas behind introduction of $\mathcal{S}$ are different from the search for perceptually meaningful, statistically stable microimage codes. First, any aggregation leads to larger counts, and consequently more robust testing. This purely technical reason to consider $\mathcal{S}$ is necessitated by our current sampling limitations (§§3.3,C.1,C.2). However, the aggregation chosen below is also going to be meaningful from the viewpoint of modeling p, since, in Chapter 4, we assign equal probabilities for patches from the same symmetry class $\mathcal{O} \in \mathcal{S}$, and test the validity of this symmetrization in Chapter 5. Thus, this aggregation also sets the stage for modeling p. The particular symmetry equivalence relation is further explained in §4.4 and also discussed in §6.4.1 in the proper algebraic context of Appendix D. It now only remains to specify what particular symmetries are entertained. Whereas in Chapter 4 we will define several models based on different sets of symmetries, we take but one example here, defined in the following subsection. Also note that this aggregation is different in principle from collapsing the rare events into a single class (§3.3) which we do in §3.6.1. Thus, this aggregation allows us to analyze the stability of p from a somewhat different perspective.

In Appendix C we will briefly return to the stability question to illustrate how a good model leads to lower estimation variability and thus more reliable testing (§C.6).

3.4.1 Rotation, Reflection, and Inversion Symmetries

We assign two patches ω_1 and ω_2 to one class $\mathcal{O}$ if and only if they can be obtained from each other by any number of the following three symmetry transformations: rotation by $\pi/2$ (e.g. counterclockwise), reflection through the secondary diagonal, and intensity inversion. The first two types are clearly geometric transformations exploiting the square nature of the patch; the choice of the rotation direction and the reflection axis is unimportant (Appendix D). Equation (3.3) illustrates the two geometric symmetries.

$$\begin{pmatrix} 4 & 3 \\ 1 & 2 \end{pmatrix} \overset{\text{rot.}}{\sim} \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} \overset{\text{refl.}}{\sim} \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix} \tag{3.3}$$

The last symmetry corresponds to replacing the original (micro)image by its "negative". With $L = 8$, for example, $\begin{smallmatrix} 4 & 7 \\ 0 & 3 \end{smallmatrix}$ is identified with $\begin{smallmatrix} 3 & 0 \\ 7 & 4 \end{smallmatrix}$. From now on we will refer to this particular partition of Ω as "$G - Symmetries$", and will later provide a more complete discussion.

3.5 Stability Analysis via Hypothesis Testing

We are going to analyze stability of the microimage distribution p across scenes (§3.6), with respect to downscaling by block-averaging (§3.7), and under global, linear rescaling of the image intensities (§3.8). This analysis is necessary to justify the ensuing modeling of p, and naturally advances our early, similar analysis ([22]) by putting it in the more rigorous framework of statistical hypothesis testing.

Thus, the goal is to demonstrate a whole group of related properties of p. First, imagine classifying the natural microimage population Σ by some global image attribute such as, e.g., type of scenery. This effectively divides Σ into several subpopulations each with its own variation of the microimage measure p. We would like to verify our hypothesis that such variations are virtually indistinguishable, that is p satisfies the property of *inter-scene* stability. This is only one of the several modes of universality of p that, if confirmed, would prove that the problem of modeling p is indeed *well-posed*, and hence promise a significant practical value. The other related types of universality of p are the absence of the effective spatial scale of natural images (*scale-invariance*) and lack of dependence of p on photometric transformations undergone by images during preprocessing.

Common to all these properties is the formulation of the corresponding statistical hypotheses. Namely, assume $\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ are two estimates of p obtained under different conditions appropriate for verification of the particular type of stability (e.g., two different scene types or two different spatial scales). The two estimates are thus assumed to come from two independent microimage subsamples. We then test the *two sample consistency* or *homogeneity of proportions* hypothesis: "$\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ are generated by the same distribution".

In order to test our stability hypotheses, we are going to apply a virtually infinite class of *Power-Divergence Tests* studied and advocated in [54]. Essentially the same tests are used for model testing in Chapter 5. These tests are based on the *Power-Divergence* quasi-distance measures $I(p, q; \lambda)$ between two probability vectors p and q:

$$I(p, q; \lambda) = \frac{1}{\lambda(\lambda+1)} \sum_k p_k \left[\left(\frac{p_k}{q_k} \right)^{\lambda} - 1 \right] \quad \lambda \in \mathbb{R}, \quad \lambda \neq -1, \ \lambda \neq 0. \qquad (3.4)$$

$I(p, q; -1)$ and $I(p, q; 0)$ are understood in the limiting sense and coincide with $D(q, p)$ and $D(p, q)$ (up to the base of the logarithm), respectively. ($D(p, q)$ is the usual Kullback-Leibler divergence [11],[40], Definition A.5.) Thus, the power-divergence tests are very natural in the sense that they connect Statistics with Information Theory. They also prove particularly convenient for our purposes because of the nature of the hypotheses we are going to test. Specifically, parameter estimation in these tests is simple in several respects in the current case of homogeneity of proportion testing as well as in the case of symmetry hypotheses. Also, these tests suggest a mechanism to assess validity of the large sample, i.e. asymptotic, results; and we will be able to take advantage of this tool. Among alternatives to these tests is the *bootstrap* [61],[62]. We do not use bootstrapping in the present work, although it would be interesting to further verify our findings by other statistical means.

The test statistics that we use for testing the stability hypotheses "$p^{(1)} = p^{(2)}$" are $T(\lambda) = N_1 I(\hat{p}^{(1)}, \hat{q}(\lambda); \lambda) + N_2 I(\hat{p}^{(2)}, \hat{q}(\lambda); \lambda)$, where N_1 and N_2 are the microimage subsample sizes and $\hat{q}(\lambda)$ are the *Minimal Power-Divergence Estimators*: $\hat{q}(\lambda) \overset{\text{def}}{=} \arg\min_{q \in \Theta^+} [N_1 I(\hat{p}^{(1)}, q; \lambda) + N_2 I(\hat{p}^{(2)}, q; \lambda)]$, where Θ^+ is the set of all positive distributions on Ω. Unlike $I(\cdot, \cdot; \lambda)$, these statistics incorporate sample sizes which allows one to build statistical tests based on universal asymptotics. Appendix B presents the key details from the general theory of these tests. The particulars relevant to testing the *two sample consistency* can be found in §B.2 and include expressions for $\hat{q}(\lambda)$ (B.7),(B.8). Depending on λ, among our tests are *Generalized Likelihood Ratio* and *Pearson's* χ^2 Tests for two sample consistency ($\lambda = 0$ and $\lambda = 1$, respectively). The choice of $\lambda = \frac{2}{3}$ is advocated as "golden mean" ([54], Appendix B) for several reasons including its suitability to detect a overall lack of fit in the presence of states with very small counts.

The general issues of obtaining $\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ are essentially the same as in the case of $\hat{p}$ (§§C.1,C.2) and the ones specific to particular stability experiments are presented below in the appropriate contexts.

3.6 Inter-Scene Stability

We split $\mathcal{I}_{im}$ into $\mathcal{I}_{im}^1$ and $\mathcal{I}_{im}^2$ by the following global criterion: $\mathcal{I}_{im}^1$ includes indoor scenes or distinct outdoor views of buildings. The rest of $\mathcal{I}_{im}$ goes in $\mathcal{I}_{im}^2$. Thus, in our experiments we say that p exhibits inter-scene stability if the null hypothesis "$\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ estimate the same distribution" is not rejected. The decision is based on the asymptotic approximation to the distribution of the test statistics $T(\lambda)$ under the null hypothesis. The theory of the power-divergence tests shows that in this case the asymptote is the χ^2 distribution with ν degrees of freedom (§B.2), where $\nu =$“$|\Omega|$”-1. The quotes are because of the adjustments caused by the state space reduction due to aggregations (e.g., $\Omega \to \mathcal{S}$).

In these and the ensuing hypothesis testing situations we ran tests several times to select “typical” results. Namely, most of the test results presented in this work correspond roughly to the median values of the test statistics observed in multiple repetitions of the same test. Further discussion of validity of our test results can be found in §C.3.

Consider first the case of 1 × 2 patches, in which Ω has 64 states ($L = 8$). A typical test run ($N_1 \approx 7000$ and $N_2 \approx 1000$) with different values of the test parameter λ is recorded in Table 2, top. Zero counts render the test statistic undefined for $\lambda < -1$ (Appendix B), hence the $\lambda = -2$ cell is blank . Although the test statistics $T(\lambda)$ seem close to each other, their variation around the asymptotic cut-off point $\chi^2_{0.95,63} = 82.53$[1] makes the situation somewhat inconclusive: Only five tests (corresponding to λ “large”) accept the null hypothesis at significance level $\alpha = 0.05$. In Appendix B we cite [54] to explain that such situations may occur when there are a small number of low count states

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	-	86.17	84.08	80.71	78.15	77.36	75.87	71.79
$P\text{-}val(\lambda)$	-	0.028	0.039	0.066	0.095	0.11	0.13	0.21
$T(\lambda)$	-	96.65	88.38	77.46	71.1	69.43	66.61	60.22
$P\text{-}val(\lambda)$	-	0.004	0.019	0.104	0.226	0.27	0.354	0.576

Table 2: Power-divergence tests for inter-scene stability. Top: Horizontal pairs. Bottom: Vertical pairs.

[1] If X is distributed as $\chi^2(\nu)$, then $\chi^2_{1-\alpha,\nu} \overset{\text{def}}{=} c$ such that $P(X > c) = \alpha$.

"spoiling" the match between the data and the null hypothesis. Thus, we identified and aggregated states with counts consistently below five (indeed, a third of them were frequently zeros) in order to convince ourselves of the overall good match between our data and the null hypothesis. The aggregate set (§3.3) is $D = \{(0,4),\ (4,0),\ (0,5),\ (5,0),\ (0,6),\ (6,0),\ (0,7),\ (7,0),\ (1,5),\ (5,1),\ (1,6),\ (6,1),\ (1,7),\ (7,1),\ (2,6),\ (6,2),\ (2,7),\ (7,2),\ (3,7),\ (7,3)\}$ and the corresponding test results are shown in Table 3, top. Similar results were obtained for the 2 × 1 case (Tables 2,3 bottom).

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	58.04	57.98	57.78	57.48	57.11	56.97	56.67	55.66
$P\text{-}val(\lambda)$	0.076	0.077	0.08	0.084	0.089	0.091	0.095	0.112
$T(\lambda)$	55.6	55.46	55.12	54.61	53.95	53.69	53.17	51.36
$P\text{-}val(\lambda)$	0.1136	0.115	0.121	0.131	0.145	0.15	0.162	0.208

Table 3: Power-divergence tests for inter-scene stability after aggregation. Top: Horizontal pairs. Bottom: Vertical pairs.

We now move to a demonstration of inter-scene stability in the 2 × 2 case.

3.6.1 The 2×2 Case

With $L = 8$ we would have $|\Omega| = 4096$, which is already comparable to our typical microimage sample sizes (see above). Clearly, no sensible testing is possible then. This is the reason for choosing $L = 4$; consequently, $|\Omega| = 256$. Now, $N_1 \approx N_2 \approx 9000$. Based on the argument of preventing low counts, and also after visual inspection of several histograms, we aggregated 136 most rare configurations in one compound state. A few examples of the collapsed micro configurations are: $\begin{smallmatrix}3&0\\0&1\end{smallmatrix}$, $\begin{smallmatrix}1&0\\3&2\end{smallmatrix}$, $\begin{smallmatrix}2&0\\0&2\end{smallmatrix}$, $\begin{smallmatrix}2&0\\3&3\end{smallmatrix}$. Based on the asymptotic cut-off point ($\chi^2_{0.95,120} = 146.57$) the null hypothesis is not rejected (Table 4, top) but the significant variation of the test statistic with λ suggested a more drastic aggregation to overcome the empty bin effect. Thus, we additionally collapsed 47 rare states. To our satisfaction, the results in Table 4, middle, indicate a overall good standing of the null hypothesis in all the eight tests ($\chi^2_{0.95,72} = 92.808$).

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	-	145.02	144.9	137.05	131.38	129.7	126.67	118.5
$P\text{-}val(\lambda)$	-	0.06	0.061	0.137	0.225	0.257	0.321	0.522
$T(\lambda)$	87.34	88.19	88.17	87.72	87.01	86.71	86.05	83.59
$P\text{-}val(\lambda)$	0.105	0.095	0.095	0.1	0.11	0.114	0.124	0.165
$T(\lambda)$	24.1	24.44	24.43	24.31	24.1	24.01	23.81	23.07
$P\text{-}val(\lambda)$	0.514	0.495	0.495	0.501	0.514	0.519	0.531	0.574

Table 4: Power-divergence tests for inter-scene stability. Top: Partially aggregated Ω. Middle: More severe aggregation of Ω. Bottom: G-symmetric aggregation.

Now we perform the power-divergence tests for two-sample consistency in order to test inter-scene stability of p relative to $\mathcal{S}$ (§3.4). In §6.4.1 we present a general formula (Proposition 6.8) for the size of $\mathcal{S}$ as a function of L. With $L = 4$, $|\mathcal{S}| = 31$ (resp., $|\Omega| = 256$). Low counts are now less of a problem: There are only six rare classes:

$$\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix}.$$

Their aggregate mass is less than 0.02 and these classes contain only 38 microimages in all. (This should be compared with the previous aggregations of 136 and 183 rare microimages.) Clearly (Table 4, bottom), independently of λ the tests support the null hypothesis.

In §C.6 we will complement this discussion by showing an increase in test reliability when the raw empirical distributions $\hat{p}^{(i)}$, $i = 1, 2$ are replaced by a model (Tables 21 and 22).

3.7 Scale Invariance

A very thorough mathematical treatment of scale invariance in terms of probability measures on abstract image spaces can be found in [50]. However, the subject of our work is more concrete, and we continue to use the hypothesis testing framework to analyze scale invariance of p. The present analysis substantiates our previous work on this subject [22].

There are many ways to define "scale invariance" depending, first of all, on how downscaling (i.e., downsampling) is performed. The most popular downscaling operators are *block-averaging* (BA), *subsampling* (SS), and *median filtering* (MF). Usually, they lead to similar results [34],[42].

Assuming image I has dimensions 2v×2h, downscaling by 2×2 block averaging transforms I into a v×h image $BA(I)$ where: $BA(I)(i,j) = 0.25[I(2i-1,2j-1) + I(2i-1,2j) + I(2i,2j-1) + I(2i,2j)]$ for $i = 1,\ldots,$v and $j = 1,\ldots,$h. Simple subsampling of I returns an v×h image $SS(I)$, retaining, for instance, the upper-left pixel in each 2×2 block: $SS(I)(i,j) = I(2i-1,2j-1)$. Finally, each pixel value $MF(I)(i,j)$ in the v×h image obtained via median filtering of I is the median of the four original intensities $\{I(2i-1,2j-1),\ I(2i-1,2j),\ I(2i,2j-1),\ I(2i,2j)\}$. We emphasize that these operations are to be performed prior to any preprocessing (e.g. histogram-equalization and uniform coarsening of the intensity range).

Next, imagine performing one of these three operations on every image from the population Σ_{im}, thus obtaining a new image population and, consequently, image distribution $\mathbb{P}'_{im}$ in the same way as $\mathbb{P}_{im}$ is obtained from Σ_{im} (§3.1). The new microimage population and corresponding distribution are Σ' and p' respectively. In principle, one could similarly define Σ''_{im}, Σ'', p'', etc. We can then define what we mean by "scale invariance" of natural images:

Definition 3.1 If p and p', p'' etc. are all equal, we will say that p is scale invariant.

Of course, in reality scaling is effectively limited to a reasonable finite range. Thus the above definition is in practice too restrictive and the strict equality must be weakened and a distance introduced. In order to be able to test scale invariance, we restrict our discussion to the two scales only, i.e., the null hypothesis is $p = p'$. We will use block-averaging to obtain $\mathcal{I}'_{im}$ from $\mathcal{I}_{im}$.

In practice, scale invariance of natural images is often reported by graphing various statistics (e.g. single pixel histograms) extracted at different scales, and then appealing to a visually close match between the two such plots [9], [34], and [35]. Thus, visual inspection has been commonly relied on. When image sample sizes are as large as tens of thousands and a virtually continuous image function (such as the logarithm of the histogram) is analyzed [34], statistical stability may be sufficiently apparent to ignore a more formal, numerical analysis of the errors. Our early and somewhat simplistic attempts to address statistical significance in experiments on stability in general, and scale invariance in particular, reflect our concern with the validity of conclusions based on relatively small samples and are documented in [22].

We now take a natural step toward rigorous verification of scale invariance by testing the hypothesis "$p = p'$". Again, we will use the Power-Divergence tests similar to the analysis of inter-scene stability (§3.6). Thus, we are going to split $\mathcal{I}_{im}$ into $\mathcal{I}^1_{im}$ and $\mathcal{I}^2_{im}$, then downscale one of them, and finally compute $\hat{p}^1_d$ and $\hat{p}^2_d$ respectively. To have comparable numbers of microimages contributing to $\hat{p}^1_d$ and $\hat{p}^2_d$ (which is desirable in two sample consistency tests §B.2), we can either divide $\mathcal{I}_{im}$ approximately as $1 : 4$, or split it evenly but sample microimages from the downscaled image subsample at the rate $4d$. In order to better control introducing dependencies in the microimage samples, we choose the former method and thus assume that we generate random samples from p and p'. Note that, at any rate, the downscaling leads to a decrease in the microimage sample size relative to the inter-scene stability analysis. We also performed similar experiments without the division of $\mathcal{I}_{im}$, namely obtaining microimage samples from the entire $\mathcal{I}_{im}$ and its downscaled version. Of course, there is then the risk of introducing dependence among the microimage samples from the two different scales.

We now present typical[2] test results based first on $1:4$ splitting of $\mathcal{I}_{im}$ (Table 5) and second on the entire $\mathcal{I}_{im}$ (Table 3.7). In the latter situation the sampling rate for the downscaled microimages will be $4d$ in order to balance the resulting microimage sample sizes. The actual sampling rate (C.2) in the first situation is $d = 0.0002$ and the same aggregations as in §C.2 are in effect for both the 1×2 and 2×1 cases. Thus, there we have 44 degrees of freedom for the asymptotic test distribution under the null. Coping with the loss of about half the sample size that we had in §3.6.1 for the 2×2 case, we now aggregate the rare patches more dramatically. Alternatively, we can lift the testing to the G-symmetric quotient space $\mathcal{S}$ (§3.4.1) and aggregate several more classes in addition to the ones we had in §3.6.1. We chose the latter method due to more meaningful aggregation. At the end, we arrive at 15 classes and thus 14 degrees of freedom; the "rare" super class absorbs less than 3% of the total mass.

In the second situation (i.e., $\mathcal{I}_{im}$ is not divided), we reduce the sampling rate d to partially compensate for the anticipated gain in inter-sample dependence (C.2). Specifically, we use $d = 0.0001$ in both the two-pixel cases, and, guided by stabilization of the test statistics with respect to the parameter λ, we take $d = 0.00006$ in the 2×2 case. *In summary, the scale-invariance hypothesis is not rejected.*

[2] Multiple test runs show satisfactory agreement with respective χ^2 asymptotics, comparable with the similar results in §3.6.

For the rest of the stability analysis we will only use the experiments based on splitting $\mathcal{I}_{im}$: Several test reruns confirm that the other set-up is inferior to this one, namely, dependence among the two microimage samples inevitably leads to the test data resembling the respective asymptotic distributions less.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	56.82	57.71	57.54	56.95	56.07	55.73	55.04	52.91
$P\text{-}val(\lambda)$	0.09	0.08	0.08	0.09	0.11	0.11	0.12	0.17
$T(\lambda)$	48.77	49.83	49.66	49.02	48.02	47.62	46.78	44.04
$P\text{-}val(\lambda)$	0.29	0.25	0.26	0.28	0.31	0.33	0.36	0.47
$T(\lambda)$	12.62	12.78	12.71	12.55	12.32	12.23	12.05	11.47
$P\text{-}val(\lambda)$	0.56	0.54	0.55	0.56	0.58	0.59	0.6	0.65

Table 5: Scale invariance tests based on $\mathcal{I}^1_{im}$ and $\mathcal{I}^2_{im}$. Top: Vertical pairs. Middle: Horizontal pairs. Bottom: 2×2 $L = 4$. Partially aggregated $\mathcal{S}$.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	57.26	58.08	58.02	57.64	57	56.74	56.18	54.3
$P\text{-}val(\lambda)$	0.09	0.08	0.08	0.08	0.09	0.09	0.1	0.1
$T(\lambda)$	49.24	50.2	50.2	49.76	49.06	48.77	48.14	45.97
$P\text{-}val(\lambda)$	0.27	0.24	0.24	0.26	0.28	0.29	0.31	0.4
$T(\lambda)$	20.87	21.01	20.99	20.93	20.82	20.77	20.67	20.31
$P\text{-}val(\lambda)$	0.11	0.1	0.1	0.1	0.11	0.11	0.11	0.12

Table 6: Power-divergence tests for scale invariance based on entire $\mathcal{I}_{im}$. Top: Vertical pairs. Middle: Horizontal pairs. Bottom: 2×2 $L = 4$. Partially aggregated $\mathcal{S}$.

3.8 Intensity Transforms

Our last experiments on stability of p involve (global) intensity transforms. Namely, regarding an v×h image I (with the original intensity, i.e. $L = 256$) as an element of the vector space $\mathbb{R}^{\mathrm{vh}}$, we linearly rescale it by some positive α, rounding the resulting values and truncating them to fit again into $\mathcal{C}_L = \{0, \ldots, 255\}$: $I \rightarrow \min(255\vec{1}, \alpha I)$. These transforms take place before the preprocessing (histogram equalization and uniform quantization). The goal here is to analyze invariance of p against the induced degradation, inevitable when $\alpha \neq 1$. Similar experiments were conducted in our early work but not in a hypothesis testing framework. Thus, any such transformation replaces the original image population Σ_{im} by one consisting of transformed images, and this finally leads to another measure p' on the microimage space Ω. The data in Table 7 provide typical test statistics for testing the hypothesis $p = p'$. We only display the results for the more important 2×2 case, with $L = 4$ and the rarest 17 classes integrated into one, exactly as in §3.7. The image sample is randomly and almost evenly divided in $\mathcal{I}^1_{im}$ and $\mathcal{I}^2_{im}$ to generate random microimage samples from p and p' under the sampling rate $d = 0.0001$. The top rows correspond to $\alpha = 0.5$ and the bottom ones to $\alpha = 1.5$. In both cases, the invariance is evident. Multiple test reruns show good agreement between the test data and the $\chi^2(14)$ asymptote.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	13.28	13.37	13.36	13.32	13.25	13.22	13.16	12.89
$P\text{-}val(\lambda)$	0.5	0.5	0.5	0.5	0.51	0.51	0.51	0.54
$T(\lambda)$	20.32	20.38	20.36	20.32	20.26	20.23	20.17	19.94
$P\text{-}val(\lambda)$	0.12	0.12	0.12	0.12	0.12	0.12	0.13	0.13

Table 7: Power-divergence tests for invariance under truncated linear intensity transforms: Top: Contraction. Bottom: Expansion.

3.9 Summary

We have introduced a mathematical framework for the analysis and modeling of natural microimage distributions. The main ingredients of this framework are:

- (Ω, p) - the probability space of microimages (matrices) of appropriate size.
- *Minimum Power-Divergence Tests* - a parametric family of statistical tests for evaluating various hypotheses involving p.
- $\mathcal{I}_{im}$ - the random sample of $N = 400$ "natural" digital images.
- $\mathcal{I}$ - the collection of all microimages extracted from $\mathcal{I}_{im}$.
- $\hat{p}^d$ - the (unconstrained) maximum likelihood estimator of p (which is based on the mean of a random microimage sample; d is the sampling rate and will often be suppressed).

Also, we have addressed *methods of aggregation of elements of* Ω *in order to prevent low bin counts.* Other technical issues to optimize statistical inference under current limitations such as the insufficient image sample are addressed in Appendix C. Most important of those issues is *generation of a sufficiently large but random microimage sample, and specifically, determining the sampling rate* d (§§C.1,C.2,C.3).

Inter-Scene Stability of the microimage distribution p has been introduced as the property of near independence of p of the particular type of natural image scenery. *This property has passed the appropriate hypothesis tests.*

Except in abstract image analysis, *Scale Invariance* can only be defined in approximate terms. Nonetheless, we have successfully tested our microimage data at two different scales: *The respective empirical distributions are largely the same.*

We have also *observed statistically insignificant variation of our microimage data under global contraction and expansion of the intensity range.*

In summary, we have not rejected the following three hypotheses:

- Inter-Scene Stability
- Spatial Scale Invariance
- Photometric Scale Invariance

Thus, not only have we corroborated our early results [22] and results reported elsewhere on stability of microimage statistics, but we have now put the corresponding analyses in a more rigorous statistical framework. This provides a firm foundation for the ensuing modeling and analysis of p.

CHAPTER 4

PROBABILITY MODELS FOR NATURAL MICROIMAGES

After a short discussion of probabilistic modeling, we define most of the models studied in this work. Some of the definitions will be presented at an intuitive level with formal explanations to follow in Chapter 6 and Appendix D. One reason for this staggered presentation is to expedite the transition to Chapter 5, where these models are evaluated in the context of statistical hypothesis testing.

4.1 A Few Words on Probabilistic Modeling

In general, one can argue that probabilistic, or stochastic, modeling is the most natural approach to the exploration of complex systems on the basis of distorted, conflicting, incomplete, or overwhelming information. Equipped with the infinite flexibility of general probability spaces, not only do we employ stochastic models to explain current observations, but we also expect these models to guide us in optimal acquisition of new data. Eventually, the most natural framework for the automation of *learning* (which, until recently, was an unchallenged prerogative of the living forms) might also be stochastic.

Among the ultimate goals of probabilistic modeling is efficient data representation in the sense of a compact yet accurate summary of observed statistics. Whether one is interested only in numerical measurements of the system (*estimation*) or also in complex *decision making* based on such measurements, it is imperative to recognize the sensitivity of the estimates of the system parameters to the particular sample. A suitable model, therefore, is needed in order to “smooth” or “regularize” this variability. Naturally then, discerning between “essence” and “noise” determines which models capture the key features of the actual data generation mechanism. Given sparse data or limited resources for analysis, one is obliged to consider only primitive models. Even then, an intelligent selection among the candidates can make a difference by “learning” some, possibly very coarse, properties of the system.

Apart from resource limitation (e.g. human labor, computer time and memory), models are commonly compared with one another in light of the Bias-Variance Dilemma (see, for example, [7],[24]). Namely, one seeks a balance between intrinsic model complexity and the amount of data available. High variance in model selection (parameter estimation) results from a model family being too complex. In terms of statistical learning, this corresponds to overfitting the data and leads to poor generalization. On the other hand, if the model family is too restrictive, the risk of large estimation bias is high. Ordering model families by their complexity is therefore very important to facilitate selection. This is one of the main aspects of the ensuing investigation of microimage probability models.

The notation we develop in this chapter will be followed throughout the work. Recall (§3.1) that we can index Ω by, for example, $k(\omega)$, $k = 1, \dots, K = |\Omega|$ as in (3.1). Thus we will identify real functions f on Ω with vectors in $\mathbb{R}^K$ via: $f(\omega_k) \equiv f_k$ for any such function (vector). We will often use the probability simplex $\Theta \overset{\text{def}}{=} \{q : q \in \mathbb{R}^K;\ q_k \geq 0,\ k = 1, \dots, K;\ \sum\limits_k q_k = 1\}$.

In order to avoid trivialities in testing hypotheses corresponding to our models, we add the positivity condition to all our model definitions. (Actually, some of our models will be originally defined to satisfy this condition.) The other motive is, of course, the assumption that the positivity also holds for p (i.e. $p_k > 0$ $k = 1, \ldots, K$). In fact, in the testing framework (Chapter 5 and Appendix B) our *saturated*, or *ground*, model subspace is $\Theta^+ \overset{\text{def}}{=} Q^+ \cap \Theta$, the interior of Θ, where $Q^+ \overset{\text{def}}{=} \{v : v \in \mathbb{R}^K;\ v_k > 0,\ k = 1, \ldots, K\}$ is the positive quadrant. When we later mention distributions on sets other than Ω (Chapter 6), the appropriate dimensions should always be clear from the context.

4.2 Modeling by Constraints

Generally, modeling includes two stages - model generation and model selection. In the first stage, a family of distributions satisfying some properties believed to be essential is proposed. These properties are often characterized as mathematical constraints, which are a common representation for prior knowledge. We will refer to the proposed family as "feasible" and will denote it by F. For the rest of this chapter, we only discuss modeling finite systems (i.e., finite state spaces).

In the second stage, a particular feasible model must be selected. Depending on the context, this may involve parameter estimation or global optimization.

From now on, by "data" we mean the empirical microimage distribution $\hat{p}$ (§§3.3,C.1). Recall that $\hat{p}$ is simply the *unconstrained* MLE of p: $\hat{p}(\omega_k) = \frac{n_k}{N_\omega}$, where the n_k's are the sample frequencies and N_ω is the sample size[1]. In the ensuing discussion of two typical types of constraints [54] we will emphasize the asymmetry of

[1] The subscript is only to distinguish N_ω, the microimage sample size, from N, the image sample size.

Kullback-Leibler divergence $D(\cdot,\cdot)$ by referring to $D(p,q)$ as the divergence *from p to q*.

4.2.1 External Constraints and Maximum Likelihood Estimation

In the external constraints framework (ECP), *prior to observing the data*, with each model we associate a subfamily $\Theta_0 \subset \Theta^+$, the positive distributions on Ω. Thus, Θ_0 plays the role of the feasible region F above and is usually engendered by recognizing patterns in multiple data sets or simply speculating about the data-generation process prior to experimentation.

One example is linear constraints: $\Theta_0 = \{q \in \Theta^+ : \ \mathbb{E}_q V_j = \mu_j, \ j = 1, 2, \ldots\}$, where both the constraining functions V_j and target values μ_j are *fixed before observing the data*. Most of our models originate in this way, and an important special case is modeling by symmetry constraints (4.7) (§4.4). Later in the work (§6.1), we treat this special case in more detail.

The other example that we use from this framework is Θ_0^{loglin}, the log-linear model with the parameter space Γ:

$$\begin{aligned} \Gamma &= \{\gamma \in \mathbb{R}^{J+1} : \sum_{k=1}^{K} e^{\gamma_0 - 1 + \sum\limits_{j=1}^{J} \gamma_j V_{jk}} = 1\}, \\ \Theta_0^{loglin} &= \{q \in \mathbb{R}^K : q_k = e^{\gamma_0 - 1 + \sum\limits_{j=1}^{J} \gamma_j V_{jk}}, \gamma \in \Gamma\}, \end{aligned} \tag{4.1}$$

Symmetry constraints on positive probability vectors can also be represented as a special case of the log-linear model (§6.1).

Except in trivial cases, the empirical distribution always lies outside the respective feasible region $F = \Theta_0$ defined by such constraints. In the case of the symmetry models, this is to say that even sampled from a truly symmetric distribution, data, due to noise, never respect the symmetries exactly.

At last, a single feasible distribution is usually selected from Θ_0 (e.g., via minimization of some distance-like measure of closeness to the data) which corresponds to estimating the appropriate model parameters. In general, statistical testing is desirable in order to evaluate the significance of the deviation of the empirical distribution from the appropriately selected feasible model.

In any such case, a classical tool for parameter estimation is *constrained maximization of the (log)likelihood function*, or constrained MLE:

$$\hat{p}_{MLE,\Theta_0} = \arg\max_{q\in\Theta_0} \sum_{k=1}^{K} n_k \log q(\omega_k) = \arg\max_{q\in\Theta_0} \sum_{k=1}^{K} \hat{p}(\omega_k) \log q(\omega_k) \tag{4.2}$$

Note also that $\hat{p}_{MLE,\Theta_0} = \arg\min_{q\in\Theta_0} \sum_{k=1}^{K} \hat{p}(\omega_k) \log \frac{\hat{p}(\omega_k)}{q(\omega_k)}$. *Thus, constrained MLE is equivalent to constrained minimization of Kullback-Leibler divergence $D(\hat{p}, q)$ (with the same constraints) from the empirical distribution.* We will sometimes refer to the constrained MLE as “ECP/MLE”.

In §6.1, we compute $\hat{p}_{MLE,\Theta_0}$ explicitly in the case of symmetry-based constraints (6.4). Given the existence and uniqueness of $\hat{p}_{MLE,\Theta_0}$, it also makes sense to introduce a *deterministic function* $\mathcal{R}$ mapping an arbitrary positive distribution ν to $\arg\max_{q\in\Theta_0} \sum_{k=1}^{K} \nu(\omega_k) \log q(\omega_k)$. This operator is a key ingredient of Chapter 6.

In the case of symmetry-based constraints, *generalized minimum power-divergence* estimators [54],(B.4),(B.5) are also available and are used for hypothesis testing along with the constrained MLE (Chapters 3 and 5). Naturally, this framework is suitable for testing hypotheses of the form “$p \in \Theta_0$” (Chapter 5, Appendix B).

4.2.2 Internal Constraints and Maximum Entropy Estimation

In the previous setup, the model generation and model selection stages are clearly separated: The feasible region F is defined independent of the data immediately producing a parametric model family Θ_0, and the model selection is essentially parameter estimation. In contrast with the previous framework, internal constraints (ICP) depend on the data. Consequently, the corresponding feasible region F is defined *relative to the empirical distribution* and hence is a *function of the data*. Namely: $F = F(\hat{p}) = \{q \in \Theta^+ : \ \mathbb{E}_q V_j = \mathbb{E}_{\hat{p}} V_j, \ j = 1, 2, \ldots\}$. Thus, F is a (convex) subset of all probability distributions on a given set with certain expectations equal to their corresponding values observed in the data. The empirical distribution is then necessarily inside this family: $\hat{p} \in F(\hat{p})$; hence the characterization "internal".

If we believe that F incorporates all the knowledge that we presently have about the distribution of interest, then selection of the particular feasible member is often driven by maximizing residual uncertainty as measured, for instance, by *Shannon's Entropy H* (Definition A.1). Thus, a natural and classical model selection principle is constrained *entropy maximization*, or the *maximum entropy extension* (MEE, §4.3). We write V for the matrix whose j^{th} row is the function V_j evaluated at each patch ω_k, $k = 1, \ldots, K$. Assuming positivity of $\hat{p}$, the V-constrained estimator of p

$$\hat{p}_{MEE,V} = \arg \max_{q \in F(\hat{p})} H(q) = \arg \max_{q \in F(\hat{p})} \sum_{k=1}^{K} q(\omega_k) \log \frac{1}{q(\omega_k)} \tag{4.3}$$

is well-defined and admits a specific exponential form. (Provided the rows of V are linearly independent, this form is essentially Gibbs (4.6).) Entropy maximization is then also said to *smooth* the empirical distribution.

We are going to consider several models that originate in this way. Central examples are *Potts* (§4.6) and *pairwise interaction* (§4.7) models.

Note that $\hat{p}_{MEE,V} = \arg \min_{q \in F(\hat{p})} \sum_{k=1}^{K} \hat{q}(\omega_k) \log \frac{q(\omega_k)}{u_k}$, where $u_k \equiv 1/K$ is the uniform distribution. Thus, *MEE under internal constraints is equivalent to constrained minimization of the Kullback-Leibler divergence $D(q, u)$ (with the same* *constraints) to the uniform distribution.* We will often refer to these two equivalent statements as "ICP/MEE". The authors in [54] suggest generalizations of these methods, replacing $D(\cdot,\cdot) = I(\cdot,\cdot;0)$ by other members of the power divergence families $I(\cdot,\cdot;\lambda \neq 0)$ (Appendix B). Thus, additive models more general than linear ($\lambda = 1$) and log-linear ($\lambda = 0$) are engendered by this framework through minimization of $I(\cdot, u; \lambda)$. This naturally extends the entropy maximization principle and, consequently, gives rise to a whole class of uncertainty measures, generalizing Shannon's entropy.

In the case of modeling p, the natural microimage distribution, we place a higher priority on selection of (internal) constraints than on alternative parameter estimation methods such as minimization of $I(\cdot,\cdot;\lambda)$ with various values of $\lambda \neq 0$. Thus, in modeling by internal constraints we adhere to the usual MEE ($\lambda = 0$), which does not seem to limit our modeling capacity.

Note also that regardless of the particular model selection method, strictly speaking, we cannot yet talk about parameter estimation in the internal constraints framework: No parametric model Θ_0 has been explicitly introduced so far. (Unlike in the external constraints case, due to its dependence on the data, F can not be identified with any fixed set Θ_0.) Hence, in order to properly analyze distributions generated by internal constraints (e.g., by making them suitable for testing with hypotheses of the form "$p \in \Theta_0$"), we have to embed them into some parametric family. By doing so, we effectively reformulate the problem relative to the external constraint framework. For instance, $\hat{p}_{MEE,V}$ above (4.3) is known (§4.3) to belong to the log-linear model family (4.1). It is then straightforward to verify that $\hat{p}_{MEE,V} = \hat{p}_{MLE,\Theta_0^{loglin}}$ provided the same V is used on both sides (Proposition A.6). Figure 6 illustrates geometrically the relation between the ICP/MEE and the log-linear ECP/MLE frameworks. The vertical line segment and the planar

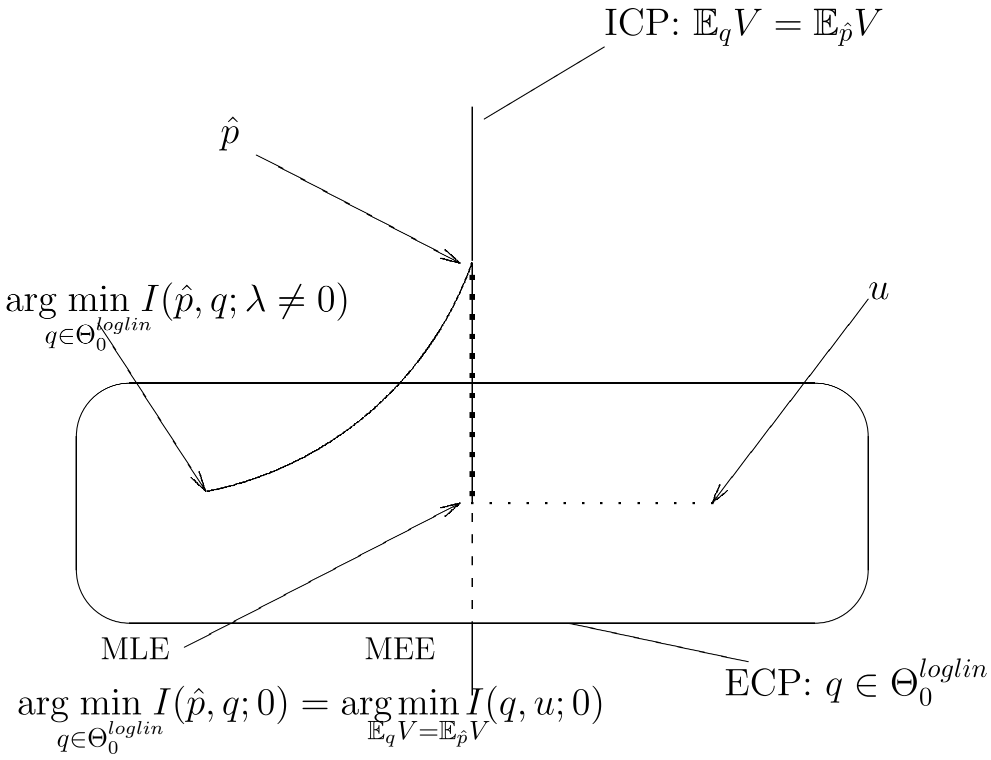

ICP: $\mathbb{E}_q V = \mathbb{E}_{\hat{p}} V$
$\hat{p}$
$\arg\min_{q\in\Theta_0^{loglin}} I(\hat{p}, q; \lambda \neq 0)$
u
MLE
MEE
ECP: $q \in \Theta_0^{loglin}$
$\arg\min_{q\in\Theta_0^{loglin}} I(\hat{p}, q; 0) = \arg\min_{\mathbb{E}_q V=\mathbb{E}_{\hat{p}} V} I(q, u; 0)$

Figure 6: Correspondence between parameter estimation in the ICP and ECP frameworks

domain represent the ICP and ECP (i.e., the resulting log-linear family) feasible families respectively. The unique intersection point corresponds to the MEE (from the ICP perspective) and the MLE (from the ECP perspective). The former is also the minimizer of the Kullback-Leibler "distance" from the the ICP segment to the uniform distribution u and the latter minimizes the same "distance" but from the empirical distribution $\hat{p}$ to the log-linear domain.

Other model selection mechanisms (e.g., constrained minimization of $I(\cdot, u; \lambda \neq 0)$) lead to different parametric models. Although transformation of internal constraints to the external constraints framework is in principle conceivable with any reasonable model selection mechanism, the resulting parametric family Θ_0 need not necessarily lead to computational gains in estimations.

In summary, for the purpose of testing hypotheses corresponding to models generated via ICP/MEE, and also in order to analyze model complexity in these cases, we will always consider these models from the ECP/MLE viewpoint.

Finally, independent of the modeling approach, by a "model" we will generally mean the appropriate parametric family Θ_0, rather than its particular member estimated from the data. The corresponding notation will be of the form: $\{p^{type}_{class}(\omega;\gamma),\ \gamma \in \Gamma\} = \Theta_0$. For example, we will write p^{g}_{symm} to refer to the model family whose members respect all geometric symmetries of the microimage set Ω. Also, parameters γ as well as their space Γ are often kept implicit, especially in this chapter, and additional relevant discussions follow in Chapter 6.

4.3 Maximum Entropy Extension

Since we consider several models in the context of ICP with model selection based on maximization of H, Shannon's entropy (Definition A.1), we now give a brief introduction to the Maximum Entropy Extension Principle (MEE) ([11],[40],[47],[67],[70]) in the finite case. We use our microimage set Ω as a prototype for any finite set.

Let μ be a (fixed) positive probability mass function on Ω[2], i.e. $\mu \in \Theta^+$, the set of all positive probability distributions on Ω. Also, let $V_0, V_1, \ldots, V_J \in \mathbb{R}^K$ be a set of linearly independent real functions on Ω, with $V_{0,1} = \cdots = V_{0,K} = 1$. Note that $V_{jk} \equiv V_j(\omega_k)$. There is no loss of generality in assuming linear independence: The excluded trivial cases of redundant or inconsistent constraint equations (the right hand side of (4.5) below) are not interesting in practice. We also write V for the $(J+1) \times K$ matrix with rows V_j for $j = 0, \ldots, J$. Hence, $\operatorname{rank}(V) = J + 1 \leq K$.

When $q \in \Theta$, we write interchangeably Vq (matrix multiplication) and $\mathbb{E}_q V$ referring to the left hand side of the internal constraints $\mathbb{E}_q V = \mathbb{E}_\mu V$ that define the feasible region

$$F(\mu) \overset{\text{def}}{=} \{q \in \Theta^+ : \ \mathbb{E}_q V_j = \mathbb{E}_\mu V_j, \ j = 1, \ldots J\} = \{q \in Q^+ : \ Vq = V\mu\}, \qquad (4.4)$$

just as in §4.2.2.

Thus, we think of V also as a $J+1$-dimensional random vector on (Ω, q) for any $q \in \Theta$. Generally, we shall assume the constraints are *consistent*, namely, the feasible family $F(\mu)$ (4.4) is nonempty. Note that we enforce this condition by having the right hand side of the constraining equations (4.4) arise as the expectation of V under a probability vector.

We say that a probability vector q^* is the maximal entropy extension of μ constrained by V if

$$q^* = \arg \max_{q \in F(\mu)} H(q), \qquad (4.5)$$

[2] Recall (§4.2) that we restrict all our modeling to strictly positive probability distributions.

In the finite case the solution always exists and is unique. (The existence can be demonstrated constructively by using, for example, *iterative proportional fitting* algorithm [54], and the uniqueness follows from the concavity of H [11],[54].) In a slightly more general statement of the problem, the condition $\mu \in \Theta^+$ is relaxed to $\mu \in \Theta$, and Q^+ in (4.5) is replaced by the non-negative quadrant, thus allowing solutions on $\partial\Theta_0$ (the boundary of Θ_0). The problem is still well-posed, although when the maximum is attained on the boundary, it does not have an exponential (i.e. *Gibbs*) form (4.6). However, the positivity of μ - as assumed here - implies that the solution to (4.5) always belongs to the interior of Θ (see, for example, [47]). The method of *Lagrange Multipliers* is then generally used to derive the general form of q^*:

$$q_k^* = \exp\left(\gamma_0 - 1 + \sum_{j=1}^{J} \gamma_j V_{jk}\right), \tag{4.6}$$

where the parameters $\gamma_0, \ldots, \gamma_J$ are uniquely determined by $\mathbb{E}_{q^*(\gamma_0,\ldots,\gamma_J)}V = \mathbb{E}_\mu V$. Except in special simple cases (e.g., §4.5), these equations require numerical methods.

4.4 Symmetry

Abstract Algebra via the *Group Action* formalism (see, for example, [18]) provides the most natural approach to analysis and modeling of symmetry. In our context of modeling the microimage distribution p on a finite set Ω, symmetries (i.e. transforms of Ω into itself) will be examined from the perspective of identifying those respected by p. Namely, we will search for symmetries under which symmetric states (microimages mapped to each other by the symmetry transform) are also p-equiprobable. Clearly, such symmetries should reduce the number of parameters in the representation of p. With no constraints[3], there are, of course, $|\Omega| - 1$ free parameters. Assuming a particular set of symmetries (external constraints), this number reduces to one less than the number of p-constancy classes under these symmetries. Naturally, we identify symmetric states into *equivalence classes*, p being constant on these classes. Given a set of symmetry transformation, we denote the set of all equivalence classes by $\mathcal{S}$ and individual classes by $\mathcal{O}$, just as we did in §3.4. We will also say that p is *invariant* under the given symmetry transforms, referring to the constancy of p on all $\mathcal{O} \in \mathcal{S}$.

Evidently, all symmetry models can be defined by sets of linear homogeneous equations of special form (6.1), (6.2), whose solutions are exactly probability distributions invariant under given symmetry transformations. Symbolically, the corresponding model families can be represented as:

$$\Theta_0 = \{q \in \Theta^+ : \; i, j \in \mathcal{O} \in \mathcal{S} \Rightarrow q_i = q_j\}. \tag{4.7}$$

Apart from statistical testing of the constancy of p on the symmetric states, an important computational task is to calculate the number and the sizes of the equivalence classes corresponding to the symmetries. Some of the computations needed to define our models can, in principle, be carried out without the group-theoretic apparatus; for others, using this elegant machinery is essential. We will try to employ only the necessary minimum from this theory. Although we will adopt the group-theoretic language in most of the work relevant to symmetries, most of the algebraic computations are moved to Appendix D. This decision reflects favoring generality of the audience over generality of the methods. Namely, despite the existence of very general and elegant (albeit rather advanced for a non-mathematician) computational methods, we will use longer, but more intuitive techniques. Also,

[3] Strictly speaking, with the normalization constraint alone.

relatively short computations are presented fully in Chapter 6. In this Chapter we continue to discuss symmetries through the basic notion of *equivalence relation*. We already used geometric and intensity inversion symmetries in §3.4.1. There, we simply aggregated patches into equivalence classes in order to make the stability analysis more reliable. Consequently, the discussion moved from p as defined on Ω to the integrated distribution induced on the coarser space of the equivalence classes. In contrast, the state space here is always Ω (instantiated primarily to the 2×2 matrices).

4.4.1 Geometry

Similar to the definitions of rotation and reflection from §3.4.1, we denote by p^h_{symm} and p^v_{symm} the families of distributions invariant under reflections across the vertical and horizontal axes respectively. We will also refer to these symmetries as “Left-Right” and “Up-Down” respectively. Thus, for instance, if Ω is the set of horizontal pixel pairs, then $p^h_{symm}(a,b) = p^h_{symm}(b,a)$ for any 1×2 patch $(a,b) \in \Omega$ (see also §5.2). These are the most fundamental and intuitive constraints we believe to be satisfied by p: It is hard to imagine a physical source that would cause, for instance, the left pixel in the pair to be consistently brighter than its right neighbor. It is also easy to imagine that the global version of “Left-Right” reflection also holds: $\mathbb{P}_{im}(I) \approx \mathbb{P}_{im}(I_{h-\text{reflected}})$. Attributes of natural scenes such as the “Blue Sky Effect” [50] surely inject some doubt about “Up-Down” symmetry, which is obviously not satisfied by (macro) images. Less clear, perhaps, is to what extent the “Up-Down” symmetry stands at the micro level. Testing the respective hypotheses (Chapter 5) will provide an answer to these questions.

In the case of 2×2 patches (assumed for the rest of this subsection), it still makes sense to entertain one or both of the above reflections. Any distribution on Ω that simultaneously respects these two symmetries will be denoted by p^{vh}_{symm}. Note, that *transitivity* of the symmetry equivalence relation forces p^{vh}_{symm} to be also invariant under rotations by π. Next, "r" will stand for the rotation by $\pi/2$. Easily verified and more clearly seen in the group-theoretic context (Appendix D), the above rotation along with reflection through any of the four axes of 2×2 patch generates one and the same set of symmetries, naturally referred to as *geometric*: These are all the geometric symmetries of the square, and p^{g}_{symm} will stand for the corresponding invariant distributions (see Equation (3.3) for a visual example).

4.4.2 Photometry

In §3.4.1 we also aggregated patches with their *negatives*, referring to the corresponding symmetry of Ω as *intensity inversion*. A distribution on Ω that assigns equal probability to states related in this way, will bear the superscript "n". Unlike the geometric symmetries, inversion is a *photometric* property and is also meaningful for the single pixel distributions. Again, the frequent, heavy presence of the sky, as well as simply "blank" areas in most of the outdoor scenes, makes the invariance of p under the intensity negation rather questionable. Moreover, the single pixel statistics suggest a violation of this symmetry [34]. We will nonetheless consider this invariance mode and run statistical tests to quantify it. In fact, the decision to equalize (of course, only approximately) the single pixel histograms was also motivated by our attempt to give the inversion symmetry "a second chance".

As already mentioned in §3.4, we will also consider models respecting all of the above symmetries. Even if this hypothesis is rejected at common levels, the reduction in model complexity should be acknowledged. We will write p^{G}_{symm} for such distributions, where the superscript simply stands for the set (more precisely *group*, see Appendix D) *generated* by all the geometric and the inversion symmetries.

These are geometric symmetries of Ω as the square-based *parallelepiped*, the patch being associated with the base.

Intermediate models will also emerge wherein some proper subsets (*subgroups*) of the G-symmetries are respected. For example, p^{vhn}_{symm} respects the Up-Down, Left-Right, and the inversion symmetries, but need not respect the rotation by $\pi/2$. Note the reverse correspondence between the set inclusion of the symmetries and that of the respective invariant distributions (see Figure 7).

4.5 Dominant Mass Constraints

It may be sensible to model highly non-uniform distributions simply by smoothing rare events (e.g. §3.3). Namely, if a set $D_J \overset{\text{def}}{=} \{\omega_{k_1}, \ldots, \omega_{k_J}\}$ of J states (J is in practice relatively small) accounts for most of the mass (determined from preliminary experimentation), then such states are represented precisely by their respective frequencies, whereas all other states are averaged. Where exactly to draw the line between the "distinguished" states and the averaged tail $T_J \overset{\text{def}}{=} \Omega \setminus D_J$ is, of course, application dependent. Such models can be viewed from both the ICP and ECP viewpoints (more on the ICP/ECP duality is in Chapter 6).

When viewed as an ICP/MEE, a dominant mass model derives from a maximum entropy extension problem (§4.3) which has, in addition to the normalization constraint function $V_0(\omega) \equiv 1$, the constraint functions: $V_j(\omega) = \mathbb{I}_{\{\omega=\omega_{k_j}\}}$ for $j = 1, \ldots, J < K$. Clearly, the privileged states D_J will then inherit their probabilities from μ, i.e. $q^*(\omega_{k_j}) = \mu(\omega_{k_j})$ for $j = 1, \ldots, J$. In this special case, the original MEE problem transforms into unconstrained entropy maximization relative to T_J, the set of the remaining states. Namely, we now maximize $H(q)$, understanding by q the conditional distribution $q(\omega|T_J)$ restricted to T_J. Recalling the well-known fact that in the finite case entropy is maximized by the uniform distribution, we arrive at:

$$q^*(\omega) \quad = \quad \exp\left(\sum_{j=1}^{J} \log \mu(\omega_{k_j}) \mathbb{I}_{\{\omega=\omega_{k_j}\}}(\omega) + \log \frac{\mu(T_J)}{K-J} \mathbb{I}_{T_J}(\omega)\right)$$

Thus, in this case the Gibbs form (4.6) has the following parameters: $\gamma_0 = 1 + \log \frac{\mu(T_J)}{K-J}$ and $\gamma_j = \log \mu(\omega_{k_j}) + 1 - \gamma_0 = \log \frac{\mu(\omega_{k_j})(K-J)}{\mu(T_J)}$ for $j = 1, \ldots, J$. Of course, there is no computational benefit in expressing the solution in this form. However, in order to test these models, we need to view them in the ECP framework via their equivalent log-linear representations. One such representation is based on the choice of parameters γ above:

$$p_{dom}(\omega;\gamma) \quad \overset{\text{def}}{=} \quad e^{\gamma_0 - 1 + \sum\limits_{j=1}^{J} \gamma_j \mathbb{I}_{\{\omega=\omega_{k_j}\}}(\omega)} \tag{4.8}$$

Evidently, the dimension of the corresponding model space Θ_0 is $d_0 = J$.

Note also that Lagrange multipliers are found trivially in this case due to *orthogonality* of $V_1, \ldots, V_J$. In Chapter 6, when we discuss duality between ECP and ICP, we will present a more general and illustrative example of this situation (Proposition 6.1).

This model can be slightly modified to produce $p_{dom\ symm}$, a "symmetry"-based model with external constraints enforcing constancy of $p_{dom\ symm}$ on the tail set T_J. The quotes above are due to that there do not appear any natural symmetry-like transformations of Ω that would induce the corresponding partition $\mathcal{S} \overset{\text{def}}{=} \{\{\omega_{k_1}\}, \ldots, \{\omega_{k_J}\}, T_J\}$.

The corresponding model family in this case is a specialization of (4.7):

$$\Theta_0^{dom} = \{q \in \Theta^+ : \ \omega_i, \omega_j \in T_J \Rightarrow q_i = q_j\},$$

and for the purpose of hypothesis testing this model will be associated with the rest of our symmetry models. The relation between p_{dom} and $p_{dom\ symm}$ is a special case of $\hat{p}_{MEE,V} = \hat{p}_{MLE,\Theta_0^{loglin}}$ (§4.2.2 and Proposition A.6) and will become clear in the context of §§6.1,6.2.

Mostly for illustration, we will test one such model, $p_{dom\ symm}$, in which the rarest states T_J will absorb about 10% of the total mass. To simplify the notation, however, we are still going to refer to this model as p_{dom}.

In general, resource limitations may directly influence control over the model complexity: An upper bound on the number of the original states distinctly represented in a dominant mass model is one such example. However, aggregating microimages based only on their rareness in the current sample may endanger further insight into the mechanism of their formation, and is certainly unnatural in a perceptual sense. Imagine that among T_J, the states with smoothed probabilities, can be representatives of perceptually very different patterns (e.g. "edge-lets", "T-junctions"). On the other hand, two "edge-lets" of the same orientation but with different contrasts may well get separated: The one with the higher gradient (and consequently lower probability mass) is likely to be smoothed (i.e., appears in T_J), whereas the other is likely to stay in D_J. Thus note, that the model deficiency may come not so much from averaging the "light" representatives of disparate classes, but more from *not combining them with their more frequent "true", or "natural", associates*. This may occur due to that, in a practical situation, having to distinguish between two rare states (i.e., belonging to T_J) may be less likely than facing a decision involving two states on the opposite sides of the "rareness" demarcation (i.e., D_J versus T_J). Of course, this also depends on the particular visual selection strategy.

Primitive dominant mass models may greatly benefit from additional constraints such as symmetries or more general relations. Depending on the type of these constraints, the proper framework may be that of ECP or ICP.

We experimented earlier [22] with *Alternate Quantization* (Chapter 2) combined with the geometric symmetries (§4.4.1). (Some of these results are reported in [22] in the context of stability analysis.) We also considered dominant mass models in those experiments. Taking into account a very small number of parameters (e.g. less than ten in the case of $|\Omega| = 4096$), the results were encouraging based on such measures as Kullback-Leibler divergence, Shannon's entropy, and L_p norms. Formal hypothesis testing was not feasible then due to insufficient image data. Other examples of additional (internal) constraints included absolute intensity difference of pixel pairs, e.g., $V(\omega) =\mid \omega_{11} - \omega_{22} \mid + \mid \omega_{12} - \omega_{21} \mid$, where $\omega = \begin{smallmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{smallmatrix}$.

4.6 Potts Model

Motivated by the ideas from statistical physics, and also by visual examination of several empirical histograms for p, we consider a simple model related to the Potts potential [6]. The term "potential" is used in the sense of *Gibbs - Markov Random Fields* [17],[25],[26],[29],[38],[41],[56]. However, the notion of Gibbs-Markov potential is not critical for understanding any of our microimage models: Our presentation is largely self-contained. On the other hand, we will appeal to the general theory of Gibbs-Markov random fields in the discussion of extending our models to larger lattices in Chapter 7.

Under the Potts models, the microimage distribution is characterized completely by the number of pairwise matches among pixel intensities, and in particular is insensitive to absolute intensities. These models are originally introduced as maximum entropy extensions (§4.3) under internal constraints (§4.2.2).

Consider the following constraint functions:

$$\begin{aligned} V_1(\omega) &= \mathbb{I}_{\{\omega_{11}=\omega_{22}\}} + \mathbb{I}_{\{\omega_{12}=\omega_{21}\}} \\ V_2(\omega) &= \mathbb{I}_{\{\omega_{11}=\omega_{12}\}} + \mathbb{I}_{\{\omega_{21}=\omega_{22}\}} \\ V_3(\omega) &= \mathbb{I}_{\{\omega_{11}=\omega_{21}\}} + \mathbb{I}_{\{\omega_{12}=\omega_{22}\}}. \end{aligned} \tag{4.9}$$

Thus, for example, $V_1 q$ is the expected number of the diagonal pixel intensity matches relative to q. The solution to the corresponding entropy maximization problem is of the form:

$$p^{vhn}_{potts}(\omega) = C(\gamma_1, \gamma_2, \gamma_3) \exp\left(\gamma_1 V_1(\omega) + \gamma_2 V_2(\omega) + \gamma_3 V_3(\omega)\right), \tag{4.10}$$

where we have replaced the cumbersome "$\exp(\gamma_0 - 1)$" by "C", the multiplicative normalization constant whose dependence on $\gamma_1, \gamma_2, \ldots$ will often be suppressed. The "*vhn*"in the superscript refers to respecting the vertical and horizontal reflection and inversion symmetries (§4.4). Note that the symmetries of the distributions are inherited from the constraint functions V. This observation also suggests that in order to enforce all the geometric symmetries, including the $\pi/2$-rotation, we should simply replace the two last constraints by their sum $V_2'(\omega) \overset{\text{def}}{=} V_2(\omega) + V_3(\omega)$. Solving the MEE, we arrive at a more restrictive family defined below in (4.11); note that the distributions p^G_{potts} fully respect the G-symmetries (§4.4).

$$p^G_{potts}(\omega) = C \exp\left(\gamma_1 V_1(\omega) + \gamma_2 V_2'(\omega)\right). \tag{4.11}$$

An interesting question arises: How much more complex is the family p^G_{symm} (§§4.4,6.4.1) than p^G_{potts}? Note that by model complexity we simply mean the number of free parameters defining the corresponding family of distributions. Example 1 from §6.2 will show that the difference is minimal in the binary case (i.e. $L = 2$). Specifically, the four G-classes lead to the complexity of p^G_{symm} being three, whereas the parameter space of p^G_{potts} is two-dimensional. Also, since the complexity of p^G_{potts} does not depend on L and the complexity of p^G_{symm} is of the order L^4 (§6.4.1), the difference can grow arbitrarily large.

Next, let us denote by $\mathcal{S}$ the set of all V-constancy classes $\mathcal{O}$, namely: $\forall \mathcal{O} \in \mathcal{S}$, $\omega,\ \omega' \in \mathcal{O} \iff V(\omega) = V(\omega')$. Another interesting observation in Example 1 will be that in the binary case, the constraints V used to define p^G_{potts} have their constancy classes $\mathcal{O} \in \mathcal{S}$ exactly equal to the G-symmetric classes (§4.4), i.e. $V(\omega) = V(\omega') \iff \omega \overset{G}{\sim} \omega'$. On the other hand, we will see that these constraints are insufficient to represent all the G-invariant functions (and hence G-invariant probability vectors) on Ω. (One dimension will be missing.) This suggests a general way to refine simple models like p^G_{potts} and p^{vhn}_{potts} by extending the linear subspaces spanned by their constraint functions V in order for the extended subspace to include all the functions respecting the constancy classes of V. This is properly discussed in §6.2. The MEE solution will then gain as many new parameters γ as the number of the additional constraints. We will refer to the resulting log-linear Potts distributions as "complete". For example, for any L, p^{vhn}_{potts} will be shown to have nine constancy classes (§6.4.2) and only four (including the normalization) constraints. We will complete these four constraints so that the new V becomes a basis for the subspace of all the functions respecting the nine-class partition $\mathcal{S}$. It will be explained in §6.2 that in situations when functions in V form a basis for all the functions respecting some partition $\mathcal{S}$, the corresponding maximum entropy extension problem under the internal constraints given by V always admits an equivalent, "symmetry"-based formulation. This allows us to define p^{vhnC}_{potts}, "C" standing for "complete", based on this nine-class partition $\mathcal{S}$. The corresponding model space Θ_0 is thus of the same form (4.7) as in the case of the symmetry models. (However, unlike the geometric and inversion symmetry transformations, the "symmetry" transformations inducing this partition are implicit and not meaningful by themselves outside this context.) The p_{potts}^{vhnC} model will also be tested in Chapter 5 with the model construction details to follow in §§6.2,6.4.2.

4.7 Pairwise Interaction

The models p_{potts}^{vhn}, p_{potts}^{G}, and p_{potts}^{vhnC} from the previous section are examples of the more general class of the "pairwise interaction" (or "pair-potential") models. The *Ising* model [6],[25],[26],[29],[38] is a well-known member of this family. Recall that the constraints V_1, V_2, and V_3 used to define p_{potts}^{vhn} and p_{potts}^{G} are sums of functions of two variables only. This is the defining feature of the pairwise interaction models. More precisely, let $\mathcal{D}$ be the linear subspace of $\mathbb{R}^K$ spanned by all functions on Ω of (at most) two variables. For example, the constant functions, $\omega_{11} - \omega_{22}$ and $\mathbb{I}_{\{\omega_{11}=0\}}(\omega)$ all lie in this subspace[4]. In general, the dimension of this family (and therefore the model complexity) is of the order L^2 (§6.4.3). For comparison, recall that the complexity of the Potts-like models is independent of L, and the general symmetry models of §4.4 all have complexities of the order L^4.

The entropy maximization method (and thus, the ICP formulation) still applies here provided a basis for $\mathcal{D}$ is found to play the role of V. Selecting a particular member from this family may in principle be viewed within the equivalent log-linear ECP framework, except that parameter estimation is not as simple as in the case of the symmetric models (Chapter 6). Instead of the full pair-potential family we will focus on its proper subset, namely the distributions p_{pair}^{g} that also respect the geometric symmetries (§4.4.1). One reason for this is that we studied the geometric

[4] Visual inspection of a function-defining formula may sometimes be misleading in determining whether the function indeed belongs to $\mathcal{D}$.

symmetries first and their presence proved significant as will be demonstrated in Chapter 5. Also, combining constraints of different origins such as symmetry and pattern of interaction illustrates the power of the duality between the ICP and ECP frameworks in the context of parameter estimation (Chapter 6).

4.8 Summary

We have used *internal* and *external constraints* to define a variety of models for p, the natural microimage distribution. The proposed models are different in origin and complexity and encode our prior knowledge about p. In Chapter 5 we will test most of these models, and in Chapter 6 we will discuss computation and complexity. These models are classified into the following four categories:

- Dominant Mass: p_{dom}
- Potts-type: p^{vhn}_{potts}, p^{G}_{potts}, p^{vhnC}_{potts}
- Pair-potential: Potts models and p^{g}_{pair}
- Symmetry-based: p^{h}_{symm}, p^{v}_{symm}, p^{n}_{symm}, p^{vh}_{symm}, p^{vhn}_{symm}, p^{g}_{symm}, p^{G}_{symm}, and pairwise interaction models

Figure 7 symbolically depicts the principal symmetries entertained in our modeling. Higher vertices in the diagram correspond to more restrictive model families.

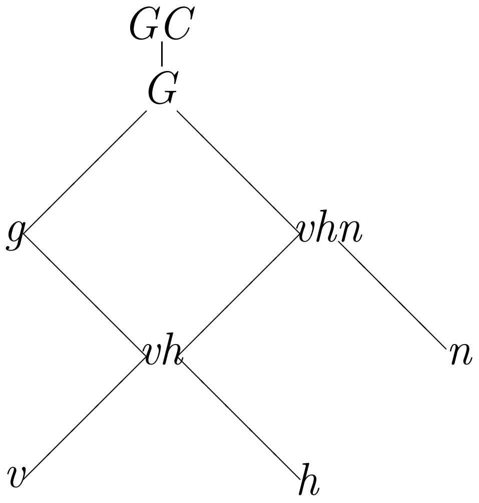


Figure 7: Inclusion relations on the symmetry sets. Higher vertices correspond to more restrictive models

CHAPTER 5

STATISTICAL TESTING OF MICROIMAGE MODELS

5.1 Introduction

The ultimate criteria for accepting a particular model should be at least partially application dependent. In particular, results of any statistical hypothesis testing, however accurate, should not be the only factor in model consideration. This opinion is commonly acknowledged: For example, based on generalized likelihood ratio tests, various modified information criteria such as AIC, BIC, and NIC have been suggested to control data-overfitting by adjusting a penalty for model complexity [46],[54],[56]. However, unless a class of potential applications is clearly specified, a universal criterion for resolving the bias-variance dilemma based on subjective penalizing of complexity would be of a limited value. This is one reason why we neither use the above criteria, nor design our own. We shall also try not to allow our subjective notions of "mathematical elegance" to influence model evaluation. Instead, we will simply try several existing techniques for testing discrete data models [28],[40],[54],[61],[62]. The rationale is that the test results should provide a general filter for further, application-oriented studies (Chapters 2,7).

In exchange for the deferment of a final "accept-reject" judgment of most of our non-trivial models, we present a careful examination of evidence based on our image data. Namely, we will discuss the "pros" and "cons" of these models based both on test results and on other complexity considerations. In fact, we have already attempted to order some of the symmetry models by set-inclusion of the corresponding model families (Figure 7). In this regard, the statistical tests based on the power-divergence measures (Appendix B), that we also used in Chapter 3, are particularly convenient: For example, two such tests, the generalized likelihood ratio test and the modified likelihood ratio test ($\lambda = 0$ and $\lambda = -1$, respectively) are related to information theory, and consequently provide a clear interface between engineering and statistics. These tests, when applied to our models, also allow relatively simple estimation of model parameters (Chapter 6 and Appendix B).

We consider how model definitions translate into statistical hypotheses and introduce the necessary notation. Then, we test two symmetric models, p^h_{symm} and p^v_{symm}, in the 1×2 and 2×1 cases respectively. The rest of this chapter is devoted to testing 2×2 models and interpreting test results.

Recall that we have assumed that p is strictly positive on all of Ω. Thus, our saturated model[1] is $\Theta^+ \overset{\text{def}}{=} Q^+ \cap \Theta$, the interior of the probability simplex Θ (§4.1). This assumption is enforced by the aggregation of rare states of Ω. We will perform the following two types of hypothesis testing:

- "Absolute": $H_0 : \ p \in \Theta_0$ and $H_a : \ p \in \Theta^+$, where $\Theta_0 \subset \Theta^+$ and $\dim \Theta_0 < \dim \Theta^+$
- Hierarchical (or nested): $H_0 : \ p \in \Theta_0$ and $H_a : \ p \in \Theta_a$, where $\Theta_0 \subset \Theta_a \subset \Theta^+$ and $\dim \Theta_0 < \dim \Theta_a < \dim \Theta^+$

Additionally, it is always the case in our testing that $d_0 > 0$, hence, prior to evaluating test statistics, we need to estimate parameters. The two situations only

[1] Recall that saturation refers to the dimension of the model being equal to that of the embedding probability simplex.

differ by alternative hypotheses: In the first case, the alternative is "absolutely unconstrained" and, in the second, it is nested strictly in between the null model and the saturated model. *The same power-divergence tests apply in both cases.* Most of our tests are from the first category and only in §5.3.1, where we assess the relative merit of nested models, do we use the second setting. In this latter case, we also refer to the alternative models Θ_a as less restrictive relative to the more restrictive null models.

Generally, the particular parameter estimation (model selection) method may suggest $f_0 : \Gamma \to \Theta_0$ and $f_a : \Gamma \to \Theta_a$, some convenient parametrizations of Θ_0 and Θ_a. Note that since Ω is finite, the positivity condition ensures that all our models[2] can be viewed as members of the exponential family (e.g., [46],[61],[62]). Thus, for example, the saturated family, Θ^+, admits the following trivial exponential parametrization:

$$\begin{aligned}\Gamma &= \{\gamma \in \mathbb{R}^K : \ \gamma_k < 0, \ k = 1, \dots, K, \ \sum_{k=1}^{K} e^{\sum\limits_{j=1}^{K} \gamma_j \delta_{jk}} = 1\} \\ \Theta^+ &= \{\theta \in \mathbb{R}^K : \ \theta_k = e^{\sum\limits_{j=1}^{K} \gamma_j \delta_{jk}}, \ \gamma \in \Gamma\}, \qquad (5.1)\end{aligned}$$

where the parameters γ_k are simply the logarithms of the corresponding probabilities, and δ_{jk} is the Kronecker symbol. Note also that $\dim \Theta^+ = \dim \Theta = |\Omega| - 1 = K - 1 = \dim \Gamma$.

The external constraint frameworks (ECP) considered in §4.2.1 naturally connects to this testing program: The null hypothesis always corresponds to the constrained (feasible) family Θ_0 and, in the absolute testing, the test statistics $T(\hat{p}, \hat{p}'; \lambda) \overset{\text{def}}{=} 2N_\omega I(\hat{p}, \hat{p}'; \lambda)$ (N_ω is the microimage sample size and $I(p, q; \lambda)$ was defined in (3.4)) provide distance-like measures of closeness between the empiri-

[2] In the case of symmetry constraints, the positivity condition is added explicitly (4.7).

cal distribution $\hat{p}$ and the appropriate (i.e., estimated) feasible distribution $\hat{p}'$ determined from the data. In the nested testing, the statistics $T(\lambda)$ also measure closeness but now between the appropriate member of the less restrictive model (i.e. corresponding to the alternative hypothesis) and $\hat{p}'$, the "best"feasible distribution (as in the absolute testing). From now on we omit reference to obvious extensions of ensuing statements to the nested testing. It is noteworthy that $T(0)$ =*const*$\times N_\omega \times$*Kullback-Leibler divergence* from $\hat{p}$, the empirical distribution, to $\hat{p}'$, the appropriately estimated feasible distribution (Appendix B), and $T(1)$ is the classical Pearson's χ^2 *goodness-of-fit* statistic.

Each of these power-divergence tests (or the statistics $T(\lambda)$) gives rise to a distinct parameter estimation issue, i.e., selection of a particular member from the constrained family Θ_0. Thus, to each λ there corresponds a *minimum power-divergence estimator* defined as a minimizer of the $T(\lambda)$ "distance" from the empirical distribution:

$$\hat{p}(\lambda) \overset{\text{def}}{=} \arg\min_{q \in \Theta_0} T(\hat{p}, q; \lambda), \tag{5.2}$$

and the minimization is generally performed relative to the appropriate parameter space Γ: $\hat{p}(\lambda) = \arg\min_{\gamma \in \Gamma} T(\hat{p}, q(\gamma); \lambda)$. However, the appropriate theory ([54] and Appendix B) states the asymptotic equivalence of all minimum power-divergence estimators under some general conditions. The *maximum likelihood estimator* (MLE) $\hat{p}(0)$ is, perhaps, the most famous such estimator, and it corresponds to the generalized maximum likelihood ratio test ($\lambda = 0$). In [54], the authors appeal to common sense suggesting to use with $T(\lambda)$ its true minimizer, but "if only the MLE ... is readily available", not to hesitate using it with other power-divergence statistics. Depending on the constraints, computation of other estimators may be more involved than the MLE, which partly explains why MLE is commonly used in practice even with tests arising from $\lambda \neq 0$.

The general symmetry models (p_{symm}) and the "completed" Potts model (p_{potts}^{vhnC}) are naturally defined within the ECP framework and the appropriate feasible family (4.7) is best understood in terms of the linear equations forcing constancy of probability vectors on symmetry classes (§6.1). Since under all such models the probability distributions are constant on the orbits $\mathcal{O} \in \mathcal{S}$, these models are completely characterized by the orbit probability masses. Thus, a natural choice for the model parameter space is $\Gamma = \Theta_{\mathcal{S}}^{+}$, the set of all positive probability distributions on the quotient space $\mathcal{S}$. Γ is then mapped by f_0 to Θ_0 by dividing the orbit masses $\gamma(\mathcal{O})$ uniformly among all $\omega \in \mathcal{O}$. A detailed explanation follows in §6.1. For now, however, it suffices to know that in these cases we can easily compute the "genuine" estimators for each of our tests §B.1. ("Genuine" simply means that, for each λ, the estimator of the free parameters is the true minimizer of the corresponding power-divergence statistic.)

The dominant mass model (p_{dom} §4.5), although initially introduced as a maximum entropy extension (MEE) in the internal constraint (ICP) framework, was then modified to a similar model but with external constraints (by imposing equality on the rare probabilities). Thus, as a special case of the symmetry models, this model is also ready for testing with minimum power-divergence tests and their genuine estimators (§B.1).

As mentioned in §4.2.2, models, such as p_{potts} and p_{pair}, originally constructed as the maximum entropy distributions under internal constraints of the form $\mathbb{E}_q V = \mathbb{E}_{\hat{p}} V$, can also be viewed from the ECP perspective based on $\hat{p}_{MEE,V} = \hat{p}_{MLE,\Theta_0^{loglin}}$ (Proposition A.6), where Θ_0^{loglin} is defined by (4.1). Recall (§4.3) that the constrained maximum entropy extension is generally of the exponential form (4.6) and it is through this exponential form that any MEE with internal constraints relates to *log-linear* modeling. Because the ICP framework was shown (§4.2.2) as not immediately suitable for hypothesis testing, we agreed to formulate the appropriate model hypotheses relative to the equivalent log-linear families. Correspondingly, the constrained MLE is used with p_{potts} and p_{pair} even with the test statistics $T(\lambda \neq 0)$. Figure 6 also illustrates that minimum-power-divergence estimators other than the MLE ($= \arg \min\limits_{q \in \Theta_0^{loglin}} T(\hat{p}, q; 0)$) generally violate the internal constraints $\mathbb{E}_q V = \mathbb{E}_{\hat{p}} V$. This is another reason, in addition to the immediate availability of the MLE (as the numerical solution to the MEE ICP), why we use the MLE instead of the "genuine" estimators with tests other than the generalized maximum likelihood ratio one ($\lambda = 0$) for assessment of the p_{potts} and p_{pair} models.

Generally, in this work:

- *When tested with power-divergence tests, all models are considered within the ECP framework.*
- *Analytical or numerical, parameter estimation is not an issue in any of our models.*

The theory of power-divergence tests ([54] and Appendix B) also provides under some general regularity conditions satisfied in our setting an asymptotic distribution for the test statistics $T(\lambda)$. It turns out that this asymptotic distribution is the χ^2 distribution, remarkably independent of the particular test, i.e., the same for all λ (Appendix B) and for all minimum power-divergence estimators; the number of degrees of freedom is the difference between the dimensions of Θ^+ (or Θ_a in nested testing) and Θ_0, the ground (or less restrictive) family and (more restrictive) model family respectively. Whereas validity of the large sample (i.e. asymptotic) results is always a question in practice, some general guidelines are available; for example, a detailed treatment of these tests appears in [54]. We have already followed two such guidelines in testing the stability of p (Chapter 3) and will continue to do so in this chapter: First, we have committed ourselves to controlling low counts by appropriate aggregations. Second, we have performed a series of tests with different values of the parameter λ aiming to achieve inter-test consistency relative to the asymptotic significance threshold (which is consequently also the same for all λ).

5.2 Bivariate Reflections

We first consider evidence for reflection (left-right and up-down) symmetries in the case of two pixel distributions (1×2 and 2×1 with $L = 8$). These experiments have served to calibrate our testing procedures in regard to the independence assumptions and other issues. There has already been some evidence of symmetry in other work. For example, Huang and Mumford have repeatedly demonstrated on large sets of natural images [34],[35] that the difference of two adjacent intensities has a symmetric distribution (i.e., the corresponding density function is even). In contrast to the symmetry of the univariate difference distribution, our claims involve the bivariate intensity distribution and are therefore stronger. Finally, in [33], results are reported in consonance with ours.

For two configurations $\omega = (a, b)$ and $\omega' = (a', b')$, $a, b, a', b' \in \{0, \dots, 7\}$, we write $\omega \overset{\text{refl.}}{\sim} \omega'$ if $a = b'$ and $b = a'$ (§4.4.1). It is easy to see that there are $\frac{L(L-1)}{2}$ reflection symmetry classes consisting of two configurations (i.e. $a \neq b$) and L *singular* classes with only one configuration each. With $L = 8$, we therefore have $\dim \Theta_0 = 28 + 8 - 1 = 35$ and $\dim \Theta^+ = 63$, subtracting one degree of freedom for normalization. Thus, in testing the null hypothesis

$$H_0: \ p \in \Theta_0 = \{q \in \Theta^+ : \omega \overset{\text{refl.}}{\sim} \omega' \Rightarrow q(\omega) = q(\omega')\}, \tag{5.3}$$

we use the χ^2 distribution with $63 - 35 = 28$ degrees of freedom in order to obtain the asymptotic P-values. The test statistic is $T(\lambda) = 2N_\omega I(\hat{p}, \hat{p}(\lambda); \lambda)$, where the microimage sample size is $N_\omega \approx 15,000$ and $I(p, q; \lambda)$ was defined in (3.4). Note that $I(\hat{p}, \hat{p}(\lambda); 0) = \log 2D(\hat{p}, \hat{p}(\lambda))$, where $D(\cdot, \cdot)$ is the Kullback-Leibler divergence.

Values of $T(\lambda)$, as well as corresponding P-values, are tabulated in Table 8 for several values of the test parameter λ. *Based on the asymptotic threshold ($\chi^2_{0.95,28} =$* 41.337*) we do not reject the null.* This result is also strengthened by the apparent consistency of the $T(\lambda)$'s for the seven values of λ: Since varying λ allows the corresponding test statistics $T(\lambda)$ to peak its sensitivity at specific departures from the null, the observed consistency appears to protect us against a range of such departures [54].

λ	-2	-1	0	0.5	0.67	1	2
$T(\lambda)$	33.278	33.868	33.624	33.215	33.043	32.676	31.396
$P\text{-}val(\lambda)$	0.2258	0.2053	0.2135	0.2279	0.2342	0.2479	0.2998
$T(\lambda)$	23.347	23.746	23.586	23.319	23.21	22.98	22.216
$P\text{-}val(\lambda)$	0.7155	0.6949	0.7032	0.7169	0.7225	0.7341	0.7712

Table 8: Power-divergence tests for bivariate reflection symmetries. Top: Up-Down. Bottom: Left-Right.

5.3 Testing Models in the 2×2 Case

Preliminary experiments revealed that additional aggregation was necessary to further protect against low counts and to stabilize the power-divergence test statistics as function of λ. Hence, we increase the size of the "rare" class from 136 patches of total mass of about 0.025 (as in §3.7) to 184 patches of total mass around 0.045. Unlike in §3.7, this aggregation fully respects the G-partition[3]. Table 16 (Appendix C) provides the adjusted model complexities. Aggregation decreases the dimensions of the parameter spaces. Thus, for example, by aggregating 184 patches we reduce the dimension of the ground space from 255 to 72. Consequently, the degrees of freedom of the asymptotic distributions are also affected; all values are given in Appendix C. To facilitate comparisons, all these tests were performed on the same microimage sample, which consists of about $11,500$ microimages (extracted at the rate $d = 0.0002$; d is discussed in §C.2).

Now, we present test results for some of the models defined in Chapter 4. Appendix C provides a complete account of all of our test results. We can classify these results as follows:

- Strong Rejection: p_{dom}, p^{G}_{potts}, p^{vhn}_{potts}, p^{vhnC}_{potts}
- Rejection: p^{n}_{symm}, p^{vhn}_{symm}, p^{G}_{symm}
- Non-rejection: p^{v}_{symm}, p^{h}_{symm}, p^{vh}_{symm}, p^{g}_{symm}, p^{g}_{pair}

Models fall into the "Strong Rejection" category if the corresponding tests consistently yield P-values virtually equal to 0. Table 9 displays two such examples:

$$\begin{aligned}\hat{p}^{G}_{potts}(\omega) &\approx 11671\exp\left[-0.26V_1(\omega) - 1.77V_2'(\omega)\right]\\ V_1 &= \mathbb{I}_{\{\omega_{11}=\omega_{22}\}} + \mathbb{I}_{\{\omega_{12}=\omega_{21}\}}\\ V_2' &= \mathbb{I}_{\{\omega_{11}=\omega_{12}\}} + \mathbb{I}_{\{\omega_{21}=\omega_{22}\}} + \mathbb{I}_{\{\omega_{11}=\omega_{21}\}} + \mathbb{I}_{\{\omega_{12}=\omega_{22}\}}\end{aligned}$$

and p^{vhnC}_{potts} (see §4.6).

[3] This means that none of our symmetry models will have a class divided in two by the aggregation: The G-partition is the coarsest among all the partitions induced by the p_{symm} models.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	4799.4	3447.3	3161.5	3073	3195.3	3297.5	3618.6	6783.6
$T(\lambda)$	1032.4	1036.5	964.74	872.8	787.97	762.85	719.88	629.82

Table 9: Strong rejection of primitive models. Top: p^{G}_{potts}, $\chi^2_{0.99,70} = 100.43$. Bottom: p^{vhnC}_{potts}, $\chi^2_{0.99,66} = 95.626$.

Note that the tests take into account both the complexity and accuracy of the model. In this regard, the dominant mass model is actually relatively accurate (as measured by the power-divergence measures) but still rather complex, whereas the Potts models have very low complexity but insufficient accuracy. Not surprisingly, the p_{dom} captures the entropy of p very well: Both are estimated to be approximately 4 bits. On the other hand, the Potts models p^{G}_{potts} and p^{vhn}_{potts} overestimate the entropy of p by about half a bit, although the entropy of the "completed" Potts model p^{vhnC}_{potts} (this model has eight free parameters) is very close to four bits. Still, the refinement of p^{vhn}_{potts} by completion (§4.6) proved insufficient: The gain in accuracy is not justified by the increase of complexity. *Since all the variations of the Potts model are strongly rejected, we conclude that information induced only from equality comparisons is insufficient and some information about absolute intensities is necessary.*

The next category represents models that, despite being rejected at the common 5%-significance level, are not totally unreasonable in the sense of yielding P-values at least on the order of 10^{-3} (see Table 10). In the next section (§5.3.1) these models will be further tested in the context of hierarchical modeling in order to assess their *relative* merit. Note that the intensity inversion is part of all these models as well as of the strongly rejected ones, suggesting its unlikely presence in the natural microworld.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	97.873	91.108	88.934	87.523	86.662	86.453	86.134	85.522
$P - val(\lambda)$	0.0025	0.0094	0.0141	0.0181	0.021	0.0218	0.023	0.0257

Table 10: The G-Symmetric model comes close to being not rejected. $\chi^2_{0.99,62} = 90.802$.

The last category are the models which stand very strongly at common significance levels. Recall that p^g_{pair} is the pair-potential model respecting all the geometric symmetries and thus is the most restrictive of these models. Without aggregation the dimension of this family is 42, whereas the dimension of the most general geometrically symmetric model is 54. Unfortunately, our aggregation cancels the difference rendering p^g_{symm} and p^g_{pair} virtually indistinguishable under the current testing regime. This is how it happens: The 36 rarest among the 55 total g-symmetric classes are aggregated into one. Thus, a 20-dimensional linear space is required to represent all the functions on the resulting set. It turns out that functions of fewer than three variables already span this subspace. Apart from this limitation, the fact that the geometrically symmetric states capable of interacting with degrees higher than two are so rare (less than 0.05) leads us to believe that the three- and four-pixel interactions may generally be insignificant in the presence of all pair-wise interactions. Also, $L = 4$ is too small for the asymptotic complexity comparison to be meaningful: Recall (§4.7) that the complexity of the pairwise interaction models grows only as L^2, whereas for all the symmetry models the growth rate is L^4, the order of $|\Omega|$. (No aggregation is assumed for the moment.) However, for example, the most general G-symmetric model, p^G_{symm}, now has only 30 free parameters (Proposition 6.8 from §6.4.1), which is significantly smaller than 54, the complexity of p^g_{pair} (determined numerically as mentioned in §§6.3,6.4.3), the least complex among the non-rejected models, but the situation is already reversed with $L = 8$.

We present only one example of test results from the non-rejected category here; the rest are found in Appendix C. In Table 11 we display the results of our "absolute winner", p^g_{pair}.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	58.914	45.594	42.651	40.953	40.085	39.934	39.816	40.637
$P - val(\lambda)$	0.27	0.76	0.85	0.89	0.91	0.91	0.91	0.89

Table 11: The pair-potential model with all the geometric symmetries is not rejected. $\chi^2_{0.9,53} = 66.548$**.**

Remark 5.1 Note that the results of the test with $\lambda = -2$ is somewhat inconsistent with the others. According to [54], tests with large negative values of λ (e.g., $\lambda = -2$) help to detect "departures involving ratios of alternative to null expected frequencies that are close to 0" in a small number of cells. Since all our models involve some averaging (symmetrization), we see larger $T(\lambda)$ values (consequently, smaller P-values) with $\lambda = -2$: Averaging $\hat{p}$ over a rare symmetry class leads to overestimation of its more rare states. However, in the case of p^g_{pair} $T(-2)$ is still well below the common significance thresholds, from which we conclude that $\hat{p}(-2)$ captures the rest of the $\hat{p}$ values very well.

Remark 5.2 As outlined in §5.1, we use the MLE with $T(\lambda)$ for all the eight tests when testing p^g_{pair}. In Appendix C (Remark C.1), we will see the results of testing essentially the same model p^g_{symm} but using the true minimal power-divergence estimators with all the tests. With the exception of $\lambda = -2$, the results are very similar. We interpret this as an indication of the asymptotic equivalence of power-divergence estimators.

5.3.1 Hierarchical (Nested) Model Testing

Sometimes models are tested relative to each other as opposed to the common ground space, i.e., the most general alternative. In practice, such “nested” or “hierarchical” testing may actually be more valuable. For instance, in modeling a large system, due to prior knowledge or other factors, one can often restrict attention to a family of distributions which, despite being high-dimensional, is nonetheless much smaller than the set of all distributions. Given this initial restriction, any further simplification should consequently be tested against the distinguished family and not against all alternatives. In this context, the null hypothesis H_0 corresponds to a subspace Θ_0 of lower-dimension than the initial family Θ_a. Therefore, there now is a non-trivial alternative hypothesis $H_a: \ p \in \Theta_a$.

We use nested testing primarily to better understand which “dimensions” are the most important in modeling p. The following example is motivated by the observations from the previous section (§5.3). Models containing photometric inversion symmetry are consistently rejected whereas their supersets without inversion symmetry are not rejected. Thus, we want to determine if the rejection of these models is entirely due to the inversion assumption. Namely, we will test p^G_{symm} against p^n_{symm} in order to see if the geometric symmetries stand as firm in this restricted case as they did earlier (§5.3). Note that p^g_{symm} not being rejected against a general alternative does not imply that p^G_{symm} will not be rejected (on the same sample) against p^n_{symm}: The various subspaces may be oriented in complicated ways relative to one another. The results below (Table 12, top) show that, indeed, if one initially accepts p^n_{symm}, then further narrowing the model space to p^G_{symm} is accepted.

Table 12, bottom, shows a strong rejection of p^G_{symm} against p^g_{symm}. Thus, the decrease in accuracy due to imposition of the inversion constraints is not justified by

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	33.139	25.351	23.264	22.409	22.366	22.465	22.758	23.968
$P-val(\lambda)$	0.16	0.5	0.62	0.67	0.67	0.66	0.65	0.58
$T(\lambda)$	50.62	47.11	46.62	46.58	46.76	46.85	47.06	47.72

Table 12: Nested testing. Top: p^G_{symm} **is not rejected in favor of** p^n_{symm}**;** $\chi^2_{0.9,26} = 35.563$**. Bottom: All P-values**$\approx 10^{-7}$**;** p^G_{symm} **is rejected in favor of** p^g_{symm}**,** $\chi^2_{0.99,9} = 21.67$**.**

the complexity reduction: The reduced nine dimensions are apparently insufficient. At last, when tested against p^G_{symm}, the general G-symmetric alternative, the Potts models are still strongly rejected.

5.4 Summary

We have tested most of our models for the microimage distribution p. From the test results, we have learned that *p respects left-right and up-down reflection symmetries* and, in the case of 2 × 2 patches, *rotational symmetry.* The left-right reflection also appears to be the most prevalent among the geometric symmetries. In contrast, *photometric inversion symmetry is consistently rejected* by itself as well as in combination with other symmetries. Also, *the most general G-symmetric model is not rejected when tested against the inversion invariant alternatives.*

The Potts models and dominant-mass model are rejected even more dramatically than the models with inversion symmetry, although their entropies are reasonably close to four bits, the estimated entropy of p. *The Potts models are also rejected against the most general G-symmetric model, suggesting some information about absolute intensities is important.*

The geometrically invariant model determined only by pairwise interactions (p^g_{pair}) *is the most restrictive among the non-rejected models.* Even with the present limitation due to aggregation of rare states, our results suggest that *three- and four-pixel interactions are insignificant in the presence of all pairwise interactions.*

We also recognize the need for larger microimage samples in order to verify extensions of these results to finer quantizations ($L > 4$).

CHAPTER 6

COMPUTATIONS AND ALGEBRAIC REPRESENTATIONS

The objective of this chapter is twofold: First, we want to discuss systematically the computational issues alluded to in Chapters 4 and 5, in particular, the algorithms that compute the minimal power-divergence estimators for our models and some results that provide additional flexibility for the baseline computations. The other objective is to obtain algebraic representations for the symmetry-based models. Both objectives are relevant for extending model computations to larger intensity ranges, and eventually, to larger patches. The latter extension is also discussed in Chapter 7.

We first recall (§§4.2,5.1, Figure 6) the well-known equivalence results related to modeling by constraints.

I Constrained MLE (ECP/MLE) is equivalent to constrained minimization of Kullback-Leibler divergence $D(\hat{p}, q)$ (with the same constraints) from the empirical distribution: $\hat{p}_{MLE,\Theta_0} = \arg\min_{q \in \Theta_0} D(\hat{p}, q)$, where Θ_0 is the model set induced by the constraints.

II MEE under internal constraints $\mathbb{E}_q V = \mathbb{E}_{\hat{p}} V$ (MEE/ICP) is equivalent to constrained minimization of the Kullback-Leibler divergence $D(q, u)$ (with the same constraints) to the uniform distribution: $\hat{p}_{MEE,V} = \arg \max_{\mathbb{E}_q V = \mathbb{E}_{\hat{p}} V} D(q, u)$, where V are the constraining functions.

III MEE/ICP as above is also equivalent to maximum likelihood parameter estimation in the log- linear model (4.1) based on the same functions V: $\hat{p}_{MEE,V} = \hat{p}_{MLE,\Theta_0^{loglin}}$

We next consider potentially useful implications of these facts. To start, we consider a special case of the correspondence between I and III, namely a transformation of a typical symmetry-based ECP/MLE situation into an equivalent ICP/MEE (Proposition 6.1).

6.1 Translation of Symmetry ECP into ICP

A natural way to analyze real functions on Ω is via the K-dimensional real vector space $\mathbb{R}^K$, which is also an algebra, with the canonical basis $\{e_k\}_{k=1}^K$. Due to our assumption that p and all our model distributions are strictly positive, we focus on Q^+ and Θ^+, the set of all positive functions on Ω and the subset of all positive distributions on Ω, respectively. Symmetry constraints involve independent (with no loss of generality), homogeneous linear equations, and are a special case of external constraints. For a trivial example, consider $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ (thus $K = 4$) partitioned into two symmetry classes $\mathcal{O}_1 = \{\omega_1, \omega_3\}$ and $\mathcal{O}_2 = \{\omega_2, \omega_4\}$. The symmetry constraints can then be represented by the equations below:

$$\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}$$

In general, we define the feasible region Θ_0 when modeling by symmetry constraints as the restriction to Θ^+ of the kernel of the corresponding linear operator. The special feature of symmetry constraints is that they require functions (vectors) to be constant on certain subsets of Ω.

Let $F: \mathbb{R}^K \to \mathbb{R}^m$ be a symmetry-defining linear operator of rank m, written as an $m \times K$ matrix in its canonical form, i.e., each row of F contains exactly two non-zero entries, 1 and -1. Then $\ker(F) = \{h \in \mathbb{R}^K : \; Fh = 0 \in \mathbb{R}^m\}$ defines the linear subspace of $\mathbb{R}^K$ consisting of all real-valued functions on Ω respecting the given symmetry. Let $\mathcal{S}$ be the set of equivalence classes induced by F on Ω. Thus, $\forall \mathcal{O} \in \mathcal{S}$ we have: $\omega, \omega' \in \mathcal{O} \iff h(\omega) = h(\omega') \; \forall h \in \ker(F)$. Let $J = |\mathcal{S}| = K - m$. For clarity of exposition, the rows of F are ordered in accordance with the partition $\mathcal{S}$; in particular, the classes in $\mathcal{S}$ are enumerated from 1 to J and F has the following block-diagonal structure:

$$F = \begin{pmatrix} B_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & B_2 & \cdots & \mathbf{0} \\ \multicolumn{4}{c}{\dots\dots\dots\dots} \\ \mathbf{0} & \cdots & \mathbf{0} & B_J \end{pmatrix}, \tag{6.1}$$

where B_j has dimension $(|\mathcal{O}_j| - 1) \times |\mathcal{O}_j|$:

$$B_j = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \multicolumn{5}{c}{\dots\dots\dots\dots\dots} \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}, \tag{6.2}$$

Since $\sum\limits_{j=1}^{J} (|\mathcal{O}_j| - 1) = K - J = m$, F is indeed $m \times K$.

Let $\mathbb{I}_{\mathcal{O}}(\omega_k)$ stand for the indicator function of class $\mathcal{O}$. Clearly, the class indicators are linearly independent and form a basis for $\ker(F)$ since $\ker(F) = \text{span}\{\mathbb{I}_{\mathcal{O}_j} :$

$j = 1, \dots, J\}$. Note that the constant functions belong to $\ker(F)$. Let $\{V_j\}_{j=1}^{J-1}$ be any $J-1$ distinct class indicators: $V_{jk} = \mathbb{I}_{\mathcal{O}_j}(\omega_k)$ for $j = 1, \dots, J-1$. Recall (§4.2) that by an empirical distribution $\hat{p}$ we mean $\hat{p}_k = \frac{n_k}{N_\omega}$. Next we state the duality of the two categories of constraints problems:

Proposition 6.1 For any empirical distribution vector $\hat{p} \in \Theta^+$, the following ECP and ICP have identical, unique solutions:

$$\arg\min_{q \in \Theta_0} D(\hat{p}, q) \quad = \quad \arg\max_{q \in \Theta^+ : \mathbb{E}_q V = \mathbb{E}_{\hat{p}} V} H(q), \tag{6.3}$$

where $\Theta_0 = \ker(F) \cap \Theta^+$.

Remark 6.2 Recall (§4.3 and [47]) that positivity of $\hat{p}$ would ensure positivity of the ICP/MEE solution even if $\partial\Theta^+$ were part of the feasible region. Note then that the ICP conditions above can be replaced by $q \in \mathbb{R}^K,\ q_k \geq 0,\ k = 1, \dots, K$ and $\mathbb{E}_q V_j = \mathbb{E}_{\hat{p}} V_j,\ j = 0, \dots, J-1$ with $V_0 = \vec{1}$. (Writing $\mathbb{E}_q$ would still be legitimate in that case since q is implicitly normalized by $V_0 q = 1$.) For compactness, we will sometimes use the matrix multiplication notation $Vq = V\hat{p}$ for the equality of expectations: $\mathbb{E}_q V = \mathbb{E}_{\hat{p}} V$.

Proof. Step 1 We begin by "solving" the ECP. The constraints $Fq = \vec{0}$ imply that we have only $J \leq K$ parameters, $J-1$ of which are "free". A natural parametrization is provided by the probability values assumed on the symmetry classes. Let q'_j be the common value of q within the class $\mathcal{O}_j$ and let $Q_j = \sum\limits_{k \in \mathcal{O}_j} q_k = |\mathcal{O}_j| q'_j$ and $N_j = \sum\limits_{k \in \mathcal{O}_j} n_k$. Recall (§4.2.1) that minimizing $D(\hat{p}, q)$ is the same as maximizing the log-likelihood function $\sum\limits_k n_k \log q_k$. Due to the constraints, we have $\sum\limits_k n_k \log q_k = \sum\limits_j N_j \log \frac{Q_j}{|\mathcal{O}_j|}$, and maximizing this is equivalent to maximizing $\sum\limits_j N_j \log Q_j$. This is already in the form of unconstrained MLE, but relative to the state space $\mathcal{S}$. The solutions to this and to the original ECP are given, respectively, by:

$$Q_j = \frac{N_j}{N_\omega}, \quad q'_j = \tfrac{Q_j}{|\mathcal{O}_j|}, \quad j = 1, \ldots, J \tag{6.4}$$

Step 2 Rewrite the solution in (6.4) in exponential form:

$$q_k = \exp\left(\sum_{j=1}^{J} \gamma'_j \mathbb{I}_{\mathcal{O}_j}(\omega_k)\right), \qquad \gamma'_j = \log q'_j. \tag{6.5}$$

This is equivalent to

$$q_k = \exp\left(\gamma_0 - 1 + \sum_{j=1}^{J-1} \gamma_j V_{jk}\right) \tag{6.6}$$

with $\gamma_0 = 1 + \gamma_J$ and $\gamma_j = \gamma'_j - \gamma'_J$ for $j = 1, \ldots, J-1$. By uniqueness of the Maximum Entropy Extension (§4.3), the γ's of (6.6) are identified with the corresponding *Lagrange Multipliers* of the ICP. Thus, the vector q in (6.5), (6.6) is the solution of the ICP. This completes the proof. $\diamond$

Remark 6.3 Implicit in the above proof is that $\vec{1} \in \ker(F)$. Thus, adjoining $\vec{1}$ to any $J-1$-dimensional subspace of $\ker(F)$ not containing $\vec{1}$ recovers the entire subspace $\ker(F)$. This helps to see why the separation of the normalization constraint from other constraints is rather artificial at this point: Suppose $\{V_j\}_{j=1}^{J}$ is an arbitrary basis of $\ker(F)$ and the conditions of the above ICP are replaced by $q \in \mathbb{R}^K$, $q_k \geq 0$, $k = 1, \ldots, K$ and $Vq = V\hat{p}$. Then the equivalence still holds, although the normalization constraint may now be implicit. In order to find new Lagrange multipliers, one would have to solve a system of linear equations corresponding to the change of bases, namely the new $\{V_j\}_{j=0}^{J-1}$ and $\{\mathbb{I}_{\mathcal{O}_j}\}_{j=0}^{J-1}$. Clearly, $\{\vec{1}, \mathbb{I}_{\mathcal{O}_1}, \ldots, \mathbb{I}_{\mathcal{O}_{J-1}}\}$, our choice of V in the proposition, trivialized this technicality.

Remark 6.4 The equivalence in Proposition 6.1 is just a special case of that between log-linear modeling with MLE (namely ECP/MLE) and ICP/MEE because symmetry constraints generalize to constraining the vector of log-probabilities to belong to some linear subspace of functions on Ω. (To avoid the degeneracy of having only one feasible probability vector, this subspace must contain constant functions.)

Computationally, the above equivalence implies that in order to find the MEE in the ICP of (6.3), one can simply obtain maximum likelihood estimates on the quotient space (of equivalence classes) and then uniformly divide their aggregate masses among the elements of each class. Consequently, instead of solving a system of nonlinear equations to find Lagrange multipliers, one *averages* the empirical distribution over the symmetry classes. The averaging operator is clearly *linear* and is exactly the operator $\mathcal{R}$ mentioned in §4.2.1. However, despite its seeming simplicity it plays an important role in the general theory of invariants [12],[52],[64],[65], one proper framework for studying symmetries. In that context it bears the name *Reynolds Operator*. We write $\mathcal{R}$ for this operator and, when necessary, use a subscript to indicate the origin of the equivalence classes. Thus, the solution to the ECP/MLE with symmetry classes $\mathcal{S}$ and empirical distribution $\hat{p}$ is $\mathcal{R}_{\mathcal{S}}\hat{p}$, or, equivalently $\mathcal{R}_V\hat{p}$, where V refers to a basis of $\mathbb{R}^{\mathcal{S}}$. We also index $\mathcal{R}$ by a group of symmetries if the group is part of the discussion.

An obvious question is to characterize the MEE/ICPs that originate in this way from the ECP/MLE. Does any ICP/MEE admit an equivalent symmetry-based ECP formulation? The answer in general is clearly "no". Otherwise, in particular, there would be no distinction between p_{potts}, Potts models, and p^C_{potts}, their "completed" versions (§4.6). However, we still hope to demonstrate the practical value of attempting such a conversion. One motivation is to enrich the stock of computational tools for modeling distributions on larger microimage spaces, for instance 3×3 patches with $L = 8$, rendering $|\Omega| = 2^{27}$. Choosing between the MEE approach and the construction of the averaging operator would be one example; of course, the choice must depend on available resources. Two concrete examples are estimating parameters in the "completed" Potts model, p^{vhnC}_{potts}, and in the geometrically invariant pair-potential model, p^{g}_{pair}.

6.2 From ICP to Symmetry-Based ECP

Before discussing a general result, let us look at a simple example:

Example 1 Consider a binary case $\Omega = M_{2\times 2}(\{0,1\}) = \{\begin{smallmatrix}0&0\\0&0\end{smallmatrix}, \begin{smallmatrix}0&0\\1&0\end{smallmatrix}, \cdots, \begin{smallmatrix}1&1\\0&1\end{smallmatrix}, \begin{smallmatrix}1&1\\1&1\end{smallmatrix}\}$, where our default enumeration of the 16 elements of Ω (3.1) becomes: $k(\omega) = 1 + \omega_{2,1} + 2\omega_{2,2} + 4\omega_{1,1} + 8\omega_{1,2}$. We then impose constraints based on *Potts*-like potentials (§4.6), i.e. $V_1 = \mathbb{I}_{\{\omega_{1,1}=\omega_{2,2}\}} + \mathbb{I}_{\{\omega_{1,2}=\omega_{2,1}\}}$, and $V_2 = \mathbb{I}_{\{\omega_{1,1}=\omega_{1,2}\}} + \mathbb{I}_{\{\omega_{1,2}=\omega_{2,2}\}} + \mathbb{I}_{\{\omega_{2,2}=\omega_{2,1}\}} + \mathbb{I}_{\{\omega_{2,1}=\omega_{1,1}\}}$. V_0 is again the normalization constraint vector $\vec{1}$. Thus, $V = \left(\begin{smallmatrix}V_0\\V_1\\V_2\end{smallmatrix}\right) = \left(\begin{smallmatrix}1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\2&1&1&0&1&0&2&1&1&2&0&1&0&1&1&2\\4&2&2&2&2&2&0&2&2&0&2&2&2&2&2&4\end{smallmatrix}\right)$. Given a positive empirical distribution $\hat{p}$, the distribution p^{G}_{potts} is defined as $\hat{p}_{MEE,V}$, the maximum entropy extension of $\hat{p}$ constrained by $\mathbb{E}_{.}V_j = \mathbb{E}_{\hat{p}}V_j$ for $j = 1, 2$. Recall that the transition from the ECP to ICP in §6.1 was made possible because the vectors $\{V_j\}_{j=0}^{J-1}$ of the internal constraints spanned the entire subspace of functions constant on the $\mathcal{O}$'s. Notice that our present V induces the following four constancy classes: $\mathcal{O}_1 = \{1, 16\}$, $\mathcal{O}_2 = \{2, 3, 5, 8, 9, 12, 14, 15\}$, $\mathcal{O}_3 = \{4, 6, 11, 13\}$, $\mathcal{O}_4 = \{7, 10\}$. For example, $\begin{smallmatrix}0&0\\0&0\end{smallmatrix} \in \mathcal{O}_1$, $\begin{smallmatrix}0&0\\1&0\end{smallmatrix} \in \mathcal{O}_2$, $\begin{smallmatrix}0&0\\1&1\end{smallmatrix} \in \mathcal{O}_3$, and $\begin{smallmatrix}1&0\\0&1\end{smallmatrix} \in \mathcal{O}_4$. However, $\mathrm{span}\{V_0, V_1, V_2\} \subsetneqq \mathrm{span}\{\mathbb{I}_{\mathcal{O}_1}, \mathbb{I}_{\mathcal{O}_2}, \mathbb{I}_{\mathcal{O}_3}, \mathbb{I}_{\mathcal{O}_4}\}$. Thus, the feasible region of this

ICP is strictly larger (by one dimension) than that of the problem with internal constraints generated by $\{\mathbb{I}_{\mathcal{O}_j}\}_{j=1}^4$. The latter surely admits the ECP/MEE formulation reversing the lines of §6.1, and the corresponding solution is $\mathcal{R}_V\hat{p}$, the version of $\hat{p}$ averaged over the four constancy classes (see Table 13). Note that this solution corresponds exactly to p_{potts}^{GC}, the "completed" version of p_{potts}^{G} (§4.6). The distribution $\hat{p}$ in the table was chosen rather arbitrarily; it is not an estimate of the microimage distribution. (The last column of this table will be explained in §6.3.) $\hat{p}_{MEE,V}$, the solution to the original MEE/ICP indeed "smoothes" $\hat{p}$ more than does the simple symmetrization, namely $H(\hat{p}_{MEE,V}) > H(\mathcal{R}_V\hat{p})$.

$k(\omega)$	$\hat{p}(\omega)$	$(\mathcal{R}_V\hat{p})(\omega)$	$\hat{p}_{MEE,V}$	q^*
1	0.1641	0.166	0.1588	0.1845
2	0.043	0.0483	0.0556	0.0299
3	0.0508	0.0483	0.0556	0.0299
4	0.0469	0.0508	0.0435	0.0693
5	0.0391	0.0483	0.0556	0.0299
6	0.043	0.0508	0.0435	0.0693
7	0.0312	0.0391	0.0318	0.0575
8	0.0547	0.0483	0.0556	0.0299
9	0.0625	0.0483	0.0556	0.0299
10	0.0469	0.0391	0.0318	0.0575
11	0.0664	0.0508	0.0435	0.0693
12	0.0430	0.0483	0.0556	0.0299
13	0.0469	0.0508	0.0435	0.0693
14	0.0391	0.0483	0.0556	0.0299
15	0.0547	0.0483	0.0556	0.0299
16	0.168	0.166	0.1588	0.1845
H	3.7748	3.7892	3.8009	3.6511

Table 13: Comparison of solutions to related ICP/MEE's.

This example clearly illustrates why it is not possible in general to translate an ICP/MEE problem into an equivalent symmetry ECP/MLE: *The subspace spanned by the constraint vectors of the ICP need not in general be large enough to include all* *possible functions respecting the constancy classes of the ICP constraint operator.* Consequently, one generally achieves extra smoothing in the original ICP/MEE due to additional freedom for mass redistribution (e.g., an extra dimension in the example above) as compared with the effect of the averaging operator which only redistributes masses within the constancy classes.

Thus, it now makes sense to call "complete" (§4.6) an ICP/MEE whose constraint vectors V span the whole space of functions constant on the constancy classes of V. In particular, the solution to this problem is given by $\hat{p}_{MEE,V} \overset{(4.2)(6.3)}{=} \hat{p}_{MLE,\Theta_0} \overset{(6.4)}{=} \mathcal{R}_V \hat{p}$, where Θ_0 consists of all positive distributions constant on the constancy classes of V. The computation of V-constancy classes in the case of p^{vhnC}_{potts} (§6.4.1) provides another, concrete example for this discussion. The difference between the original and "completed" models in that case is more noticeable both in terms of the complexity comparison and the fit to the empirical $\hat{p}$: The entropy of $\hat{p}^{vhnC}_{potts}$ is closer to that of $\hat{p}$ than to the entropy of $\hat{p}^{vhn}_{potts}$.

Of course, such situations are only interesting when the size of the state space is much larger than the number of constancy classes of the constraint operator V.

We conclude this section by noticing:

Proposition 6.5 The maximum entropy distribution $\hat{p}_{MEE,V}$ constrained by V always belongs to the "symmetry" subspace Θ_0, i.e. it always respects the constancy classes of its constraints in that $\mathcal{R}_V \hat{p}_{MEE,V} = \hat{p}_{MEE,V}$.

Proof. This can be seen directly from the exponential form of $\hat{p}_{MEE,V}$ (4.6). This is also a consequence of an important information inequality: First, notice that for any distribution q from the feasibility region $F(\hat{p}) = \{q \in \Theta^+ : \ qV = \hat{p}V\}$, its symmetrized version $\mathcal{R}_V q$ (i.e., averaged over the constancy classes) also belongs to $F(\hat{p})$ since $Vq = V\mathcal{R}_V q$[1]. But then the smoothed vector has at least the amount of uncertainty (entropy) of the initial vector. Thus, the solution must also be $\mathcal{R}_V$-invariant. The just quoted inequality $H(q) \leq H(\mathcal{R}_V q)$ follows from the so-called *log sum inequality* [11]:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum\limits_{i=1}^{n} a_i}{\sum\limits_{i=1}^{n} b_i}, \quad a_i,\ b_i > 0 \quad i = 1, \ldots, n,$$

which is in turn an immediate application of Jensen's inequality to the convex function $x \log x$. The above inequality needs to be applied to every $\mathcal{O}$, constancy class of V, producing:

$$\sum_{\omega_i \in \mathcal{O}} q(\omega_i) \log \frac{1}{q(\omega_i)} \leq \left(\sum_{\omega_i \in \mathcal{O}} q(\omega_i) \right) \log \frac{|\mathcal{O}|}{\sum\limits_{\omega_i \in \mathcal{O}} q(\omega_i)}.$$

Summing over all such orbits yields $H(q) \leq H(\mathcal{R}_V q)$. ◇

6.3 Factorization effect in ICP/MEE

Next, we discuss reducing an ICP/MEE problem to a "similar" one but formulated relative to the quotient space of the constancy classes of the constraining operator. The motivation is to reduce the dimension of the space in which the ICP is to be solved. For example, consider modeling p by p^g_{pair} (§4.7). The constraint functions $\{V_j\}_{j=1}^{J}$ then form a basis for $\mathcal{D}$, the linear subspace spanned by all geometrically invariant functions on Ω of (at most) two variables. With $L = 4$ there are 55 g-symmetric orbits, significantly fewer than the 256 states in Ω. In order to obtain the Lagrange multipliers in the MEE, one needs to solve numerically the exponential constraint equations. In this case, solving these equations directly

[1] Representing $\mathcal{R}_V$ as a matrix similar to (6.7) immediately verifies $V = V\mathcal{R}_V$.

necessitates manipulating 256-dimensional vectors[2]. On the other hand, the probability distributions on the quotient space $\mathcal{S}$ reside entirely inside a 55-dimensional space. It is naturally tempting to "tunnel through" the coarser space in order to end up at least "close" to (if not right at) the true $\hat{p}_{MEE,V}$. How would one do it? *In the following discussion, suppose $\mathcal{S}$ is* any *partition of Ω such that V is constant on each $\mathcal{O} \in \mathcal{S}$ (i.e., not necessarily the partition of the constancy classes of V).*

One intuitive approach to the above question is to replace V and the original (empirical) distribution $\hat{p}$ by their projections $\tilde{V}$ and $\tilde{p}$, where $\tilde{p}$ represents $\hat{p}$-probabilities integrated over $\mathcal{S}$-classes and $\tilde{V}$ is obtained from V as follows: For each $\mathcal{O} \in \mathcal{S}$, we retain a single column of V representing the value of V on $\mathcal{O}$. Let us denote a solution to an ICP relative to the quotient space by $\tilde{q}^*$. To return to the original space, we split $\tilde{q}^*$ uniformly within the classes and thus arrive at some (symmetric) distribution q^*. Provided the averaging operator $\mathcal{R}_{\mathcal{S}}$ is easy to construct, all the necessary operations are matrix multiplications, and in general are computationally cheap. Thus we seek an ICP in the new space that would result (following the above program) in the same solution as the original ICP/MEE. In particular, if $\tilde{q}^*$ solves the ICP/MEE determined by $\tilde{V}$ and $\tilde{p}$ in the quotient space, how close would q^* be to the solution of the original ICP/MEE? Some answers are provided by Theorem 6.6 but first we need to introduce some notation.

Let V be an $M \times K$ matrix whose rows play the role of constraint functions and are assumed to be linearly independent. V is also assumed to contain a constant non-zero row corresponding to the normalization constraint. Suppose $|\mathcal{S}| = J$. In order to clarify the set-up, let J' be the number of the constancy classes of V (i.e., the size of the coarsest partition respected by V). Note that $M \leq J' \leq J \leq K$ since

[2] With $L = 4$, computational problems are not actually encountered, but the situation changes markedly with $L = 8$, in which case there are 4096 states. We will explain a more important computational concern in §6.4.3 in the context of constructing $\mathcal{D}$.

$\operatorname{rank}\tilde{V} = \operatorname{rank} V = M$, which also equals the rank of the matrix of unique columns of V. Let π_1 be the $J \times K$ binary matrix whose rows are the $\mathcal{S}$-class indicators $\mathbb{I}_{\mathcal{O}_j}$ for $j = 1, \ldots J$. Therefore $\tilde{p} = \pi_1 \hat{p}$. Let π_2 be the $K \times J$ matrix whose j-th column is the j-th class indicator reduced by the class size: $\frac{\mathbb{I}_{\mathcal{O}_j}}{|\mathcal{O}_j|}$. Thus, $\tilde{V} = V\pi_2$. The averaging operator $\mathcal{R}_{\mathcal{S}}$ can then be represented as the $K \times K$ matrix:

$$\mathcal{R}_{\mathcal{S}} = \pi_2 \pi_1, \tag{6.7}$$

corresponding to the composition of the "integration" (π_1) and "differentiation" (π_2) operators. (We do not distinguish here between linear operators and their matrix representations.) Also, let $s_j = \frac{|\mathcal{O}_j|}{K}$, the relative size of the j-th class, and write $\tilde{s}$ for the corresponding J-dimensional probability vector. We will generally use the superscript $\sim$ to denote objects relevant to the quotient space $\mathcal{S}$.

Theorem 6.6

$$\arg\max_{q \in Q^+ : Vq = V\hat{p}} H(q) \quad = \quad \pi_2 \left(\arg\min_{\tilde{q} \in \tilde{Q}^+ : \tilde{V}\tilde{q} = \tilde{V}\tilde{p}} D(\tilde{q}, \tilde{s}) \right) \tag{6.8}$$

Proof. Recall (Proposition 6.5) that in maximizing entropy, q necessarily satisfies $\mathcal{R}_{\mathcal{S}} q = q$; hence

$$\begin{aligned}
\arg\max_{q \in Q^+ : Vq = V\hat{p}} H(q) &= \arg\max_{q \in Q^+ : \substack{\mathcal{R}_{\mathcal{S}} q = q \\ V\mathcal{R}_{\mathcal{S}} q = V\hat{p}}} H(\mathcal{R}_{\mathcal{S}} q) \\
&= \arg\max_{q \in Q^+ : \substack{\mathcal{R}_{\mathcal{S}} q = q \\ V\pi_2\pi_1 q = V\pi_2\pi_1\hat{p}}} H(\pi_2 \pi_1 q) \\
&= \pi_2 \left(\arg\max_{\tilde{q} \in \tilde{Q}^+ : \tilde{V}\tilde{q} = \tilde{V}\tilde{p}} H(\pi_2 \tilde{q}) \right) \\
&= \pi_2 \left(\arg\max_{\substack{\tilde{q} \in \tilde{Q}^+ \\ \tilde{V}\tilde{q} = \tilde{V}\tilde{p}}} \sum_{j=1}^{J} \sum_{\omega_k \in \mathcal{O}_j} \frac{\tilde{q}(\mathcal{O}_j)}{|\mathcal{O}_j|} \log \frac{|\mathcal{O}_j|}{\tilde{q}(\mathcal{O}_j)} \right) \\
&= \pi_2 \left(\arg\min_{\substack{\tilde{q} \in \tilde{Q}^+ \\ \tilde{V}\tilde{q} = \tilde{V}\tilde{p}}} \sum_{j=1}^{J} \tilde{q}(\mathcal{O}_j) \log \frac{\tilde{q}(\mathcal{O}_j) K}{|\mathcal{O}_j| K} \right)
\end{aligned}$$

$$= \pi_2 \left[\arg \min_{\substack{\tilde{q} \in \tilde{Q}^+ \\ \tilde{V}\tilde{q} = \tilde{V}\tilde{p}}} \left(D(\tilde{q}, \tilde{s}) - \log K \right) \right]$$

$$= \pi_2 \left(\arg \min_{\tilde{q} \in \tilde{Q}^+ : \tilde{V}\tilde{q} = \tilde{V}\tilde{p}} D(\tilde{q}, \tilde{s}) \right).$$

Implicit in the second equality is that $V = V\mathcal{R}_{\mathcal{S}} = V\pi_2\pi_1$, which restates the fact that averaging V over subsets of its constancy classes has no effect. The third equality follows from the parametrization of $\Theta_0 = \{q \in \Theta \subset \mathbb{R}^K : \mathcal{R}_{\mathcal{S}} q = q\}$, the K-dimensional $\mathcal{S}$-symmetric distributions by the J-dimensional distributions Θ^J. Clearly, $\pi_1|_{\Theta_0}$, the restriction of the linear map π_1 to Θ_0 yields an isomorphism[3]: $\Theta_0 \cong \Theta^J$, where $\pi^{-1}_{1|_{\Theta_0}} = \pi_{2|_{\Theta^J}}$, the restriction of the linear map π_2 to Θ^J. Finally, in the fourth equality the summation over $\{\omega_1, \ldots, \omega_K\}$ is broken down into the external summation over $\mathcal{S}$ and the internal summation over $\mathcal{O}$'s, the individual constancy classes. $\diamond$

Remark 6.7 Note that the above problem is in general different from the MEE on the quotient space:

$$\arg \max_{\tilde{q} \in Q^+ : \tilde{V}\tilde{q} = \tilde{V}\tilde{p}} H(\tilde{q}) = \arg \min_{\tilde{q} \in Q^+ : \tilde{V}\tilde{q} = \tilde{V}\tilde{p}} D(\tilde{q}, \tilde{u}) \tag{6.9}$$

Thus, $\pi_2 \left(\arg \min_{\tilde{q} \in Q^+ : \tilde{V}\tilde{q} = \tilde{V}\tilde{p}} D(\tilde{q}, \tilde{u}) \right)$ need not in general equal the solution in the theorem. A sufficient condition for the two solutions to be the same is to have equal size classes, i.e. $\tilde{s} = \tilde{u}$.

Another sufficient condition is that $J = M$. In that case, $\tilde{V}$ is a non-singular square matrix and, consequently, the only solution to $\tilde{V}\tilde{q} = \tilde{V}\tilde{p}$ is $\tilde{p}$. Hence the optimization degenerates and the solution to the original problem is given by $\mathcal{R}_{\mathcal{S}}\hat{p}$.

[3] This is exactly the parametrization that allows us to compute the minimum-power divergence estimators $\hat{p}(\lambda)$ for our symmetry-based models and verify the (Birch's) regularity conditions required for the corresponding power-divergence testing (Appendix B).

Note also that $\mathcal{S}$ must then necessarily be the set of V constancy classes, implying $\mathcal{R}_{\mathcal{S}} = \mathcal{R}_V$. This is exactly the special case in which an ICP/MEE transforms into an equivalent ECP/MLE and no optimization need be attempted. Specifically, $\hat{p}_{MEE,V} = \hat{p}_{MLE,\Theta_0}$ (4.2) independently of $\tilde{s}$, where Θ_0 is the feasible region of the equivalent ECP with V-induced symmetries. Clearly, Proposition 6.1 now becomes a corollary to Theorem 6.6.

This suggests two options for the reduction of dimensionality in ICP/MEE. The first (6.8) provides the exact solution to the original problem by minimizing the Kullback-Leibler distance to the distribution of class sizes. Note that this distribution becomes automatically available once the partition $\mathcal{S}$ has been identified. The computations required are essentially identical to those of MEE: Solving (numerically) a system of exponential equations to find the Lagrange multipliers. The only difference is reweighting of the summands of the equations according to the class sizes. The second option (6.9) is to disregard any nonuniformity of the class sizes, and will not in general yield the true solution. The data of Example 1 (see Table 13) show that this may lead to a poor approximation: The solution to the "factor"-ICP/MEE "pulled-back" from $\mathcal{S}$ to Ω is q^* which has entropy even smaller than entropy of the empirical distribution. This is clearly a result of the highly non-uniform distribution of the class sizes.

In the case of our symmetric models, the proportion of singular orbits (of size less than the order of the appropriate symmetry group) becomes negligible as L grows and $\tilde{s}$ is increasingly flattened. As a concrete example, consider $L = 8$ in which case $|\Omega| = 4096$ but the number of G-symmetry classes is only 346 (Proposition 6.8). Imagine estimating parameters of p^G_{pair}. Omitting some technical details concerning identifying V in this case, it may already seem discouraging that one needs to solve a system of about 50 non-linear equations[4]. Still, manipulating vectors of dimension 346 instead of 4096 makes a difference. In addition, $\tilde{s}$ is almost uniform: $H(\tilde{s}) = 8.3203 \approx 8.4346 = H(\tilde{u})$, suggesting little difference between the exact (6.8) and the approximate (6.9) solutions in this case.

6.4 Quotient Spaces of Invariant Classes

We next revisit some of our models defined in Chapter 4. Since we have explicitly introduced symmetries in most of the models (p_{symm}, p^{C}_{potts}, and p^{g}_{pair}), computation of the symmetry classes is a central issue in estimating model parameters. As a first step, in §6.4.1 we define the symmetry transformations (§4.4) using the group-theoretic formalism (Appendix D). We then discuss the structure of the quotient space $\mathcal{S}$ in the case of the G-invariant model p^{G}_{symm} and complete Potts model p^{vhnC}_{potts} (§6.4.2). In §6.4.3 we also explain how we compute $\pi :\ \Omega \to \mathcal{S}$, a map indexing symmetry classes, and lead to the problem of finding analytical representations for the probability functions of the symmetric models (§6.5).

6.4.1 G-Symmetries

For simplicity, let us relabel the matrix entries $\{\omega_{2,1},\, \omega_{2,2},\, \omega_{1,2},\, \omega_{1,1}\}$ as $\{\omega_1,\, \omega_2,\, \omega_3,\, \omega_4\}$. The context should always disambiguate between this enumeration and the indexing of Ω (3.1), where we also use subscripts. Recall from §4.4 that the geometric symmetries consist of rotation and reflections. Let us start with the rotation group. This group is generated by the *four-cycle* $(1, 2, 3, 4)$, representing the counterclockwise $\pi/2$ rotation r. Hence, the entire rotation group G_r is $\{1, r, r^2, r^3\}$.

[4] $\operatorname{rank} V = 48$ was only computed numerically. The same numerical computations showed that the G-symmetric partition is indeed the coarsest one respected by V.

Adding the *reflections* through all four axes of the square to this group, we arrive at the *dihedral group* D_8[5], whose symmetries cover all the geometric symmetries we have used in our modeling (e.g., p^g_{symm} and p^g_{pair}). This group has the following *presentation* in terms of its two *generators* and the defining relations between them: $\langle r, s | r^4 = s^2 = 1, rs = sr^3 \rangle$. Let us agree to represent by s the reflection through the diagonal $\omega_1 - \omega_3$. We use the convention that in a composite symmetry transformation the action develops from right to left; for example, rs means that the diagonal reflection precedes the rotation. The subgroups of D_8 that correspond to the left-right and up-down reflections are $\langle rs \rangle$ $\langle sr \rangle$ respectively. Combined, the two groups additionally generate the symmetry with respect to rotation by π, namely the resulting group $\langle rs, sr | (sr)^2 = (rs)^2 = 1 \rangle$ contains $r^2 = rssr$. These are maximal geometric symmetries of the models p^{vh}.

The last symmetry required to generate G is symmetry with respect to the *photometric inversion*. To simplify the ensuing discussion involving this symmetry, we translate the range $\{0, \ldots, L-1\}$ to center it at 0. Then, the transformation corresponding to this symmetry is a true negation: $i(\omega)_l = -\omega_l$, for $l = 1, \ldots, 4$. Finally, the group G generated by all the symmetries above has presentation $\langle r, s, i | r^4 = s^2 = i^2 = 1, si = is, ri = ir, rs = sr^3 \rangle$. Therefore, $G \cong D_8 \times C_2$, where C_2 is the *cyclic* group of order two which is evidently isomorphic to $\langle i \rangle$.

The following proposition presents the exact number and sizes of $\mathcal{S}$, the set of G-equivalence classes (or, simply, G-orbits). A proof of these statements is given in Appendix D. The case of L odd is practically not interesting and hence not covered here (its computations would also require several minor modifications). Similar results for the other symmetry (sub)groups could also be computed analyt-

[5] We follow the notation of [18] in which D_{2n} stands for the group of all symmetries of a regular n-gon. Another popular notation for this group is D_n.

ically. Instead, we induced the necessary information from the action of the averaging operator $\mathcal{R}_{\mathcal{S}}$, first encountered in §4.2.1 and further discussed in §§6.3,6.4.3. Alternative ways of "orbit counting" are also discussed in §6.5.

Proposition 6.8 The size of the partition $\mathcal{S}$ of $\Omega = M_{2\times 2}(\mathcal{C}_L)$ $(L = 2n)$ into the G-invariant classes (orbits) is $|\mathcal{S}_L| = \frac{L^4+2L^3+6L^2+4L}{16} = n^4 + n^3 + \frac{n(1+3n)}{2}$. Among them, there are L orbits of size two, $\frac{L^2}{4}$ orbits of size four, $\frac{2L^3+3L^2-10L}{8}$ orbits of size eight, and $\frac{L^4-2L^3-4L^2+8L}{16}$ orbits of size 16.

This proposition and its proof (Appendix D) suggest the following asymptotic result for any subgroup $\mathcal{H} \leq G$: The leading term of $|\mathcal{S}_L|$ is $\frac{|\Omega|}{|\mathcal{H}|}$, i.e., $\frac{|\mathcal{S}_L||\mathcal{H}|}{|\Omega|} \to 1$ as $L \to \infty$. In particular, the complexity of the corresponding symmetric models p_{symm} grows as L^4 $(= |\Omega|)$.

6.4.2 Completion of the Potts Model p^{vhn}_{potts}

Table 14 presents sizes of the nine constancy classes of the Potts operator $V = (V_1, V_2, V_3)^t$:

$$\begin{aligned} V_1(\omega) &= \mathbb{I}_{\{\omega_4=\omega_2\}} + \mathbb{I}_{\{\omega_3=\omega_1\}} \\ V_2(\omega) &= \mathbb{I}_{\{\omega_4=\omega_3\}} + \mathbb{I}_{\{\omega_1=\omega_2\}} \\ V_3(\omega) &= \mathbb{I}_{\{\omega_4=\omega_1\}} + \mathbb{I}_{\{\omega_3=\omega_2\}}. \end{aligned}$$

In Table 14, $P^l_L = \frac{L!}{l!}$ stands for the number of permutations of L elements taken l at a time. Recall (§4.6) that V determines p^{vhn}_{potts}, and then the completed Potts

V_1, V_2, V_3	0,0,0	0,0,1	0,0,2	0,1,0	0,2,0	1,0,0	1,1,1	2,0,0	2,2,2
$\lvert\mathcal{O}\rvert$	P^4_L	$2P^3_L$	P^2_L	$2P^3_L$	P^2_L	$2P^3_L$	$4P^2_L$	P^2_L	P^1_L

Table 14: Constancy class sizes of the Potts constraint operator V.

model, p^{vhnC}_{potts}, corresponds to expanding V to include the five "missing" dimensions. Namely, we add to V five more constraint vectors so that the expanded set becomes a basis for the linear subspace of all functions on Ω that are constant on the constancy classes of V. Thus, p^{vhnC}_{potts} is originally the MEE under the constraints that the probabilities of the nine constancy classes equal their $\hat{p}$ estimates. In practice, it may be more convenient if the role of constraining functions is assumed by the nine class indicators. We can also characterize p^{vhnC}_{potts} as a symmetry ECP (§§6.1,6.2). This is, in fact, how we treat this model for statistical testing based on the minimum power-divergence tests. One way or the other, the nine-class partition is characterized by basic combinatorics used to derive the results in the table and shows that the complexity of p^{vhnC}_{potts} (which is eight due to the normalization condition) is independent of L.

6.4.3 On Computation of Symmetry Partitions

In order to work with models invariant under symmetry transformations such as those represented by the group G and possibly more general ones, it is important to understand the dependence of parameter estimation on the particular representation for $\mathcal{S}$, the space of invariant (symmetric) classes[6]. For example, the maximum likelihood estimator for symmetry-based models (e.g., p^{g}_{symm}, p^{vhnC}_{potts}) is just the orbit-averaging operator $\mathcal{R}$ (6.7) applied to the empirical distribution $\hat{p}$. Although expressions for other minimum power-divergence estimators (see (B.4) and (B.5)) involve non-linear averaging, computation of the symmetry partition $\mathcal{S}$ is essential in all cases.

We discuss two options for computing this partition in the case of the geometric and inversion symmetries. It should later become clear that this discussion is

[6] We now abandon the default $L = 4$ and, when necessary, will put L in the subscript.

also relevant to certain other symmetry transformations. The first option is rather straightforward and involves simulating the action of the appropriate symmetry group on Ω. For small Ω's, for example $|\Omega| = 4096$ ($L = 8$), the inefficiencies (e.g., redundant computations) of this method are insignificant relative to available computational resources. Moreover, purely algorithmic modifications could improve efficiency. With $|\Omega|$ as small as 256, the baseline computations performed in Matlab [49] are virtually instantaneous on a *Sun Ultra-2* workstation and hence no improvements are necessary. However, for larger microimage spaces, due to spatial supports and/or finer quantization, a second option involving analytic representations may be more sensible. The second half of this section discusses potential benefits associated with this alternative and in §6.5 a concrete example demonstrates the main ideas of this second approach in the context of the p^G_{symm} model.

In the first approach, the baseline algorithm starts with all patches unlabeled and then runs over Ω, assigning labels to the elements of the partition $\mathcal{S}$. Patches whose class has already been computed are skipped. If an unlabeled patch ω is encountered, the class counter is incremented, and the algorithm loops over the entire symmetry group (except for the identity element). Successively applying to ω the symmetry transformations from the group produces indices of the other members of ω's class. The corresponding patches are then labeled appropriately. Thus, the algorithm constructs the desired map $\pi : \ \Omega \rightarrow \mathcal{S}$. Given this map, the operators π_1, π_2 and hence the averaging operator $\mathcal{R}$ are determined and the MLE is then constructed. Computer languages that support vectorization (e.g., Matlab [49]) are particularly convenient for using π to compute all minimum power-divergence estimators. One source of inefficiency in this algorithm is that it ignores the notion of singularities (classes smaller than the order of the group) and unnecessarily relabels singular orbits.

Note also that in the case of the group G, and any of its subgroups, the symmetry transformations (§6.4.1) can be expressed as *matrix multiplications.* To see this, identify 2×2 patches with points in $\mathbb{R}^4$. (Recall that we have ordered the patch components as $\{\omega_{2,1},\ \omega_{2,2},\ \omega_{1,2},\ \omega_{1,1}\}$ and translated the intensity range to $\{-\frac{L-1}{2}, \ldots, -\frac{1}{2}, \frac{1}{2}, \ldots, \frac{L-1}{2}\}$, assuming L is even.) Clearly, our symmetry transformations r, s, and i extend to transformations on the whole vector space $\mathbb{R}^4$. (This is possible mainly because Ω is *invariant* under the extended transformations r, s, and i.) In proper terms, the given G-action admits a linear representation on $\mathbb{R}^4$ ($\mathbb{R}^{n^2}$ in the case of $n \times n$ patches): $\rho : G \hookrightarrow GL(\mathbb{R}^4)$, where $GL(\mathbb{R}^4)$ is the group (under composition) of the nonsingular linear transformations from $\mathbb{R}^4$ to itself. Then, with the standard basis, ρ has the following matrix representation (via $GL(\mathbb{R}^4) \cong GL(4, \mathbb{R})$):

$$r \overset{\rho}{\mapsto} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad s \overset{\rho}{\mapsto} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad i \overset{\rho}{\mapsto} \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{6.10}$$

This correspondence *allows us not to distinguish between G and its matrix representation in the ensuing discussion.* This also shows that the computations needed to fill in a class in the above algorithm are simply matrix multiplications.

The second approach to calculating $\mathcal{S}$ is somewhat more analytic. Roughly, the idea is to represent $\mathcal{S}$ (i.e., construct π) by obtaining algebraic expressions for the class indicators $\mathbb{I}_{\mathcal{O}}$, $\mathcal{O} \in \mathcal{S}$. Particular computer implementations could, of course, significantly deviate from the main idea in order to improve efficiency. However, we do not concern ourselves with such details here. Instead, we only provide the main ingredients from which π can be assembled in various ways. We will also relate this discussion to the more general theme of finding analytic, *Gibbs* representations for symmetry-based models. G will provide a concrete example for illustrating these ideas. In fact, the emphasis of §6.5 will quickly shift from the alternative construction of π (required to compute model estimators)[7] to obtaining analytic forms for the model distributions.

Another benefit of replacing the orbit indicators in equations such as (6.5) by explicit functions of the patch components is that the latter representation is more suitable to identify *the degree of interaction among pixels in the patch.* For example, consider computing p^g_{pair} defined in §4.7. We defined $\mathcal{D}$ as the linear subspace spanned by all functions on Ω of (at most) two variables. Using an arbitrary basis for $\mathcal{D}$ to produce internal constraints determines p_{pair} as the maximum entropy extension of the empirical distribution (independent of the choice for the basis). However, we are ultimately interested in a smaller subspace $\mathcal{D}^g$ of g-symmetric functions of two variables: $\{V_m\}_{m=1}^M$, an arbitrary basis for $\mathcal{D}^g$ playing the role of internal constraint vectors defines p^g_{pair} via entropy maximization. Here, $M = \dim \mathcal{D}^g$.

Presently we construct the constraint functions[8] $\{V_m\}_{m=1}^M$ by first obtaining V', a large, possibly redundant, set of vectors spanning $\mathcal{D}$. A crude way to do this is by exhibiting indicators of all $\binom{4}{2}L^2 = 6L^2$ possible states $(a, b) \in \mathcal{C}_L \times \mathcal{C}_L$ of every pair of variables (ω_1, ω_2), (ω_1, ω_3), and so on. (Notice that fixing any two variables produces L^2 linear independent members of $\mathcal{D}$; hence the complexity of p_{pairs} is of order L^2.) Thus, we have: $\text{span}\{V'_\alpha\}_{\alpha\in\mathcal{A}} = \mathcal{D}$, and V'_α is of the form $\mathbb{I}_{\{\omega_i=a,\omega_j=b\}}(\omega)$, where α is some suitable multi-index running over $\mathcal{A}$, the set of all the variable pairs with their corresponding states.

[7] We will focus on (constrained) MLE.

[8] Recall (§§4.3,6.2) that we compactly represent functions on Ω as matrix rows.

The notation of §6.3 is helpful to follow the next step: We project each V'_α onto the space of functions on the g-symmetric orbits, i.e. $\pi_1 : V'_\alpha \mapsto \tilde{V}'_\alpha$, where the projection π_1 simply integrates vectors over the g-invariant classes. This results in $\operatorname{span}\{\tilde{V}'_\alpha\}_{\alpha\in\mathcal{A}} = \tilde{\mathcal{D}}^g$, the space of pair-interacting Ω-functions lifted to the space of functions on the g-invariant classes.

We then eliminate possible redundancy of $\{\tilde{V}'_\alpha\}_{\alpha\in\mathcal{A}}$ in order to produce $\{\tilde{V}_m\}_{m=1}^{M}$, a basis for $\tilde{\mathcal{D}}^g$. The elimination is performed numerically and hence is susceptible to errors. In particular, using linear algebraic tools provided by Matlab [49] already may lead to unreliable results in the case of $|\Omega| = 4096$ (i.e. $L = 8$). Supposing the basis is accurate, we get $V_m = \pi_2\tilde{V}_m$, a set of internal constraints defining p^g_{pair} via entropy maximization. recall that the linear pull-back map π_2 divides each component of $\tilde{V}_m$ uniformly over the corresponding class members.

Of course, once the basis $\{\tilde{V}_m\}_{m=1}^{M}$ is obtained, Theorem 6.6 applies immediately to produce $\pi_2\left(\underset{\tilde{q}\in\tilde{Q}^+:\tilde{V}\tilde{q}=\tilde{V}\tilde{p}}{\arg\min} D(\tilde{q},\tilde{s})\right)$, the solution that is generally advantageous from the computational perspective. Recall that $\tilde{s}$ above is the distribution of the g-orbit sizes.

On the other hand, polynomial expressions for bases of $(\mathbb{R}^\Omega)^g$, the space of the g-symmetric functions, might allow one to recognize and to eliminate the dimensions of the higher interactions in a way that is computationally cheaper than the above construction of $\mathcal{D}$. The idea is to use *symbolic* as opposed to *numeric* computations to obtain a polynomial form for some basis of $\mathcal{D}^g = \mathcal{D} \cap \left(\mathbb{R}^\Omega\right)^g$. Computational algebraic geometry (see, for example, [12],[13],[65]) might then provide algorithms to implement this idea, avoiding numerical methods altogether. Consequently, the constraint functions V would be computed exactly.

Avoiding explicit class referencing in candidate probability models may be also more suitable for extending symmetric models to continuous intensity ranges (e.g.

$[0,1]$) A similar but more concrete goal is to minimize the dependence of the models on L: Presently, the symmetry partition $\mathcal{S}$ needs to be recomputed for different values of L.

Finally, in Chapter 7 we discuss extending our models to larger patches. *Explicit enumeration of individual orbits is then infeasible, and the analytic representation of the appropriate* Gibbs potentials *is the only option.*

6.5 In Search for Analytical Representation of p_{symm}

We start by discussing the existence of a vector-valued, polynomial patch function whose range is in one-to-one correspondence with $\mathcal{S}$, the set of G-orbits. For the moment, G stands for an arbitrary subgroup[9] of the appropriate general linear group. Eventually we specialize to the group of our symmetry transformations (§6.4.1) on 2×2 patches.

Note that polynomials allow one to represent any real function on a finite set (specifically, Ω). This is different from our working representation of such functions by real, finite dimensional vectors. The two representations are, of course, closely related, which provides additional flexibility for modeling probability distributions on Ω. The relation between the two types of representations becomes apparent when Ω is identified with a real *affine variety.* Computational algebraic geometry then provides powerful tools for studying various properties of polynomial functions (essentially all the functions) on Ω. However, we will try to keep the following discussion self-contained; in particular, we assume no familiarity with results from algebraic geometry or invariant theory. This does not limit the dis-

[9] This generality allows inclusion of patch transformations other than the ones from §6.4.1, which may in principle be useful for models on larger patches.

cussion of how to search for analytic representations of symmetry-based models. However, computational algebraic geometry is the proper framework to carry out these ideas in practice.

Now, suppose that Ω is indeed embedded in the appropriate real vector space W (e.g., $W = \mathbb{R}^4$ in the case of 2×2 patches). Consider $\mathbb{R}[x_1, \ldots, x_m]$, the *ring* (or, *algebra*) of all polynomials in $m = \dim W$ variables with real coefficients. Alternatively, we sometimes write $\mathbb{R}[x]$ or $\mathbb{R}[W]$. Whenever it is clear from the context, we will denote n-vectors and m-tuples of indeterminates by single letters as with $x \leftrightarrow x_1, \ldots, x_m$. The action of G on W by matrix multiplication (6.10) corresponds to G acting on $\mathbb{R}[W]$ as follows:

$$(gf)(v) = f(\rho(g^{-1})v), \quad \text{where} \;\; g \in G \quad \text{and } f \in \mathbb{R}[W], \tag{6.11}$$

where the linear representation ρ was introduced in §6.4.3 in the context of (6.10). The subring of $\mathbb{R}[W]$ consisting of all polynomials that are invariant under this G action will be denoted by $\mathbb{R}[W]^G$. (Also, when using superscripts with objects invariant under some group action, we may often reference only the group provided the action is clear from the context.)

Definition 6.9 Polynomials $f_1, \ldots, f_N$ from $\mathbb{R}[W]^G$ are said to be *fundamental integral invariants* associated with the above G-action on $\mathbb{R}[W]$ if any other G-invariant polynomial f ($\in \mathbb{R}[W]^G$) can be expressed as a polynomial in $f_1, \ldots, f_N$. We will also refer to such fundamental invariants as *generators*.

The following well-known fact is fundamental for our discussion and follows from more general results in *Invariant Theory* [12],[52], [64] and [65]. Nonetheless, below we give a short, basic proof that also leads to a construction of G-invariant class indicators in terms of fundamental G-invariants (Theorem 6.15).

Proposition 6.10 Let W be an m-dimensional real vector space ($W \cong \mathbb{R}^m$ with the standard basis) and let ρ be a (m-dimensional) representation of a finite group G, $\rho : G \hookrightarrow GL(W)$. Then there is a bijection between $\mathcal{S}_{\mathbb{R}} \overset{\mathrm{def}}{=} W/G$, the orbit set of the G action on W (associated with ρ), and the image of $(f_1, \ldots, f_N) : W \to \mathbb{R}^N$, fundamental invariants of ρ.

Proof. First observe that, indeed, there always is a finite system of such generators. This fact was proved by Hilbert for fields of characteristic zero, and later extended for certain fields of positive characteristic by Noether ([20] and [64]). We briefly comment on the general problem of exhibiting such generators later in §6.5.1, where we find them for our special case $\mathbb{R}[x_1, \ldots, x_4]^G$.

The G-invariance of $f_1, \ldots, f_N$ means that these functions are constant on the orbits of $\mathcal{S}_{\mathbb{R}}$. Thus we have a well-defined map from $\mathcal{S}$ onto the image of $(f_1, \ldots, f_N)$. Therefore, we need only prove that, given any two distinct orbits $\mathcal{O}_1, \mathcal{O}_2 \in \mathcal{S}$, the corresponding values of $(f_1, \ldots, f_N)$ must be distinct. We show this by exhibiting a G-invariant polynomial f that takes distinct values on $\mathcal{O}_1$ and $\mathcal{O}_2$, and then conclude that the values assumed by at least one of the N generators on these orbits must be distinct.

The finite size of the orbits allows the following crude construction of f:

$$\tilde{f}(x) = \prod_{g \in G} \sum_{l=1}^{m} \left[x_l - \rho(g)(\omega)_l\right]^2, \qquad \omega \in \mathcal{O}_1 \tag{6.12}$$

$$f(x) = \sum_{g \in G} (g\tilde{f})(x), \tag{6.13}$$

where $g\tilde{f}$ is computed according to (6.11). The definition (6.12) ensures that $\tilde{f}(v) = 0$ (and consequently $f(v) = 0$) if and only if $v \in \mathcal{O}_1$. In (6.13) , we average $\tilde{f}$ over its orbit in order to guarantee G-invariance. Note that f separates $\mathcal{O}_1$ from the rest of the orbits, since for each $g \in G$ the only roots of $g\tilde{f}$ are the points in $\mathcal{O}_1$. In particular, f assumes distinct values on $\mathcal{O}_1$ and $\mathcal{O}_2$. ◇

Remark 6.11 Up to normalization by the group order, the operator in (6.13) is none other than the *Reynolds operator*. Recall (§6.2) that we defined essentially the same operator in (6.7), the only difference being that earlier the functions on Ω were represented by real vectors rather than by polynomials.

The following definition that appears in [12] extends the discussion to $k[x_1, \ldots, x_m]$, polynomials over an arbitrary field k of zero characteristic:

Definition 6.12 Given a finite matrix group $G \subset GL(m, k)$, where the field k has characteristic zero, the following k-linear map $\mathcal{R}_G : k[x_1, \ldots, x_m] \to k[x_1, \ldots, x_m]$ is called the Reynolds operator:

$$\mathcal{R}_G(f) = \frac{1}{|G|} \sum_{g \in G} gf, \quad \text{for} \quad f \in k[x_1, \ldots, x_m] \tag{6.14}$$

The averaging feature of this operator is formally expressed by the fact that $\mathcal{R}_G(f) \in k[x_1, \ldots, x_m]^G$ $\forall f \in k[x_1, \ldots, x_m]$. The following property further underlines the correspondence with probabilistic averaging: $\forall f \in k[x_1, \ldots, x_m]$ and $\forall h \in k[x_1, \ldots, x_m]^G$, $\mathcal{R}_G(hf) = h\mathcal{R}_G(f)$. The probabilistic interpretation is that a random variable which is measurable relative to the σ-algebra on which conditioning is performed can be factorized through the conditional expectation.

6.5.1 Fundamental G-invariants

We now specialize to our group of symmetries generated by the geometric transformations (r and s) and the intensity inversion (i). Before we propose a particular set of invariant generators for $\mathbb{R}[x_1, x_2, x_3, x_4]^G$, let us recall that, according to (6.10) and (6.11), the G action on $\mathbb{R}[x_1, x_2, x_3, x_4]$ can be concisely expressed via the action of r, s, i, generators of G, on x_1, x_2, x_4, x_4, canonical generators of $\mathbb{R}[x]$:

$$\begin{array}{llll} rx_1 = x_2; & rx_2 = x_3; & rx_3 = x_4; & rx_4 = x_1; \\ sx_1 = x_1; & sx_2 = x_4; & sx_3 = x_3; & sx_4 = x_2; \\ ix_k = -x_k,\ k = 1, 2, 3, 4 & & & \end{array} \tag{6.15}$$

Theorem 6.13 The following set of polynomials generates $\mathbb{R}[x_1, x_2, x_3, x_4]^G$:

$$\begin{aligned} f_1(x) &= (x_1 + x_3)(x_2 + x_4), \\ f_2(x) &= x_1x_3 + x_2x_4, \\ f_3(x) &= x_1^2 + x_2^2 + x_3^2 + x_4^2, \\ f_4(x) &= x_1x_2x_3x_4, \\ f_5(x) &= (x_1^2 + x_3^2)(x_2^2 + x_4^2). \end{aligned} \tag{6.16}$$

A proof of the theorem is given in §D.1. Although standard algorithms exist to compute such generating sets in a systematic fashion (see, for example, [64] and [65]), we base our proof on a very intuitive approach, which, in particular, does not require familiarity with algebraic geometry or invariant theory. The general theory (e.g., [64]) also shows that $m \leq N \leq \binom{m+|G|}{|G|}$. In our case the above upper bound, due to Noether, is $\binom{4+16}{16} = 4845$. This is too large for a direct implementation of the corresponding algorithm to find such generators. Our case turns out to be special, however, in that we nearly achieve the lower bound determined by $\dim \mathbb{R}^4 = 4$. This small number of generators encourages one to use them in practice for orbit-indexing.

Also, we obviously need not "index" the entire $\mathcal{S}_{\mathbb{R}}$, which is in fact uncountable. For any L, we need only index $\mathcal{S}_L$, a finite subset of all real G-orbits. For this task some of the above generators turn out to be redundant as can be seen from the following statements, which can be verified algorithmically.

Proposition 6.14 For the even size intensity ranges $\mathcal{C}_L$ with $L \leq 12$ and for the ranges of size $L = 3, 5, 7$, the four generators f_1, f_2, f_3, f_4 (6.16) suffice to enumerate all the orbits of $\mathcal{S}_L = M_{2\times 2}(\mathcal{C}_L)/G$. Moreover, for $L = 2, 4, 6$, or 8, the three generators f_1, f_2, f_4 are sufficient.

An immediate conjecture is that f_1, f_2, f_3, f_4 are sufficient in general if L is even. In Appendix D we briefly discuss an analytical approach to verification of the main results and the conjecture.

Finally, we summarize our reasoning. We state results only for our special choice of G but generalizations to other linear groups acting on Ω are apparent.

Theorem 6.15 Any strictly positive G-invariant probability mass function on Ω admits the Gibbs form in terms of the invariant generators only:

$$p^G(\omega) = \exp\left(\sum_{|\alpha|\leq M} a_\alpha f_1^{\alpha_1}(\omega) \times \cdots \times f_5^{\alpha_5}(\omega)\right), \tag{6.17}$$

where $\alpha = (\alpha_1, \ldots, \alpha_5)$ is a multi-index of nonnegative integer components and $|\alpha| = \alpha_1 + \ldots + \alpha_5$. A crude bound M on the maximum degree is $2(|\mathcal{S}_L| - 1) \max_{n=1,\ldots,5} \deg f_n = 8(|\mathcal{S}_L| - 1)$.

Proof. One can quickly exhibit one such representation expressing G-orbit indicators in terms of the fundamental generators $f_1, \ldots, f_5$ and then substituting the results in the exponential representation based on the indicators (6.20). First, let $\omega^* \in \mathcal{O} \in \mathcal{S}$ and write $[\omega]$ for the orbit represented by ω. We mimic (6.12) and (6.13) to produce:

$$\tilde{f}(x) = \prod_{\substack{\mathcal{O}' \in \mathcal{S} \\ [\omega] \neq \mathcal{O}}} \left[(f_1(x) - f_1(\omega))^2 + \cdots + (f_5(x) - f_5(\omega))^2\right] \tag{6.18}$$

$$\mathbb{I}_{\mathcal{O}}(x) \quad = \quad \frac{\tilde{f}(x)}{\tilde{f}(\omega^*)}. \tag{6.19}$$

The result follows upon substituting (6.19) into

$$p^G(\omega) \quad = \quad \exp\left(\sum_{\mathcal{O}\in\mathcal{S}} \gamma_{\mathcal{O}} \mathbb{I}_{\mathcal{O}}(\omega)\right), \tag{6.20}$$

where $\gamma_{\mathcal{O}} = \log \frac{p^G(\mathcal{O})}{|\mathcal{O}|}$, $\mathcal{O} \in \mathcal{S}$, are the model parameters. The degree bound follows immediately from this construction and from (6.16). ◇

Recall that for intensity ranges of even size, $|\mathcal{S}_L|$ was computed in Proposition 6.8. Therefore, the above upper bound for $L = 8$ is already large: $M = 8 \times 345 = 2760$. Although Proposition 6.14 suggests that only four generators are required if L is even, the value of the direct construction used in the proof above is still in proving the existence rather than providing an efficient construction. One way to make these ideas more useful (via approximations) is discussed next in §6.6. However, the fact that the complexity of the general symmetry models is of order L^4 strongly argues for focusing in practice on symmetric subfamilies of lower complexities. Pair-potential families are one example; recall that their complexity is of order L^2. Consequently, a relatively small fraction of the exponential terms in (6.17) need to be retained in order to represent p^G_{pair}. Computational algebraic geometry becomes indispensable for obtaining efficient algorithmic characterizations of the "low-interaction" terms in that case. Finally, this presents an interesting direction for additional research.

6.6 Sequential MEE

Restricting the degree of interaction among the pixels in the patch is one way to lower the complexity of models such as p^g_{symm}. Recall (Chapter 5) that we have found evidence that $p \in p^g_{pair}$. This suggests that parameter spaces of dimensions lower than that of p^g_{symm} may generally be sufficient for accurate estimation of p. In §5.1 we commented on the leading role of applications in accepting particular models. Thus, a reasonable scenario for a multi-stage modeling might start by accepting constraints "surely" satisfied by p (i.e., not rejected by the appropriate statistical tests) and then provide a convenient mechanism to index families of candidates for further, application-driven reduction of the model complexity. By "convenience" we mean, first of all, flexibility to expand the model parameter space on "demand": For example, higher accuracy and, consequently, a less restrictive model, may be requested if, say, the application receives additional resources. The application, of course, may itself be a subtask of a larger optimization problem.

It is conceivable that in a real situation one generally finds only "few" truly satisfied constraints. Recall that in the case of the models based on the geometric and intensity inversion symmetries, the number of constraints is $|\Omega| - |\mathcal{S}|$ (§6.1). Although this number is of order L^4, the order of $|\Omega|$, the complexities of the models are of the same order. Thus, little reduction of dimensionality is achieved. (Hence the quotes in "few".) One can then argue that for large L, the constrained family is still *non-parametric* since its dimension is of the same order as the dimension of the saturated model. Perhaps a proper term for pairwise interacting Gibbs models, whose complexity is of order L^2, should then be *semi-parametric*. Unfortunately, "truly" parametric models that have complexities independent of L are likely to be too restrictive as, for example, p_{potts}.

We now outline a variation of *minimax learning* [68],[69],[70] that provides a mechanism for incremental expansion of model spaces. The expansion occurs entirely within the symmetric family accepted on the basis of statistical hypothesis testing. (An example is p^g_{symm}.) Based on particular goals, several modifications of the baseline program are conceivable. One subtle distinction from the mainstream of minimax learning applications is the purely algebraic nature of our filter banks: We are going to consider log-linear models based on linear spaces spanned by monomials $f_1^{\alpha_1}(\omega) \times \cdots \times f_N^{\alpha_N}(\omega)$ in fundamental generators of the underlying symmetry group.

Suppose some symmetry group G induces $\mathcal{S}$, a partition of Ω into G-invariant classes. Assume that a set of G-invariant fundamental polynomial generators $f_1(\omega), \ldots, f_N(\omega)$ is available. Assume also that $B = \{F_1, \ldots, F_{|\mathcal{S}|}\}$, a basis for the G-symmetric functions $(\mathbb{R}^\Omega)^G$ is found, where $F_s,\ s = 1, \ldots, |\mathcal{S}|$ is the evaluation on Ω of some monomial $f_1^{\alpha_1}(\omega) \times \cdots \times f_N^{\alpha_N}(\omega)$. Note that such B always exists: Expressions in (6.18) provide one example. Also note that to each of the $2^{|\mathcal{S}|}$ subsets of B there corresponds a linear subspace spanned by vectors of that subset. One of such subsets, for example, spans the subspace of pair-interacting G-invariant functions, similar to $\mathcal{D}^g$ discussed in §6.4.3. There may be many other potentially suitable subspaces of dimensions significantly lower than $\dim B = |\mathcal{S}|$.

Suitability of these subspaces for capturing p may, for example, be assessed by power-divergence tests applied to the corresponding log-linear models, similar to our tests for p^g_{pair}. Application-dependent adjustments for model complexity are certainly an option. However, a global search for suitable models of this form is clearly infeasible as $2^{|\mathcal{S}|}$ is likely to be prohibitively large in practice. If, however, C is specified as a bound on the subspace dimension (hence, model complexity) and is small relative to $\dim B$, it may be feasible to search for a “best” subspace of $\dim \leq C$.

A local search is another alternative: Start with $B_0 = \emptyset$ and suppose in the n^{th} step a subspace B_n has been chosen as best in the sense of, for example, minimizing $D(\hat{p}, \hat{p}_{MEE,B_n})$, Kullback-Leibler divergence from the empirical distribution

$\hat{p}$ (hence, "minimax"). In the next step, B_{n+1} is created by adjoining F^*_{n+1} to B_n according to

$$F^*_{n+1} \quad = \quad \arg\min_{F \in B \setminus B_n} D(\hat{p}, \hat{p}_{MEE, B_n \cup \{F\}}). \tag{6.21}$$

Note that at the end ($n = |\mathcal{S}|$), $\hat{p}_{MEE,B_n}$ would always become $\hat{p}_{MLE,\Theta_0}$, the symmetry-constrained MLE, where Θ_0 is the set of all positive G-symmetric distributions.

We performed similar experiments based on this greedy strategy in the case of G-symmetries and $L = 4$ and $L = 8$. The goal, of course, was not to produce $\hat{p}_{MLE,\Theta_0}$, which is cheaply available as $\mathcal{R}_G\hat{p}$ in these cases. Instead, we wanted to see which F^*'s would be selected in the first steps. In fact, we did not limit our B to bases for the space of G-symmetric functions on Ω but instead allowed redundant sets of monomials in $f_1, \ldots, f_5$ (Theorem 6.16)[10]. (As a purely technical observation, it can be shown that the algorithm maintains linear independence of B_n's automatically, which also implies that $F \in B \setminus B_n$ in (6.21) can be replaced by $F \in B$.)

These experiments were based on an image set much smaller than our present $\mathcal{I}_{im}$ and may need to be repeated to verify the results based on a more stable $\hat{p}$. Nonetheless, they demonstrated certain consistency of the greedy strategies for the two quantizations after the intensity ranges were appropriately rescaled for standardization. Recall also that in finding MEE one solves numerically for γ's, the Lagrange multipliers (§4.3). Significant regularity was also observed in our experiments in the sense that adding new F^*'s resulted in rapidly decaying adjustments of the previously computed γ's. Finally, a variation of this approach might also be useful for "on-line" generation of suitable models of prescribed complexity.

[10] We also explored several different *monomial orders* [12] in our experiments.

CHAPTER 7

CONCLUSION

We have extensively studied microscopic parts of digital, grey level images of natural scenes. Mathematically, we represent such *natural microimages* by $\omega \in \Omega$, matrices of quantized raw image intensities. The main object of our investigation is p, the unknown probability distribution induced on Ω by the natural microimage population Σ. *We have built a proper statistical basis for meaningful analysis and modeling of p* (Chapter 3). Namely, we base our statements involving p on the grounds of statistical hypothesis testing (Chapters 3 and 5). Refuting scepticism about meaningfulness of image analysis based on raw intensities, and despite sampling limitations, we demonstrate several non-trivial properties of the natural image microworld. The list begins with *inter-scene stability.* This is precisely the property of the natural microimage distribution to allow estimation from relatively small samples that is nearly independent of global attributes of natural scenes.

Due to relatively small microimage samples available, we focus on two- and four-pixel patches and rather coarse intensity quantizations. These limitations are *not principal* since they are not necessitated by computational issues. In fact, rapidly increasing availability of large databases of natural images imminently makes such limitations obsolete. Consequently, we anticipate rapid intensification of investigations similar to ours in the near future. Such investigations are likely to focus on *specific imaging domains* (e.g., particular terrains, artistic photography, or *range* imagery). In this regard, our present experience provides valuable information that relates the amount of detail one could infer about the natural image microworld to (micro)image sample sizes and other sampling issues.

Our local and coarse microimage modeling setting embeds naturally in a more general, global framework of *random fields*. For this reason, we view our microimage measure $p = p(\omega; m, n, L)$ (albeit often implicitly) as a function of both $m \times n$, the dimensions of the microimage support, and L, the size of the intensity range. In order to better understand the place of this work in the field of natural image statistics we recall that the fundamental hypothesis of this field is the existence of $\mathbf{P}$, a *universal* probability measure explaining formation of images of natural scenes. The meaning attached to such "explanation" is rather profound and may itself require an explanation. The idea is, of course, to model families of natural (micro)image populations (such as our Σ) in such ways that the appropriate modeling frameworks are always consistent with (or "derivable" from) $\mathbf{P}$. A simple example to illustrate the idea is as follows: Suppose $\mathbf{P}$ is in some family of a continuously-valued, continuously-supported random fields. We want to be able to deduce properties of discrete, local subfields of $\mathbf{P}$ from $\mathbf{P}$. And conversely, we also want to use our coarse and local model-based observations to update our knowledge about the family containing $\mathbf{P}$.

Also, one seeks an abstract common ground (in the form of families of $\mathbf{P}$) so that, not only would it be consistent with more concrete, already working mathematical frameworks, but it would also yield these frameworks *consistent with one another*. Furthermore, the desired consistency must be insensitive to variations in transformations (i.e., calibration of imaging devices, preprocessing) which finer or more global measures undergo in the course of producing their coarser or local relatives, respectively. Evidently, providing appropriate imaging applications with a clear, universal hierarchy of computationally-developed alternatives is at the core of this theory.

This work reflects our belief that the natural approach to fulfilling such an ambitious program is "bottom-up", more precisely, "coarse-to-fine" and "local-to-global". Thus, starting at the "bottom", we have discovered the *inter-scene stability* of $p(\cdot;1,2,8)$, $p(\cdot;2,1,8)$, and $p(\cdot;2,2,4)$, which is an example of inter-measure consistency referred to above. Another type of such consistency is *scale invariance*, which, with the exception of $\mathbf{P}$, can only be defined in approximate terms. *We also exhibit virtual indistinguishability between estimates of* p *taken at two different scales.* Unlike in other research on scale invariance in natural images, including our early efforts, *these results are more rigorous due to the hypothesis testing framework in which they have been obtained.*

Our findings at these levels of spatial localization and intensity quantization provide us with tools and experience necessary to extend and expand $(\Omega, p(\cdot; m, n, L))$, the subject of our analysis and modeling, to the natural distributions on microimages with larger spatial supports and finer intensity ranges. Moreover, the discovery of *geometric symmetries* and *distinct patterns of pixel interaction in the patch* (Chapter 5) allows us to focus future research on a considerably smaller subset of microimage distributions that respect such symmetries and patterns of pixel interactions.

Also, our findings complement similar research in which symmetries of multivariate (up to tri-variate) linear filter-response statistics of two-, and four-pixel patches of fine intensity natural images are reported [33], [34], [35].

As a future project that could further enrich natural microimage phenomenology, we imagine performing our hypothesis testing experiments on synthetic images generated, for example, according to Poisson disc models [42]. An expectation is that synthetic microimages corresponding to models with the "right" scaling parameters will pass two-sample consistency tests when tested in conjunction with natural microimages.

Obviously, we do not limit our investigation to phenomenology and are concerned with computationally efficient applications of the discovered properties of the microimage distribution. A particular framework is that of defining local image features using tree structured *vector quantization* of microimages (Chapter 2). A possible direction for future research is to consider globally optimal strategies [21]. Replacing on-line microimage classification by an off-line VQ computation based on the models presented in this work defines another dimension for the evolution of this research.

Certainly, the inter-scene stability observed in this work is not an absolute law, and with finer intensity quantizations the difference in microimage statistics obtained from samples of disparate imagery may register as significant. Not only would this not diminish our modeling results[1], but it would also open perspectives for their application in, for example, image segmentation. In fact, model-based and application-oriented analysis of local statistics of segmented natural images is already in progress [33],[35]. Discrimination of terrains or textures based on our models is, of course, also conceivable, since models result in more robust estimation.

Naturally, a particular application will require revision of our testing results in order to adjust the balance between the model complexity and the model accuracy in the manner optimal for the application. Perhaps, even the primitive models

[1] We strongly believe that universality of models based on fundamental properties of the microworld (e.g., the geometric symmetries) extends to intensity quantizations much finer than those tested in this work.

strongly rejected by our tests may prove adequate for some tasks wherein a higher accuracy is unnecessary or a higher complexity is unaffordable.

Whether it is supervised microimage classification or any other practical context, extending our models to larger patches is an imperative. Direct replication of (manual or computer-based) computations of symmetric models may already prove difficult for patches as large as 4×4. However, such computations may also be unnecessary. In order to substantiate this last hypothesis, we are prepared to advance our preliminary experiments on modeling microimage distributions on large patches as *maximum entropy extensions* under constraints that equate their 2×2 marginals to certain 2×2 model distributions (e.g., p^G_{pair}). The idea behind such intentions is as follows. The general theory of *Gibbs - Markov Random Fields* (GMRF) [14],[17],[25],[26],[29],[38],[41],[56] explains that in order to correctly specify a discrete Markov random field on a finite lattice with the square eight-site neighborhood system (second-order neighbors), it is essentially sufficient to define the *Gibbs potential* on the *cliques*, which, in this case, are all 15 non-empty subsets of the 2×2 patch (*maximal clique*). “Essentially” refers to the extra conditions that such definitions must satisfy in order for the resulting Gibbs - Markov field to be translation invariant. These are precisely the conditions of *horizontal* and *vertical reflection* invariance that we have accepted in testing our microimage distribution $p(\cdot, 2, 2, 4)$. These conditions also yield the maximum entropy characterization.

Based on our tests, p is also rotationally invariant, which allows us to consider a smaller class of *isotropic* GMRF’s. Finally, the most restrictive hypothesis we have accepted (Chapter 5) is that p is geometrically invariant and depends only on pair-wise interactions of pixels in the patch. Under this hypothesis, one- and two-pixel cliques completely characterize GMRF’s. Thus, we have one readily available class of models for the natural distribution on large microimages.

We have also argued in Chapter 6 that explicit functional forms for Gibbs potentials are highly desirable for efficient applications of such models. In fact, we have proved the existence of representations of Gibbs potentials for symmetric models in terms of appropriate *invariant polynomial generators* and outlined an efficient algorithm to compute such representations. A set of invariant polynomial generators has been computed for the model with geometric and intensity inversion symmetries. Such invariant generators can similarly be found for all our less restrictive symmetry models. However, without the *pair-wise interaction* condition, general symmetry models are still too complex to be useful in the GMRF framework of modeling p on larger lattices. Finally, the rapid development of computational algebraic geometry over the last decade has brought about a multiplicity of symbolic algorithms that may prove useful for designing efficient methods to control the degree of interaction in polynomial Gibbs potentials with desired symmetries.

APPENDIX A

STATISTICAL AND INFORMATION THEORETIC PRELIMINARIES

For a detailed presentation of *Information Theory* and its relation to *Statistic* see, for example, [11] and [40]. Let p be a discrete probability measure on a finite set Ω: $p \in \Theta = \{q \in \mathbb{R}^K : q_k \geq 0,\ k = 1, \ldots, K,\ \sum_{k=1}^{K} q_k = 1\}$, where $K = |\Omega|$.

Definition A.1 The *Shannon Entropy* is

$$H(p) = -Const \sum_i p_i \ln(p_i), \quad Const > 0$$

with the convention to extend the function $x \ln(x)$ by continuity at 0. When $Const = 1/\log_2 e$, (i.e. the base 2 log is used), the measurement unit is *bits*. When the choice of $Const$ is not important we write *log* without specifying its base.

Definition A.2 The entropy of a discrete random variable (vector) X is defined to be the entropy of the (joint) probability distribution of X:

$$H(X_1, \ldots, X_n) = -\sum_{x_1,\ldots,x_n} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \log(\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n))$$

Definition A.3 The conditional entropy of X given an event B of positive probability is:

$$H(X|B) = H(\mathbb{P}_{X|B}) = -\sum_x \mathbb{P}(X = x|B) \log(\mathbb{P}(X = x|B))$$

Definition A.4 The conditional entropy of X given another random variable Y is:

$$H(X|Y) = -\sum_{x,y} \mathbb{P}(X = x, Y = y) \log(\mathbb{P}(X = x|Y = y))$$

Definition A.5 The *Kullback-Leibler Divergence* or pseudo-distance from p to q, discrete probability distributions and on a common set, is defined by:

$$D(p, q) = \sum_i p_i \log \frac{p_i}{q_i}$$

We now state and prove the well-known equivalence between constrained entropy maximization and the maximum likelihood parameter estimation in the corresponding log-linear model [54],[70].

Proposition A.6 Let $\hat{p}_k = \frac{n_k}{N}$ be a positive empirical distribution, where n_k is the number of observations of the k^{th} state in a fixed size sample of N observations. Let V be a $(J + 1) \times K$ constraint matrix with columns corresponding to the states in Ω. Let $V_0, V_1, \ldots, V_J$, the rows of V, be linearly independent; in particular $J + 1 \leq K$. Suppose also that $V_0 \equiv 1$ corresponds to the normalization constraint. Define $F(\hat{p})$ as $\{q \in \Theta^+ : \ \mathbb{E}_q V_j = \mathbb{E}_{\hat{p}} V_j, \ j = 1, \ldots, J\}$, where Θ^+ is the set of positive distributions on Ω. (Equivalently, $F(\hat{p}) = \{q \in Q^+ : \ Vq = V\hat{p}\}$, where Q^+ is the positive quadrant of $\mathbb{R}^K$.)

Let $\hat{p}_{MEE,V} = \arg \max\limits_{q \in F(\hat{p})} H(q)$ be the unique maximum entropy extension of $\hat{p}$ (§4.3). Thus (4.6),

$$\hat{p}_{MEE,V\ k} = \exp\left(\gamma_0 - 1 + \sum_{j=1}^{J} \gamma_j V_{jk}\right), \tag{A.1}$$

where the parameters $\gamma_0, \ldots, \gamma_J$ are uniquely determined by the constraints:

$$V\hat{p}_{MEE,V}(\gamma_0, \ldots, \gamma_J) = V\hat{p}.$$

Let Θ_0 be the corresponding log-linear model:

$$\begin{aligned}\Theta_0 &= \{q \in \mathbb{R}^K : q_k = e^{\gamma_0 - 1 + \sum\limits_{j=1}^{J} \gamma_j V_{jk}},\ \gamma \in \Gamma\},\\ \Gamma &= \{\gamma \in \mathbb{R}^{J+1} : \sum_{k=1}^{K} e^{\gamma_0 - 1 + \sum\limits_{j=1}^{J} \gamma_j V_{jk}} = 1\}.\end{aligned}$$

Finally, let $\hat{p}_{MLE,\Theta_0}$ be the Θ_0-constrained maximum likelihood estimator:

$$\hat{p}_{MLE,\Theta_0} = \arg\max_{q \in \Theta_0} \sum_{k=1}^{K} n_k \log q_k = \arg\max_{q \in \Theta_0} \sum_{k=1}^{K} \hat{p}_k \log q_k \tag{A.2}$$

Then,

$$\hat{p}_{MEE,V} = \hat{p}_{MLE,\Theta_0}$$

Proof. In §4.3 we discussed the existence and uniqueness of $\hat{p}_{MEE,V}$ and now show that if p_{MLE,Θ_0} exists, then the two are indeed equal. The existence of p_{MLE,Θ_0} in this context is also a well-known fact and we will give its short proof as well.

First, we reparametrize Γ by $g : \mathbb{R}^J \to \Gamma$ as follows:

$$g(\gamma_1, \ldots, \gamma_J) = (1 - \log \sum_{k=1}^{K} e^{\sum\limits_{j=1}^{J} \gamma_j V_{jk}}, \gamma_1, \ldots, \gamma_J).$$

Consequently, this reparametrizes Θ_0 as

$$\begin{aligned}\Theta_0 &= \{q(\gamma_1, \ldots, \gamma_J) \in \mathbb{R}^K : q_k = \frac{e^{\sum\limits_{j=1}^{J} \gamma_j V_{jk}}}{Z(\gamma_1, \ldots, \gamma_J)},\ (\gamma_1, \ldots, \gamma_J) \in \mathbb{R}^J\},\\ Z(\gamma_1, \ldots, \gamma_J) &= \sum_{k=1}^{K} e^{\sum\limits_{j=1}^{J} \gamma_j V_{jk}},\end{aligned}$$

which allows us to transform the Θ_0-constrained maximization of the log-likelihood function (A.2) into the equivalent unconstrained maximization of the same function below:

$$\hat{p}_{MLE,\Theta_0} = \arg\max_{(\gamma_1, \ldots, \gamma_J) \in \mathbb{R}^J} \left(-\sum_{k=1}^{K} n_k \log \sum_{k=1}^{K} e^{\sum\limits_{j=1}^{J} \gamma_j V_{jk}} + \sum_{k=1}^{K} n_k \sum_{j=1}^{J} \gamma_j V_{jk} \right)$$

The necessary conditions for an extremum to occur at $(\gamma_1^*, \ldots, \gamma_J^*)$ require setting to 0 all J partial derivatives $\frac{\partial}{\partial \gamma_j}$ of the parametrized log-likelihood function, which results in:

$$-N \frac{\sum\limits_{k=1}^{K} V_{jk} e^{\sum\limits_{j=1}^{J} \gamma_j^* V_{jk}}}{\sum\limits_{k=1}^{K} e^{\sum\limits_{j=1}^{J} \gamma_j^* V_{jk}}} + \sum_{k=1}^{K} n_k V_{jk} = 0, \qquad j = 1, \ldots, J. \tag{A.3}$$

Dividing both sides of (A.3) by N shows that if $q(\gamma_1^*, \ldots, \gamma_J^*)$ exists, it satisfies $\mathbb{E}_{q(\gamma_1^*,\ldots,\gamma_J^*)} V_j = \mathbb{E}_{\hat{p}} V_j,\ j = 1, \ldots, J$. Since the same equations characterize $\hat{p}_{MEE,V}$, the existence and uniqueness of $\hat{p}_{MEE,V}$ automatically gives the existence and uniqueness of $q(\gamma_1^*, \ldots, \gamma_J^*)$ that is apparently equal to $\hat{p}_{MEE,V}$.

However, in order to regard $q(\gamma_1^*, \ldots, \gamma_J^*)$ as $\hat{p}_{MLE,\Theta_0}$, we still need to show that $q(\gamma_1^*, \ldots, \gamma_J^*)$ is indeed corresponds to a global maximum of the log-likelihood function.

Taking the second derivatives $\frac{\partial^2}{\partial \gamma_i \partial \gamma_j}$ of the reparametrized log-likelihood function produces a well-known result: $\mathcal{H} = -N \text{Cov}_{q(\gamma_1,\ldots,\gamma_J)}(\tilde{V}, \tilde{V})$, where $\mathcal{H}$ is the *Hessian* matrix evaluated at $(\gamma_1, \ldots, \gamma_J)$ and $\text{Cov}_{q(\gamma_1,\ldots,\gamma_J)}(\tilde{V}, \tilde{V})$ is the covariance matrix of the random vector $\tilde{V} = (V_1, \ldots, V_J)$ relative to $q(\gamma_1, \ldots, \gamma_J)$. Finally, we show that $\mathcal{H}$ is indeed negative definite for any $(\gamma_1, \ldots, \gamma_J) \in \mathbb{R}^J$, the sufficient condition for $q(\gamma_1^*, \ldots, \gamma_J^*)$ to be the global maximum of the log-likelihood function.

Suppose $\xi \in \mathbb{R}^J$ and $\xi \neq 0$ and let us suppress the subscript $q(\gamma_1, \ldots, \gamma_J)$ in the covariance matrix and other ensuing expectations. Next, we show that the quadratic form $\sum\limits_{1 \le i,j \le J} \text{Cov}(V_i, V_j) \xi_i \xi_j$ is strictly positive (for all $(\gamma_1, \ldots, \gamma_J) \in \mathbb{R}^J$):

$$\sum_{1 \le i,j \le J} \text{Cov}(V_i, V_j) \xi_i \xi_j = \sum_{1 \le i,j \le J} \text{Cov}(\xi_i V_i, \xi_j V_j) =$$

$$\sum_{1 \le i,j \le J} \mathbb{E}\,(\xi_i V_i - \mathbb{E}\xi_i V_i)(\xi_j V_j - \mathbb{E}\xi_j V_j) = \mathbb{E} \sum_{1 \le i,j \le J} (\xi_i V_i - \mathbb{E}\xi_i V_i)(\xi_j V_i - \mathbb{E}\xi_j V_j) =$$

$$\mathbb{E}\left[\sum_{j=1}^{J}(\xi_j V_j - \mathbb{E}\xi_j V_j)\right]^2 \geq 0$$

Since $\Theta_0 \subset \Theta^+$, the only way for the equality in the last expression to occur is if $\sum_{j=1}^{J} \xi_j V_j = \sum_{j=1}^{J} \mathbb{E}\xi_j V_j$. But this would imply linear dependence of $V_0, \ldots, V_J$, contradicting our assumption of the linear independence of these vectors.

$\diamond$

APPENDIX B

POWER-DIVERGENCE STATISTICS AND TESTS

This chapter discusses the Power-Divergence Statistics and based on them statistical tests [54]. The unifying entity for this family of statistics is the one-parameter family of the *Power-Divergence* quasi-distance measures $I(p, q; \lambda)$[1] between two probability vectors p and q:

$$I(p, q; \lambda) = \frac{1}{\lambda(\lambda+1)} \sum_k p_k \left[\left(\frac{p_k}{q_k} \right)^{\lambda} - 1 \right] \quad \lambda \in \mathbb{R}, \quad \lambda \neq -1, \ \lambda \neq 0 \qquad \text{(B.1)}$$

Extending the above definition to $\lambda = -1$ and $\lambda = 0$ by continuity, we obtain $I(p, q; -1) = D(q, p)$ and $I(p, q; 0) = D(p, q)$, the usual Kullback-Leibler divergences (up to the base of the logarithm) [11],[40], Definition A.5. Positivity of $I(p, q; \lambda)$ is perhaps the most basic among the properties making $I(p, q; \lambda)$ into a distance-like measure [54]. This property is a simple consequence of Jensen's inequality applied in combination with the convexity of $x^{\lambda+1} - 1$ (in the general case of $\lambda \neq -1$ and $\lambda \neq 0$) and $-\log x$ (when $\lambda = -1, 0$).

The general hypothesis testing framework for the power-divergence statistics based on the measures above is essentially determined by that of modeling by external constraints 4.2.1,5.1. Namely, Ω is a finite state space equipped with Θ^+, the set of positive probability distributions on Ω. If Ω represents a real system

[1] In fact, $I(p, q; \lambda)$ is a straightforward modification of the *Hellinger integral* (see, for example, [63]), from which it inherits most of its properties.

to be modeled probabilistically, i.e., by identifying it with a (partially) unknown distribution $p \in \Theta^+$, then statements about such a system are formulated in terms of H_0 and H_a, the *null* and *alternative* hypotheses respectively:

- $H_0 : \; p \in \Theta_0 \subset \Theta^+$, and $d_0 = \dim \Theta_0 < d = \dim \Theta^+$
- $H_a : \; p \in \Theta_a \subset \Theta^+$, and $d_a = \dim \Theta_a$

Generally, Θ_0 is represented by a parametrization $f : \; \Gamma \to \Theta_0$. (Since Ω is finite, an exponential parametrization is always available (5.1).) The only exponential parametrizations we use in the present work are those arising from entropy maximization under internal constraints (§§4.2.2,4.3, Proposition A.6) and resulting in the log-linear models (4.1). Thus, f is then composed of exponential and linear maps. With the symmetry-based models, we use linear parametrizations in which the parameters are the probability values on the symmetry classes (§§5.1,6.1). Under these latter parametrizations $f = \pi_2$, using the notation of §6.2 to denote by π_2 the map that divides orbit masses uniformly among orbit elements.

We also restrict our hypothesis testing to the following two special cases, namely, $\Theta_a = \Theta^+$, which is *testing with unspecified alternatives*, and $\Theta_0 \subset \Theta_a \subset \Theta^+$, $d_0 < d_a < d$, *nested testing* (§5.1). Moreover, it is always the case in our testing that $d_0 > 0$, hence, prior to evaluating test statistics, one needs to estimate parameters.

Let $X_1, \ldots, X_N$ be a random sample from $p \in \Theta^+$. An important ingredient of the power-divergence tests is the family of the (generalized) power-divergence estimators parametrized by $\lambda \in \mathbb{R}$. Provided a solution to the minimization below exists and is unique[2], the estimators are defined as follows:

$$\hat{p}(\lambda) \overset{\text{def}}{=} \arg\min_{q \in \Theta_0} I(\hat{p}, q; \lambda), \tag{B.2}$$

[2] In most practical situations the uniqueness follows from the convexity of $I(\hat{p}, \cdot; \lambda)$.

where $\hat{p} = (\ldots, p_k, \ldots)$ is the empirical distribution of the sample, i.e, $\hat{p}_k = \frac{n_k}{N} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}_{\{X_n=k\}}$. Similarly, these can be rewritten by transferring the minimization to the parameter space as follows: $\hat{p}(\lambda) = \arg\min_{\gamma\in\Gamma} I(\hat{p}, f(\gamma); \lambda)$. Note also that $\hat{p}$ is also the (unconstrained) *maximum likelihood estimator* (MLE) of p:

$$\hat{p} = \arg\max_{q\in\Theta^+} \prod_k q_k^{n_k} = \arg\max_{q\in\Theta^+} \sum_k n_k \log q_k.$$

In the case of unspecified alternatives, power-divergence statistics are of the form:

$$T(\hat{p}, \hat{p}'; \lambda) \quad = \quad 2NI(\hat{p}, \hat{p}'; \lambda),\ \ \lambda \in \mathbb{R},$$

for some estimator $\hat{p}'$ of p. The condition that

$$T(\lambda) \underset{N\to\infty}{\Rightarrow} \chi^2(d - d_0) \tag{B.3}$$

makes these statistics valuable in practice by effectively defining the class of (constrained) estimators $\hat{p}'$ to be used with corresponding tests. A large and well-known class of estimators called *best asymptotically normal* (BAN) satisfy this condition [54]. Moreover, under very general (Birch's) regularity conditions on Θ_0 (more precisely, on its parametrization (Γ, f)), the minimum power-divergence estimators are BAN [54]. Among such regularity conditions is the *positivity* of $f(\gamma^*)$ for any $\gamma^* \in \Gamma$ actually observable under H_0. Restricting our modeling to Θ^+ (§4.1), which is enforced by our prevention of zero counts, (§§3.3,3.4) guarantees that this condition is always satisfied in the model testing experiments. Other important examples of these conditions are the differentiability and non-singularity of f. Our two choices of f (i.e., the exponential and linear maps above) satisfy all these conditions rather trivially.

That the estimators $\hat{p}(\lambda)$'s are BAN implies, in particular, that $\hat{p}(\lambda_1)$ with any λ_1 may be substituted into $T(\lambda_2)$ with any λ_2 and the asymptotic condition (B.3) still holds. This last property is referred to as *asymptotic equivalence* of the minimum power-divergence estimators.

Nested testing requires a simple modification in which $\hat{p}$ is replaced by $\hat{p}_a$, some appropriately constrained BAN estimator:

$$T(\hat{p}_a, \hat{p}'; \lambda) \;=\; 2NI(\hat{p}_a, \hat{p}'; \lambda) \underset{N\to\infty}{\Rightarrow} \chi^2(d_a - d_0).$$

Correspondingly, in order to use minimum power-divergence estimators (constrained by Θ_a) as $\hat{p}_a$, Birch's regularity conditions must also be imposed on Θ_a. Just as before, the regularity conditions are always satisfied in these cases.

Several values of λ are special in that the statistics, estimators, or divergence measures corresponding to them have long been used and have distinct names. Among such classical statistics are, first of all, the *generalized log-likelihood ratio* ($\lambda = 0$), *Pearson's goodness of fit*, ($\lambda = 1$), their *modified versions* ($\lambda = -1$ and $\lambda = -2$ respectively) and also the *Freeman-Tukey statistic* ($\lambda = -0.5$) [54].

Besides the Kullback-Leibler divergence D that corresponds to $\lambda = -1$ and $\lambda = 0$, another distinguished example of power-divergence measures is $I(p, q; -0.5)$, the *Hellinger Distance*.

Among the most popular minimum power-divergence estimators is the (constrained) MLE. In [54], the authors appeal to common sense and suggest using with $T(\lambda)$ its true minimizer, but "if only the MLE ... is readily available", but not to hesitate using the MLE with other power-divergence statistics. Depending on the constraints, computation of other estimators may be more involved than the MLE, which partly explains why the MLE is commonly used in practice even with tests arising from $\lambda \neq 0$.

Despite the asymptotic equivalence, power-divergence tests may differ significantly due to finite sample sizes. In order to select a particular member of the family in practice, it is important, among other issues, to understand beforehand what kinds of departure from the null hypothesis are most significant from the modeling viewpoint. Thus, for example, according to [54], tests with large negative values of λ help to detect “departures involving ratios of alternative to null expected frequencies that are close to zero in one or two cells”.

If the primary goal is to detect overall lack of fit, “choosing $|\lambda|$ small is advisable” [54]. Also, “comparing the computed values of the power-divergence statistic for different values of λ... can help to assess the extent of departure from the model”. In particular, having a range of test statistics consistently lying on one side relative to the asymptotic significance level increases reliability of the decision based on such tests.

The choice of $\lambda = -1$ is also advocated for nested (hierarchical) testing of models with external constraints as it leads to the additive partitioning of the test statistics: $I(\hat{p}, \hat{p}_m(-1); -1) = I(\hat{p}, \hat{p}_l(-1); -1) + I(\hat{p}_l(-1), \hat{p}_m(-1); -1)$, where the subscripts l and m stand, respectively, for “less” and “more” restrictive models. Under $H_{0,m}$, all three admit the χ^2 asymptotics and the same additivity holds for the respective degrees of freedom: $d - d_m = (d - d_l) + (d_l - d_m)$.

Models arising from internal constraints and parameter estimation by entropy maximization (ICP/MEE) are identified with their log-linear representations in the ECP/MLE framework (4.2.2,5.1). This allows one to apply nested testing to such models. If d_m constraining functions V^m include a subset of $d_l < d_m$ constraining functions V^l, then the corresponding feasible regions satisfy $F_m \subset F_l$. (The normalization constraint is not being counted.) The transformation of the two corresponding ICP/MEE to the ECP framework evidently reverses this latter relation: $\Theta_{0,l}^{loglin} \subset \Theta_{0,m}^{loglin}$. Now, if $\hat{p}(\lambda)_m$ and $\hat{p}(\lambda)_l$ correspond to $\Theta_{0,m}^{loglin}$ and $\Theta_{0,l}^{loglin}$, respectively, then choosing $\lambda = 0$ leads to $I(\hat{p}, \hat{p}_l(0); 0) = I(\hat{p}, \hat{p}_m(0); 0) +$

$I(\hat{p}_m(0), \hat{p}_l(0); 0)$, another information-partitioning property. Under $H_{0,l}$ (i.e., "$p \in \Theta_{0,l}^{loglin}$"), all three statistics have the χ^2 asymptotics with $d - d_l$, $d - d_m$, and $d_m - d_l$, the respective degrees of freedom.

Also, in modeling by internal constraints, the (generalized) minimization of power-divergence leads to a whole class of additive models each for the particular value of λ: $\hat{p}_{ICP}(\lambda) = \arg \min\limits_{q \in \Theta_0(\hat{p})} I(q, u; \lambda)$ ([54], §4.2.2), where u is the uniform distribution. Two central examples are, of course, $\lambda = 0$ and $\lambda = 1$, *log-linear* and *linear* models respectively. Since for each λ there corresponds an uncertainty measure (where Shannon's entropy is the case $\lambda = 0$), an analog of the MEE is conceivable for $\lambda \neq 0$. Although uncertainty measures other than Shannon's entropy are sometimes advocated for specific tasks[3], Shannon's entropy remains the universal choice due to its elegant mathematical characterization that connects *Information* and *Complexity Theories* with *Statistical Physics.*

Some issues of $\beta(q, \lambda) = P(H_0 \text{ is rejected} | q \in \Theta_a)$, the *power* of the power-divergence tests, are discussed in [54]. For example, it is noted that if Θ_a is uniformly bounded away from Θ_0, then for any $q \in \Theta_a$ and for any λ real, $\beta(q, \lambda) \to 1$ as the size of the sample, on which the decision to reject is based, increases. *Pittman Asymptotic Relative Efficiency* and *Bahadur Efficiency* are also discussed among other issues related to the test power. These issues are clearly important from the theoretical point of view, and their implications for our testing situations may need to be studied further, especially if the goal is purely phenomenological aspects of our hypotheses (e.g., microimage symmetries). Thus, for example, for a fixed sample size Monte-Carlo simulations may be performed to estimate the test power for a set of alternative distributions q typically observed in our experiments. In addition

[3] For example, utility of the measure corresponding to $\lambda = -0.5$ was studied in [45]

to non-parametric estimation, non-central chi^2 and normal approximations are also reasonable [54]. If, on the other hand, the goal of modeling is more application-oriented, then it is likely that any test results, however reliable from the testing point of view, will still need revision in the context of the appropriate application.

B.1 Minimum Power-Divergence Estimators for Symmetry Constraints

In the case of symmetry constraints defining Θ_0 (§§4.2.1,5.1,6.1), $\hat{p}(\lambda)$ is immediately available in a closed form:

$$\hat{p}(\lambda)_k = \frac{\left(\frac{\sum\limits_{j\in\mathcal{O}} \hat{p}_j^{\lambda+1}}{|\mathcal{O}|}\right)^{\frac{1}{\lambda+1}}}{\sum\limits_{\mathcal{O}'\in\mathcal{S}} |\mathcal{O}'| \left(\frac{\sum\limits_{j\in\mathcal{O}'} \hat{p}_j^{\lambda+1}}{|\mathcal{O}'|}\right)^{\frac{1}{\lambda+1}}}, \quad \forall k \in \mathcal{O} \in \mathcal{S} \tag{B.4}$$

The case of $\lambda = -1$ must again be understood in the limiting sense, and the corresponding estimator is simply the geometric orbit-average of the empirical distribution $\hat{p}$:

$$\hat{p}(-1)_k = \frac{\left(\prod\limits_{j\in\mathcal{O}} \hat{p}_j\right)^{\frac{1}{|\mathcal{O}|}}}{\sum\limits_{\mathcal{O}'\in\mathcal{S}} |\mathcal{O}'| \left(\prod\limits_{j\in\mathcal{O}'} \hat{p}_j\right)^{\frac{1}{|\mathcal{O}'|}}}, \quad \forall k \in \mathcal{O} \in \mathcal{S} \tag{B.5}$$

For $\lambda < -1$, $\hat{p}$ must be strictly positive (i.e. no zero counts), and $\lambda = -1$ requires that at least one orbit have non-zero counts for all of its members. Note that $\lambda = 0$ indeed reduces (B.4) to the simple (arithmetic) averaging over the orbits: $\hat{p}(0)_k = \hat{p}_{(MLE,\Theta_0)k} = \frac{1}{|\mathcal{O}|} \sum\limits_{j\in\mathcal{O}} \hat{p}_j$, $\forall k \in \mathcal{O}$.

B.2 Testing Two Sample Consistency

Just as generalized likelihood ratio and Pearson's χ^2 statistics can be used to test if two random samples come from the same distribution [3],[37],[54], any other member of the power-divergence family is suitable for the same purpose.

Let $X_1^{(1)}, \dots, X_{N_1}^{(1)}$ and $X_1^{(2)}, \dots, X_{N_2}^{(2)}$ be two random samples from p_1 and p_2, two positive distributions on the same state space Ω, $|\Omega| = K < \infty$. $H_0 : \; p_1 = p_2$, the *two sample consistency* (or, *homogeneity of proportions*) hypothesis can then be tested against the general alternative $H_a : \; p_1 \neq p_2, \; p_1, p_2 \in \Theta^+$.

In order to parametrize the subspace corresponding to H_0, for the moment we introduce a product state space $\bar{\Omega} \overset{\text{def}}{=} \Omega \times \Omega$ equipped with the set of positive probability distributions $\bar{\Theta}^+$. The null and its alternative are then reformulated relative to the extended framework as follows: $\bar{H}_0 : \; \bar{p} = p_1 \otimes p_2 \in \bar{\Theta}_0$ and $\bar{H}_a : \; \bar{p} \in \bar{\Theta}^+$, where

$$\bar{\Theta}_0 \quad \overset{\text{def}}{=} \quad \{q \in \bar{\Theta}^+, \; \sum_{\omega \in \Omega} q(\omega', \omega) = \sum_{\omega \in \Omega} q(\omega, \omega'), \; \forall \omega' \in \Omega\} \tag{B.6}$$

Then the corresponding test statistic is $T(\lambda) = N_1 I(\hat{p}^{(1)}, \hat{p}'; \lambda) + N_2 I(\hat{p}^{(2)}, \hat{p}'; \lambda)$. Here, $\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ are the empirical distributions obtained from the two samples, and the minimum power-divergence estimators can again be used as $\hat{p}'$. Due to linearity of the constraints defining $\bar{\Theta}_0$, Birch's regularity conditions are again satisfied and hence the minimum-power-divergence estimators (B.7),(B.8) are again *best asymptotically normal.*

$$\hat{q}(\lambda)_i = \frac{\left(N_1 (p_i^{(1)})^{\lambda+1} + N_2 (p_i^{(2)})^{\lambda+1}\right)^{\frac{1}{\lambda+1}}}{\sum_{k=1}^{K} \left(N_1 (p_k^{(1)})^{\lambda+1} + N_2 (p_k^{(2)})^{\lambda+1}\right)^{\frac{1}{\lambda+1}}}, \quad i = 1, \dots, K, \quad \lambda \neq -1 \tag{B.7}$$

$$\hat{q}(-1)_i = \frac{(p_i^{(1)})^{\frac{N_1}{N_1+N_2}} (p_i^{(2)})^{\frac{N_2}{N_1+N_2}}}{\sum_{k=1}^{K} (p_k^{(1)})^{\frac{N_1}{N_1+N_2}} (p_k^{(2)})^{\frac{N_2}{N_1+N_2}}}, \quad i = 1, \dots, K, \tag{B.8}$$

Consequently, with any such estimator $T(\lambda) \underset{\min\{N_1,N_2\}\to\infty}{\Rightarrow} \chi^2(K-1)$. It can be shown that the K homogeneous equations in the right side of (B.6) contain a subset of $K-1$ independent equations, which are also independent of the normalization constraint $\sum\limits_{\omega,\omega'\in\Omega} q(\omega,\omega') = 1$. Therefore, $\dim\bar{\Theta}_0 = K^2 - K$ and there are indeed $K-1$ degrees of freedom.

APPENDIX C

MICROIMAGE SAMPLING AND SUPPLEMENTARY RESULTS

C.1 Microimage Sampling and Alternate Distribution

Since we allow the natural image population Σ_{im} to contain images of variable sizes (§3.1), our definition of the microimage distribution p (3.2) has a sensible alternative (C.1):

$$p^{alt}(\omega) \overset{\text{def}}{=} \tfrac{1}{|\Sigma_{im}|} \sum_{I \in \Sigma_{im}} \tfrac{n(\omega,I)}{S(I)} = \mathbb{E}_{im} \frac{n(\omega, I)}{S(I)} \tag{C.1}$$

(It is straightforward to verify that $p^{alt}(\omega)$ is indeed a probability mass function on Ω.) One may interpret $\frac{n(\omega,I)}{S(I)}$ as defining the conditional probability of ω given image I, $p^{alt}(\omega|I)$, so that $p^{alt}(\omega) = \sum_{I \in \Omega_{im}} p^{alt}(\omega|I)\mathbb{P}_{im}(I)$. On the other hand, p does not lead to a meaningful conditional measure. Practically, the difference between the two definitions is simple: Sampling according to p amounts to first mixing all patches from all images from Σ_{im}, and then selecting one patch at random from Σ, the resulting microimage population. Sampling under p_{alt}, on the other hand, consists of first sampling a random image from Σ_{im}, and then sampling a random patch from that image. Obviously, the difference would be nonexistent if all the images in Σ_{im} were of the same size.

We will shortly return to the discussion of choosing between p and p^{alt}, and for now only say that our experience from this and other works indicates little difference between the two in terms of their key properties (such as symmetry relations).

We now seek an answer to the following basic question: *How can we generate a sufficiently large and random sample from p?* One reason why we want such a sample is to base our statistical inference on the theory of *Power-Divergence Tests* (see [54], Appendix B, Chapters 3,5).

First, we assume that we obtained our $N = 400$ images through random sampling (without replacement) from the population Σ_{im}[1]. Presently, considerably larger image samples are available and the decision to limit $\mathcal{I}_{im}$ by 400 images was somewhat arbitrary. We anticipate that with $N = 1000$, most of the technical difficulties faced in this work (e.g., aggregation of rare states (§3.3)) become obsolete. However, prior to the experimentation, little knowledge was available about how large a random (micro)image sample would be needed to avoid these difficulties altogether. On the other hand, while coping with these technicalities some interesting properties of natural images were observed (see §C.2 below).

Note that we know neither $|\Sigma_{im}|$, the size of the image population, nor $|\Sigma|$, the size of the microimage population. This makes our situation somewhat more complicated than that of *cluster sampling* [15],[53]. Recall, that, if at least the average number of microimages per image, $\mathbb{E}_{im}S = \frac{|\Sigma|}{|\Sigma_{im}|}$, were known, the *single stage cluster-sampling estimator* $\frac{\sum_{i=1}^{N} n(\omega, I_i)}{N\mathbb{E}_{im}S}$, which is unbiased for $p(\omega)$, would be available. Of course, replacing $\mathbb{E}_{im}S$ with its sample estimate $N^{-1}\sum_{i=1}^{N} S(I_i)$ produces

[1] Recall that when the sample size is very small relative to the population size, which will almost always be the case in our experiments, the effect of non-replacement in the sampling is negligible.

the consistent (asymptotically unbiased) estimator for $p(\omega)$ defined in (C.2).

$$\hat{p}_N(\omega) \overset{\text{def}}{=} \frac{\sum_{i=1}^{N} n(\omega, I_i)}{\sum_{i=1}^{N} S(I_i)}, \tag{C.2}$$

The consistency of $\hat{p}_N$ is a simple consequence of the *Law of Large Numbers.* Also note that, due to variable image sizes, the relative frequencies (C.3) obtained by extracting a single random microimage $\omega(I)$ from each of the N images do not in general lead to a consistent estimator for p. They do, though, estimate $p^{alt}(\omega)$ with no bias[2]:

$$\hat{p}^{alt}(\omega) \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{\omega(I_i)=\omega\}} \xrightarrow[N\to\infty]{\text{a.s.}} p^{alt}(\omega). \tag{C.3}$$

Unfortunately, although $\hat{p}_N$ is a sample mean for $\mathcal{I}$ (the set of all the patches extracted from every image of $\mathcal{I}_{im}$), we cannot treat $\mathcal{I}$ as a random sample from p. That is to say, the $\sum_{i=1}^{N} S(I_i)$ patch indicators are not *i.i.d.* random variables. The reason is, of course, high dependence among microimages within an image: In order to realize the extent of this dependence note, for example, that any exhaustive collection of non-overlapping microimages already determines its image completely. Again, this occurs because we are not sampling directly from the population of interest, namely Σ (natural microimages), but instead from the population Σ_{im} of natural images. (We consider this “administrative” inconvenience as rather minor when compared to similar problems in social sampling.) Thus, since $N = 400$ is certainly “small” relative to both current web resources and average image size $\mathbb{E}_{im}S$, we expect $\hat{p}_N$ to be significantly biased towards our image sample. In the next section (§C.2) we present evidence of long-range dependence in natural images that further explains our concern about bias in $\hat{p}_N$.

[2] This is again due to the law of large numbers and the fact that under this sampling, the probability of $\{\omega(I_i) = \omega\} = p^{alt}(\omega)$ (C.1).

In the case of p^{alt}, the obvious modification of $\hat{p}_N$ would be $\hat{p}^{alt}_N(\omega)= \frac{1}{N}\sum_{i=1}^{N}\frac{n(\omega,I_i)}{S(I_i)}$. This estimator is already unbiased. We will carry out some experiments (see §C.4) to test microimage symmetries under p^{alt}, and, in particular, using $\hat{p}^{alt}_N$. *However, our principal measure on* Ω *is* p. At this point, we mention only half of the reason why we prefer p to p^{alt}; and after we propose a "satisfactory" estimator for p, we will complete this argument. Recall that, regardless of the microimage measure, we need to estimate a 255-dimensional vector in the case of 2×2 patches (with $L = 4$, $|\Omega| = 256$). We have assumed that $\mathcal{I}_{im}$ is a random sample from Σ_{im}, thus subsampling one patch per image in order to obtain $\hat{p}^{alt}$ produces a random microimage sample from p_{alt}. This microimage sample is as large as the image sample, namely $N = 400$, and consequently accuracy of estimating 255 parameters in this case is doubtful. Using $\hat{p}^{alt}_N$ instead, which is also unbiased for p^{alt}, does not resolve the issue either: Relative to the number of parameters, $N = 400$ is just too small to rely on large sample results needed for hypothesis testing. Also, $\hat{p}^{alt}_N$ is a random vector on Ω_{im}, and in particular it is not constructed as mean of a random microimage sample. For this reason, using $\hat{p}^{alt}_N$ in order to test hypotheses about p^{alt}, one would need a framework (see §C.4) different from the power-divergence tests.

We now explain how we generate a sufficiently large but (approximately) random sample from p. It will then also become clear why the same approach is not desirable in the case of p^{alt}.

Since problems with $\hat{p}_N$ stem from a potentially strong bias relative to p, we attempt to reduce the bias by randomizing $\hat{p}_{400}$. The approach is intuitive and simple: Extract at random a relatively small fraction d (the "sampling rate") of all the microimages from each of the N images of our sample. (Except for the unavailability of $\mathbb{E}_{im}S$, this is *two stage cluster sampling* [10],[53].) With d small enough the independence assumption should be valid due to large spatial distances between individual microimages, and we can still generate "large" samples. Let M be the total number of microimages of the given size in our image sample, i.e., $M = |\mathcal{I}| = \sum_{i=1}^{N} S(I_i)$. Then the effective microimage sample size $N_\omega$is approximately dM, and we denote the microimage sample mean estimator of p by $\hat{p}_d$ (C.4):

$$\hat{p}_d(\omega) \overset{\text{def}}{=} \frac{1}{N_\omega} \sum_{i=1}^{N_\omega} \mathbb{I}_{\{\omega_i = \omega\}} \tag{C.4}$$

Relative to the distribution $\hat{p}_N$ (assumed fixed for the moment, i.e., replacing Σ by $\mathcal{I}$), this way of subsampling microimages is essentially *simple stratified random sampling* [10],[15],[55],[62]. (The strata are, of course, the individual images in $\mathcal{I}_{im}$.) Note that instead of stratified, one could also use simple (sub)sampling, i.e., just sampling dK microimages with (or without) replacement from $\mathcal{I}$. $\mathcal{I}_{im}$ being fixed, both schemes are unbiased for $\hat{p}_N$[3]. The difference between these schemes (see [10]) is unimportant in our situation of estimating p as compared, for example, with the bias in $\hat{p}_N$ relative to p (and the computational costs of the two schemes are almost the same).

Notice also that we can think of the proposed subsampling as effectively breaking down original images into subimages (still large enough to have a global semantic interpretation) of almost equal size and sampling one microimage from every subimage. It might even be reasonable to assume that the new collection of (sub)images is still a random sample from $\mathcal{I}_{im}$.

Now imagine carrying the idea of sparse microimage subsampling to p^{alt}. Instead of sampling with the fixed rate d (i.e., with probability proportional to the image size), one would need to extract an equal number of microimages from each image

[3] Strictly speaking, stratification may have a bias due to rounding of $dS(I)$, but such bias is obviously negligible.

to obtain $\hat{p}_d^{alt}$, defined by analogy with $\hat{p}_d$ (C.4) as microimage sample mean. The previous argument of dividing images into several subimages suggests a potential difficulty in controlling microimage sample independence: For independence to be valid in all images of $\mathcal{I}_{im}$, the number of patches extracted from each image would need to be determined by the size of the smallest image(s) of $\mathcal{I}_{im}$. This would result in under-sampling relative to the case of $\hat{p}_d$, where the number of extracted microimages per image is proportional to the image size.

On the other hand, some of our earlier experiments [22] showed that coarse properties (e.g. symmetry relations) of microimage statistics are not sensitive to whether a constant fraction or a constant number of microimages are subsampled from variable-sized images as long as the subsampling is very sparse. Thus, we find the difference between p and p^{alt} to be too abstract and artificial for practical purposes, and expect it to prove insignificant when tested on sufficiently large samples.

Finally, then, the estimator of choice is $\hat{p}_d$, obtained by randomly subsampling $dS(I)$ (more precisely, $\lfloor dS(I) \rfloor$) patches from each image $I \in \mathcal{I}_{im}$. *Summarizing the above considerations, we have argued that $\hat{p}_d$, a randomized version of $\hat{p}_N$, is a more satisfactory estimator of p than $\hat{p}_N$ due to ameliorating the bias in sampling microimages from Σ.* We also conjecture that the bias in $\mathcal{I}_{im}$ will largely disappear with $N \approx 10,000$, namely the scale of $S(I)$ (as estimated from $\mathcal{I}_{im}$). Should this be true, and in the near future we will have the resources to check it, the whole issue of adjusting $\hat{p}_N$ will become obsolete.

C.2 Sampling Rate

We now turn to the question of selecting "the right" value for d, the microimage sampling rate, the first step of which is a demonstration of the long-range dependence in natural images. Let $\rho(\omega_1, \omega_2)$ count the number of identical pixels in ω_1 and ω_2. Considering (sub)populations of all patch pairs from $\mathcal{I}$ that are r pixels apart in their respective images makes the random variable $\rho(\omega_1, \omega_2)$ a function of r. We then estimate probabilities of the events $\rho(\omega_1, \omega_2; r) = n$, where $n = 0, 1, \ldots,$ the number of pixels in the patch. The most illustrative observations correspond to the probabilities that two patches in the pair are identical; the relevant estimates as well as their linear fits are presented in Figure 8. The "independence" benchmarks are obtained when patches in the pair are sampled from different images.

This phenomenon suggests that the bias in $\hat{p}_N$ does not disappear abruptly when the sampling rate falls below some critical level, but instead decreases with d gradually. The dependence introduced by sampling more than one patch per image is likely to behave similarly. According to the graphs of Figure 8, we would roughly have to sample one 2×2 patch from a 200×200 image in order to secure full independence and hence produce a truly random microimage sample. Of course, such sample would be too small (comparable to $N = 400$). Naturally then, we want to assess the within sample dependence effects on our testing schemes. Namely, we would like to know how large d could be so that the non-randomness effect is negligible for the purpose of our testing.

We begin with the following crude reasoning: For simplicity, assume that patches are subsampled at points of a regular grid. There are $9,801$ 2×2 patches in a 100×100 image. Extracting, say, four such microimages (i.e. $d \approx 0.0004$) would correspond to a 2×2 grid, and two extracted patches would be more than 30 pixels

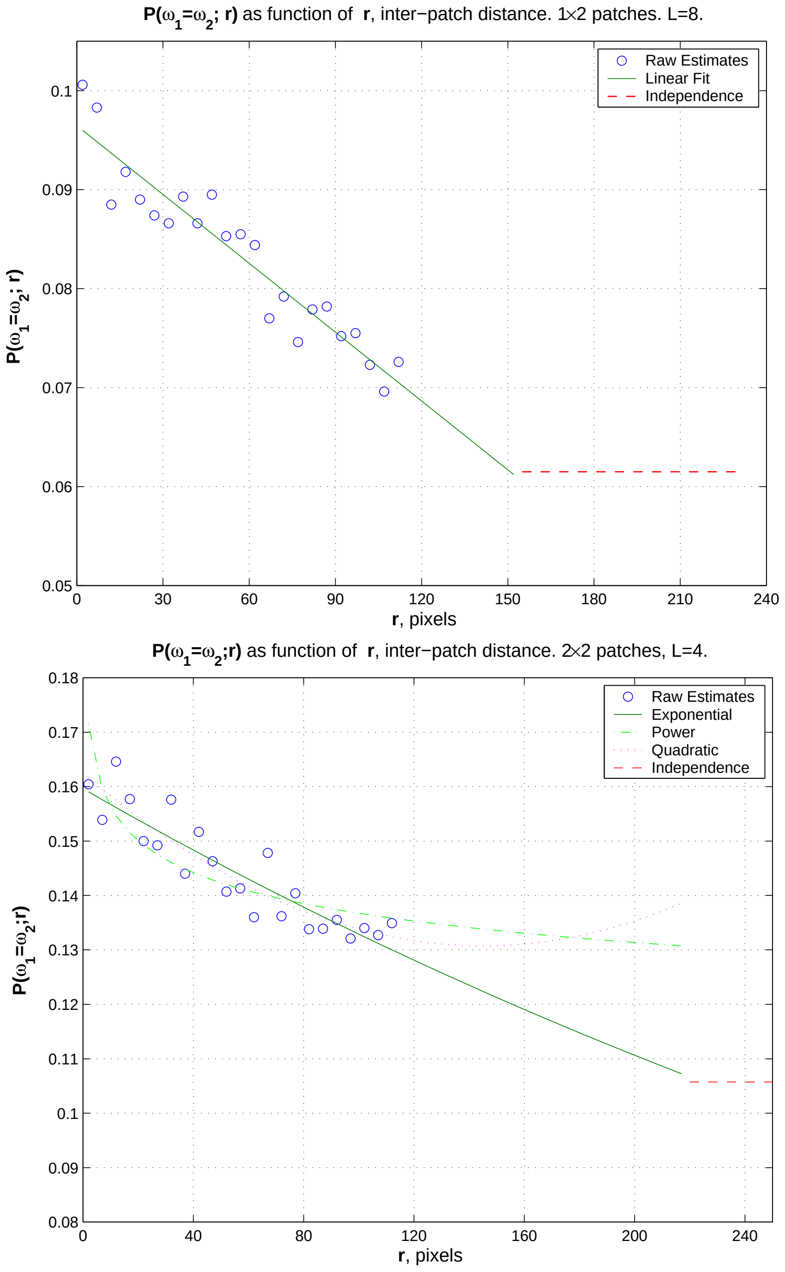


Figure 8: Long-range dependence in natural images. Top: Horizontal pairs. Bottom: The 2×2 case

apart on average. Recalling the argument (§C.1) about virtual subdivision of large images, we can regard the four patches as arising from four 50×50 subimages (one patch per image) which, we would like to assume, represent a random sample from $\mathbb{P}_{im}$.

We now offer an experimental argument in support of the assertion that d of the order of magnitude of 10^{-4} still allows us to generate sufficiently random microimage samples. Naturally, the argument is in the context of hypothesis testing with the power-divergence tests.

Divide the image sample $\mathcal{I}$ *at random* into $\mathcal{I}^1_{im}$ and $\mathcal{I}^2_{im}$ and generate two microimage samples one from each of the new image subsamples. Similar to the set-up of the stability hypothesis testing (§3.5), we formulate a two sample consistency hypothesis "$\hat{p}^{(1)}_d$ and $\hat{p}^{(2)}_d$ estimate the same distribution", where $\hat{p}^{(1)}_d$ and $\hat{p}^{(2)}_d$ are the empirical distributions of the two microimage samples. If the microimage sample sizes are sufficiently large and the non-randomness effects are weak, the distribution of the power divergence test statistic $T(\lambda)$ should resemble the χ^2 distribution with the appropriate degrees of freedom, which is theoretically predicted under the null hypothesis (§B.2). In short, supposing the null holds, we vary d and rerun the tests several times (for each value of d). Next, for each of these values of d we plot the histogram of the test statistic. We then inspect the histograms to obtain a decent match with the asymptote. Note that the number of test runs cannot be arbitrarily large (our experiments use 100) since we implicitly assume independent test statistics. Also, in order to assess validity of using the asymptotic results, we appeal to the property of asymptotic equivalence of the power divergence tests (see Appendix B): Test statistics $T(\lambda)$ evaluated on a common microimage sample and yielding nearly the same values independently of the test parameter λ (e.g., Table 8) indicate that their distributions are indeed close the theoretical asymptote [54].

The case of 1×2 patches is considered first, just like in §3.6. ($L = 8$ and Ω has 64 states.) Using $d \approx 0.0004$ generates about 14,000 patches for each of the two microimage samples. A typical test run with different values of the test parameter λ is recorded in Table 15. Relative to the asymptotic cut-off point $\chi^2_{0.95,63} = 82.53$, the test statistics $T(\lambda)$ are indeed close to each other, thus indicating validity of the asymptotic. The lowest bin count recorded in these experiments is six, whereas five is commonly thought to be sufficient [54].

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	64.47	65.11	65.06	64.76	64.22	63.99	63.5	61.68
$P\text{-}val(\lambda)$	0.4251	0.403	0.4048	0.4152	0.4335	0.4414	0.4587	0.5236

Table 15: Power-divergence tests for two sample consistency. Horizontal pairs.

We also collected one hundred observations of the test statistic for each of the eight values of λ used in Table 15 and compared the empirical distributions with that of the expected $\chi^2(63)$. Whereas in all the eight cases, in more than 60 out 100 runs the null was not rejected, the test statistic histograms only vaguely resembled the (theoretical) $\chi^2(63)$ curve. We attribute this to dependence among the hundred observations: 100 runs of the test necessitates a 100-fold increase in the sampling rate d.

With $d \approx 0.0001$, on the other hand, we run into the empty bin problem. Despite this difficulty, the top graph of Figure 9 indicates a reasonable match with $d \approx 0.0001$ between the $\chi^2(63)$ curve and the test histogram with $\lambda = 0.67$ (strongly advocated in [54] as "golden mean"). Nonetheless, we decided to lump the rare states (counts consistently below five) together: the aggregate set D is exactly the same as in §3.6. The empirical histograms now better match the appropriate theoretical one ($\chi^2(44)$) as can be seen from Figure 10 (top and middle). Similar results were obtained for the 2×1 case (Figure 10, bottom).

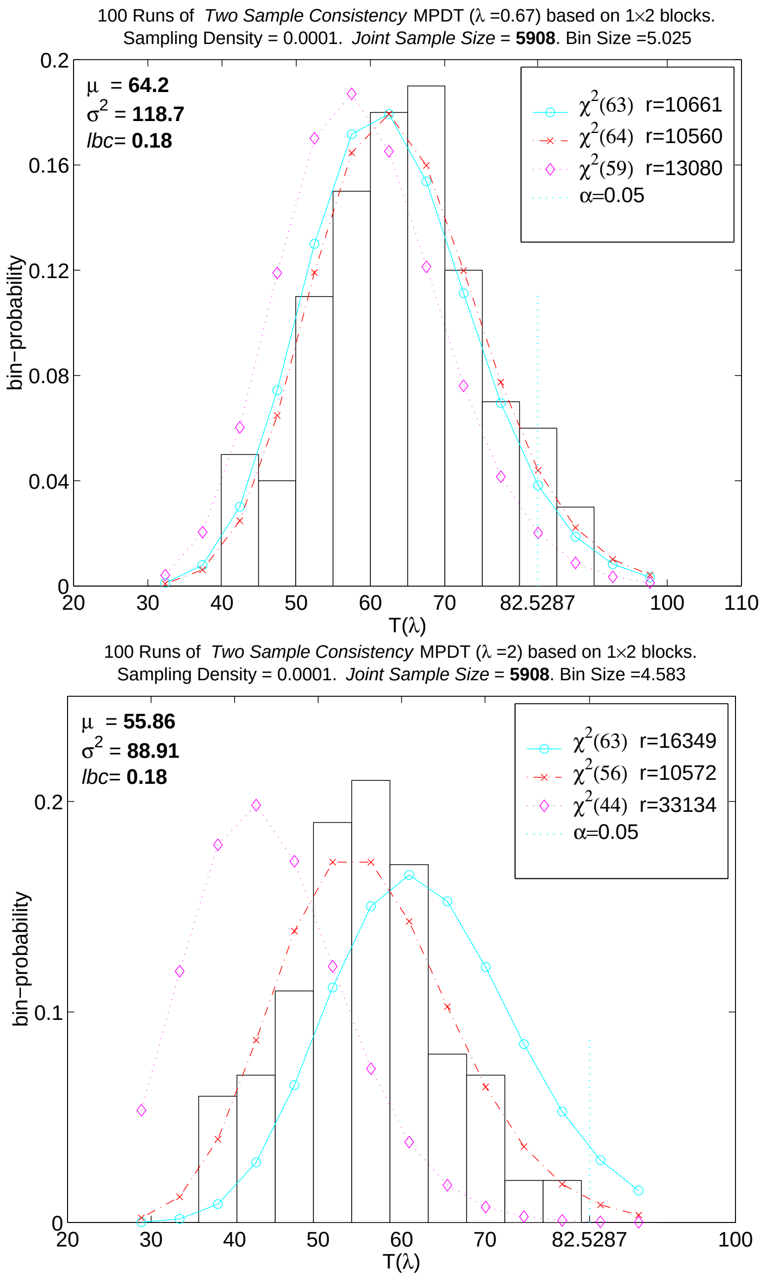
In these and ensuing graphs, the index r in the legend is a crude indicator of the error in the sense of (Pearson's χ^2) goodness of fit, wherein χ^2- distributions are fit to the 100 observations of the test statistic. In addition to the asymptotic curve, we display two more χ^2 curves corresponding to estimation of the degrees of freedom using the sample mean and using one-half the sample variance (both rounded to nearest integers). The two estimated curves usually give tight bounds on the location of the actual best fit.

The average minimum count is reported as "lbc" in the upper left corner under the sample mean and variance. The top caption reports the value of λ along with the sampling rate d and microimage sample size. Observations of the test statistics are grouped optimally in ten bins whose size is also reported in the top caption.

In the 2×2 case ($L = 4$), we aggregate the same 136 rare states as in §3.6.1. Again, based on the asymptotic cut-off point ($\chi^2_{0.95,120} = 146.57$) the null stands firm, although it is obviously harder to validate this conclusion by test repetition. The test results are presented in Figure 11.

Under the G-symmetric aggregation (§§3.4,3.6.1) $d = 0.0001$ allows sufficient flexibility for repeating the tests one hundred times and obtaining a good match with the appropriate asymptotic $\chi^2(25)$ distribution (Figure 12).

In summary, we elicit 10^{-4} as an upper bound on d under which our microimage samples are sufficiently random.



100 Runs of *Two Sample Consistency* MPDT (λ =0.67) based on 1×2 blocks. Sampling Density = 0.0001. *Joint Sample Size* = **5908**. Bin Size =5.025

100 Runs of *Two Sample Consistency* MPDT (λ =2) based on 1×2 blocks. Sampling Density = 0.0001. *Joint Sample Size* = **5908**. Bin Size =4.583

Figure 9: Low count effect in validation of two sample consistency tests. Top: $\lambda = 0.67$. Bottom: $\lambda = 2$

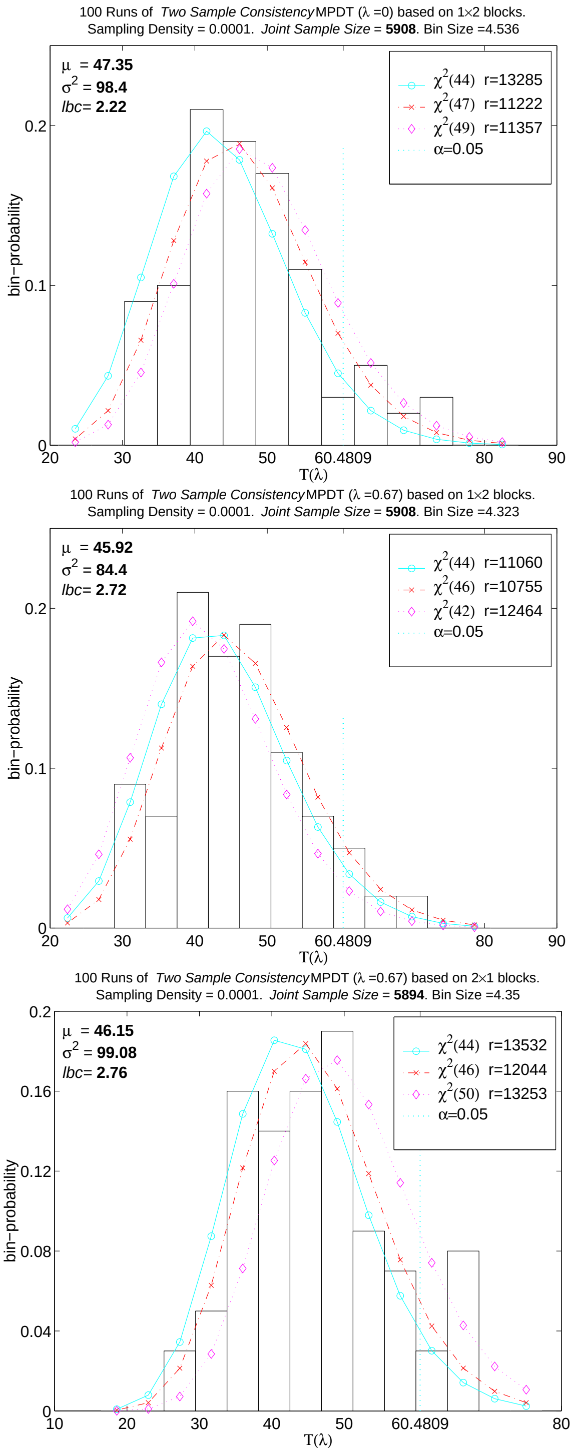
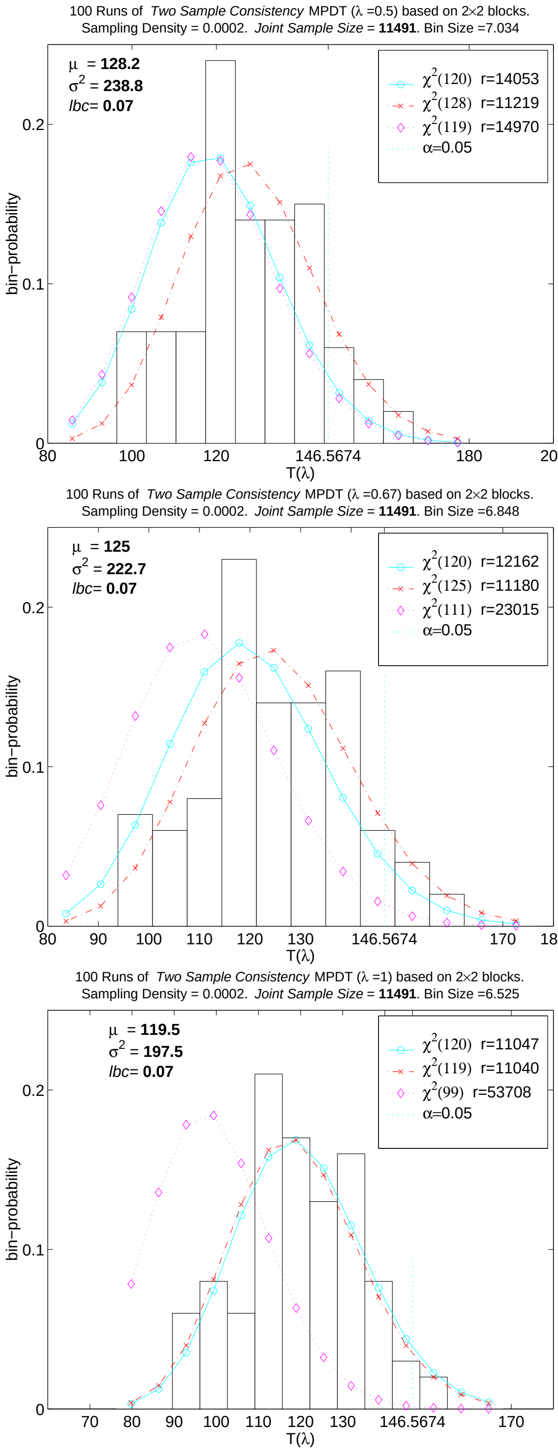


Figure 10: More satisfactory validation of two sample consistency tests after aggregation. Top and middle: Horizontal pairs with $\lambda = 0$ and $\lambda = 0.67$ respectively. Bottom: Vertical pairs, $\lambda = 0.67$

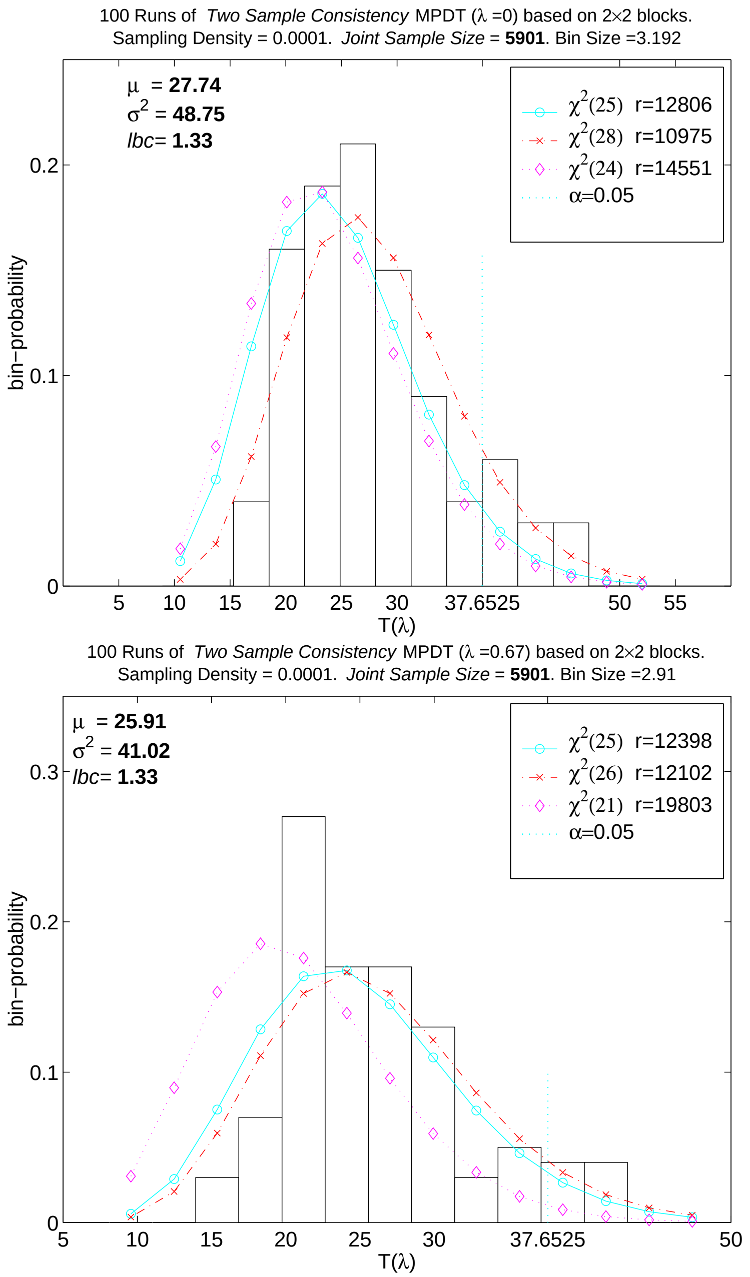


100 Runs of *Two Sample Consistency* MPDT (λ =0.5) based on 2×2 blocks.
Sampling Density = 0.0002. *Joint Sample Size* = **11491**. Bin Size =7.034

100 Runs of *Two Sample Consistency* MPDT (λ =0.67) based on 2×2 blocks.
Sampling Density = 0.0002. *Joint Sample Size* = **11491**. Bin Size =6.848

100 Runs of *Two Sample Consistency* MPDT (λ =1) based on 2×2 blocks.
Sampling Density = 0.0002. *Joint Sample Size* = **11491**. Bin Size =6.525

Figure 11: Validation of two sample consistency tests for 2×2 patches with $L = 4$. Top: $\lambda = 0.5$. Middle: $\lambda = 0.67$, and bottom: $\lambda = 2$



100 Runs of *Two Sample Consistency* MPDT (λ =0) based on 2×2 blocks. Sampling Density = 0.0001. *Joint Sample Size* = **5901**. Bin Size =3.192

100 Runs of *Two Sample Consistency* MPDT (λ =0.67) based on 2×2 blocks. Sampling Density = 0.0001. *Joint Sample Size* = **5901**. Bin Size =2.91

Figure 12: Validation of two sample consistency tests on the G-symmetric classes. Top: $\lambda = 0$. Bottom: $\lambda = 0.67$

C.3 Validity of Testing of Symmetry Hypotheses

We choose the case of the bivariate reflection symmetries (§5.2) for our final discussion of the sampling rate and the test validity. Two indications of the validity of the results in §5.2 are as follows. First, the lowest count in these experiments is two, which is generally acceptable with the most common power-divergence tests covered by our range of λ. In particular, the divergence test for $\lambda = \frac{2}{3}$ is advocated for its relative insensitivity to low bin counts ([54]). Second, we observe very consistent statistics from seven distinct power-divergence tests (Table 8). Specifically, the values of $T(\lambda)$ for $\lambda = -2, -1, 0, 0.5, 0.67, 1, 2$ always lie much closer to each other than to the asymptotic threshold $\chi^2_{0.95,28}$. The following discussion further supports our previous conclusion that $d \approx 10^{-4}$ is satisfactory.

Note that testing of the left-right reflection symmetry on horizontally adjacent pairs on 1×2 patches helps us to calibrate testing of hypotheses on other patch spaces. The reason is that the left-right symmetry is a priori the most compelling. Thus supposing this hypothesis holds, we then repeat tests in order to compare a simulated test statistic histogram with the theoretically predicted $\chi^2(28)$ curve. Recall (beginning of §C.2) that small values of d are desired to enforce sample independence. However, they also lead to small sample sizes and, consequently, to questionable large sample approximations. Correspondingly, large values of d could eventually introduce strong dependence in the sample, also undermining the use of the asymptotics.

An extreme case ($d = 0.25$) with high intra-sample dependence is presented in the top graph of Figure 13. This value of d leads to microimage overlaps being very likely. A serious departure from the $\chi^2(28)$ curve is apparent. In fact, when running the tests multiple times, the statistics $T(0)_i$ ($i = 1, \ldots, 100$) are also likely

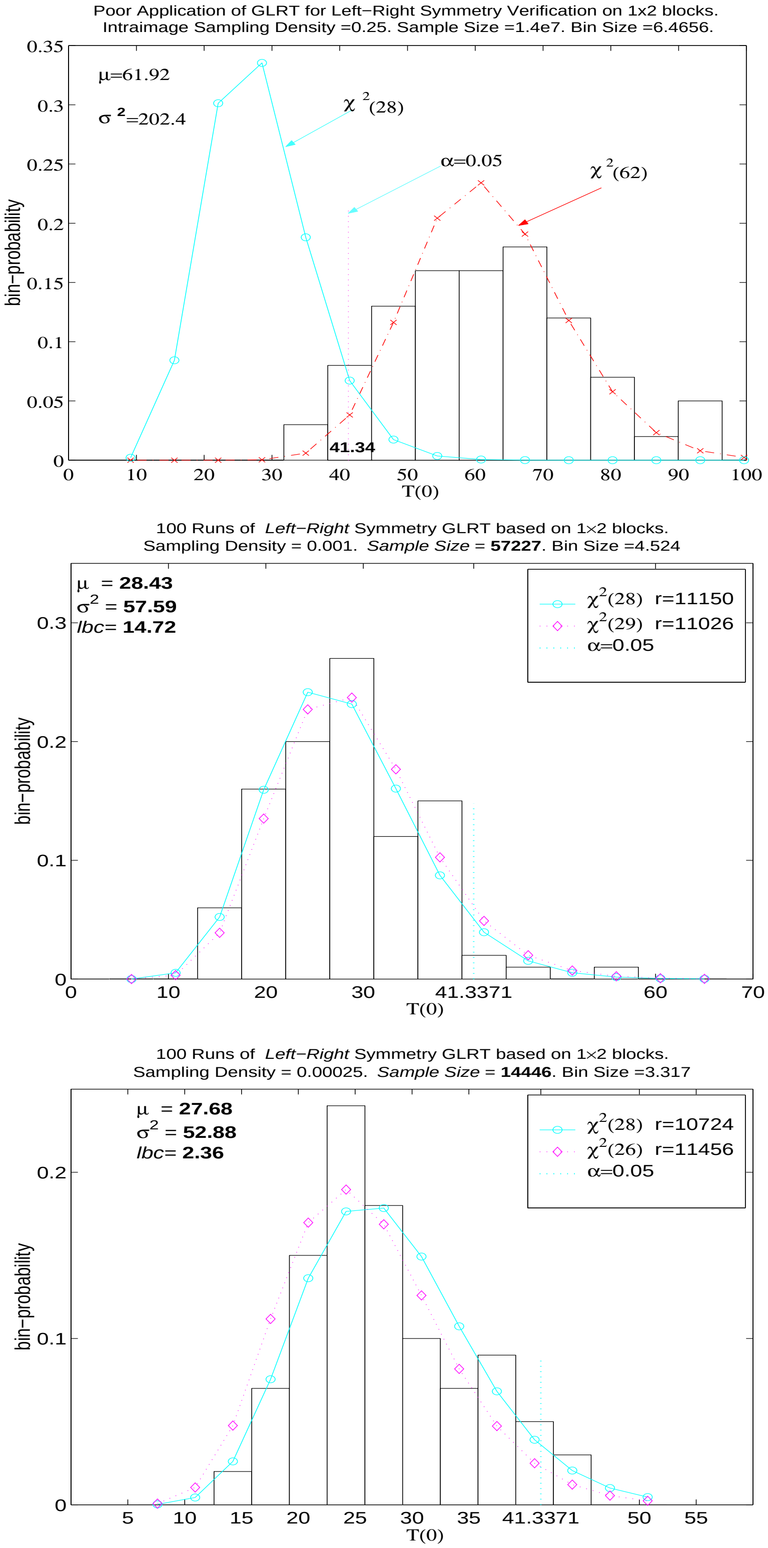


Figure 13: Significance of the independence assumption in GLRT. The asymptotic threshold $\chi^2_{0.95,28}$ is marked at ≈ 41.337

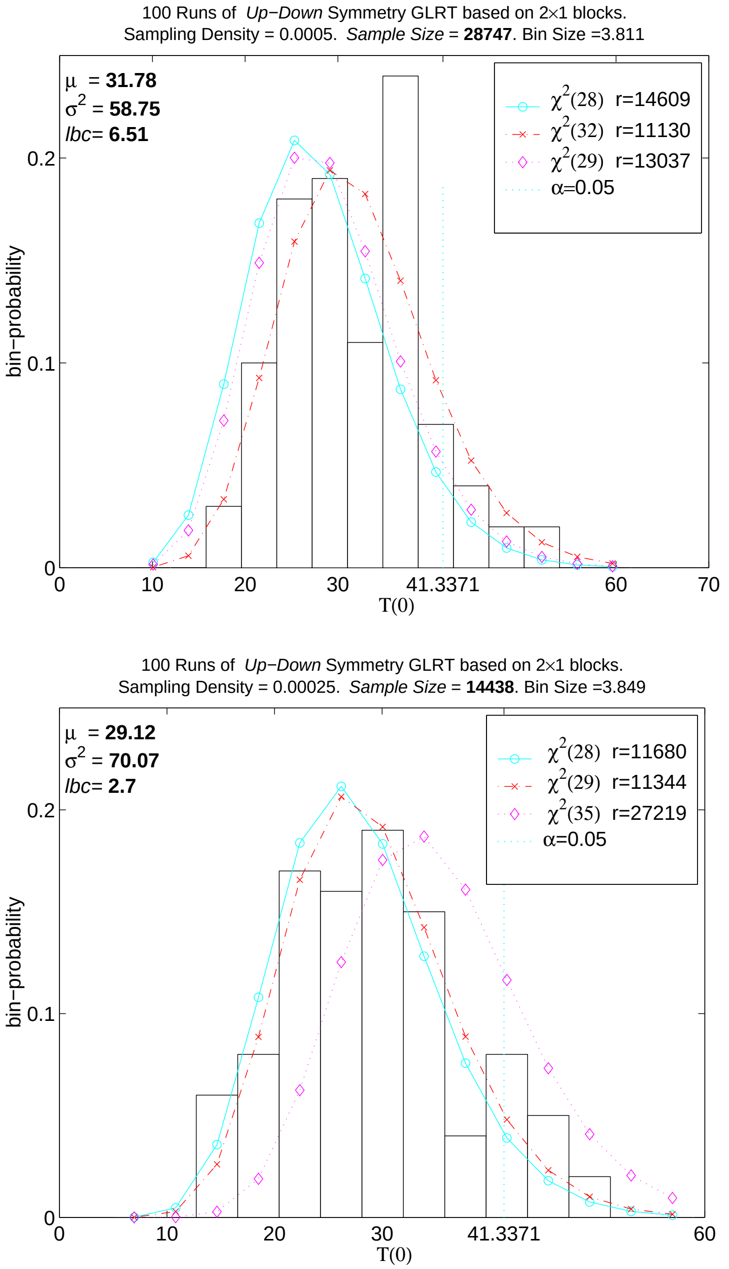
to be dependent. Consequently, the sample histogram may poorly estimate the true distribution of $T(0)$. *Thus we were seeking a range of values for d that would allow sufficiently large but random microimage samples as well as several (we use 100) independent test runs.* Gradually decreasing d (e.g. $d = 0.001$, as in the middle graph of Figure 13), we arrive at $d = 0.00025$ (Figure 13, bottom). Decreasing d further led to zero counts and considerable departures from the χ^2 shape. For instance, two estimates of the number of degrees of freedom (i.e. 28), namely half the sample variance and the sample mean, then diverged. In these and other tests (i.e. $\lambda \neq 0$) optimality of $d \in (0.0001, 0.0005)$ was also confirmed by minimization of the fit-error index r (which corresponds to the Pearson's goodness of fit between the tests histograms and the asymptote); see the legends in the right of Figure 13.

The graphs in Figure 14 uphold the hypothesis of up-down reflection symmetry on 2×1 blocks. However, the match between the empirical and asymptotic distributions is slightly less convincing than that in the case of left-right symmetry. This can partly be explained by the weaker presence of this symmetry in the microworld as compared with left-right symmetry. In fact, we argued in §4.4 that the latter may also be present at macroscopic scales, clearly an unreasonable assumption in the case of up-down symmetry due to the heavy presence of the sky and other vertical anomalies in natural images.

C.4 Another Interpretation of Symmetry

Recall (§C.1) that we chose the microimage measure p over the alternative definition p_{alt}. In this section we further support this choice by examining the validity of hypothesis testing based on p_{alt}.

In fact, instead of seeking a large random sample in order to work with the



100 Runs of *Up−Down* Symmetry GLRT based on 2×1 blocks.
Sampling Density = 0.0005. *Sample Size* = **28747**. Bin Size =3.811

100 Runs of *Up−Down* Symmetry GLRT based on 2×1 blocks.
Sampling Density = 0.00025. *Sample Size* = **14438**. Bin Size =3.849

Figure 14: Validation of tests for the up-down symmetry on 2×1 **patches. The asymptotic threshold** $\chi^2_{0.95,28}$ **is marked at** ≈ 41.337

empirical distributions corresponding to either p or p_{alt}, we will define a vector-valued random variable Z in order to measure deviations from the exact symmetry for each of the 28 symmetry classes $\{(a,b),(b,a)\}$, $0 \leq a < b \leq 7$. Specifically, we first define a deterministic, multidimensional function z on the image space Ω_{im} as follows: Enumerate all the classes from $1-28$ in some order and suppose class $\{(a,b),(b,a)\}$ has index m. Then

$$z^m(I) \stackrel{\text{def}}{=} n((a_m, b_m), I) - n((b_m, a_m), I). \tag{C.5}$$

Thus the components of z are the 28 differences between the number of pairs of type (a,b) and the number of pairs of type (b,a). This definition leads to a 28-dimensional random variable $Z(I) = (Z^1(I), \ldots, Z^{28}(I))$ on the image probability space $(\Omega_{im}, \mathbb{P}_{im})$.

Consider the null hypothesis $H_0 : \mathbb{E}_{im} Z = \vec{0}$. Note that this formulation of symmetry completely eliminates the need for microimage subsampling by allowing every (non-singular) microimage to contribute to the appropriate component of Z. Thus, the relevant space to sample is Ω_{im}, the image space, instead of Ω, the space of microimages . In particular, the sample size is now invariably $N = 400$. We assume that we are given i.i.d. random vectors $Z_1, \ldots, Z_{400}$ and that $N = 400$ is sufficiently large for the sample average (vector sum) to be normally distributed. This latter assumption could be questioned, especially in the 2×2 case where Z already has 120 dimensions in the case of reflection symmetries. More complex symmetries lead to even higher dimensions of Z, namely $|\Omega| - |\mathcal{S}|$, where $|\mathcal{S}|$ is the appropriate set of symmetry classes $\mathcal{O}$.

Recognizing the possible insufficiency of our sample size, we nonetheless test the two-pixel reflection symmetries as measured by Z, using the asymptotic normality in order to obtain 28 simultaneous confidence intervals:

$$\hat{Z}^m \pm \sqrt{\chi^2_{1-\alpha,28}\frac{\hat{\sigma^2_m}}{N}} \tag{C.6}$$

The above approximation for the level-α confidence region is an extension of Scheffé's method [31]. (For the original Scheffé's method, see, for example, [61] and [62].)

Note that we could similarly test the hypothesis $H_0 : \mathbb{E}_{im}\frac{Z}{S} = \vec{0}$, recalling that $S(I)$ is the number of microimages of a given size (currently pixel pairs) in image I. This is none other than the hypothesis that p_{alt} respects reflection symmetry.

The graphs in Figures 15 and 16 display almost perfect results in the left-right case but the overall rejection of the up-down symmetry hypotheses due to seven of the 28 pairs. In order to understand the source(s) of asymmetry, we attempted to visualize the "abnormal" pairs. However, due to the frequent and spatially dispersed presence of these classes in most of our images, associating them with particular larger structures proved difficult. A subjective opinion is that the rejection of the up-down reflection symmetry is mostly due to the insufficient sample size. However, it is noteworthy that the observed vertical asymmetry has a distinct and stable signature. First, rejecting $(a, a+1)$ is certainly less surprising than rejecting $(a, a+2)$; recall that this minimal gradient under uniform $3-$bit quantization corresponds to 32 grey levels in the original 256-level intensity range. Second, due to blue sky all such pairs have the brighter pixel on top. We also suspect that redefining Z to include only *non-overlapping* patches would mask this effect, even with sample sizes on the order of 1000.

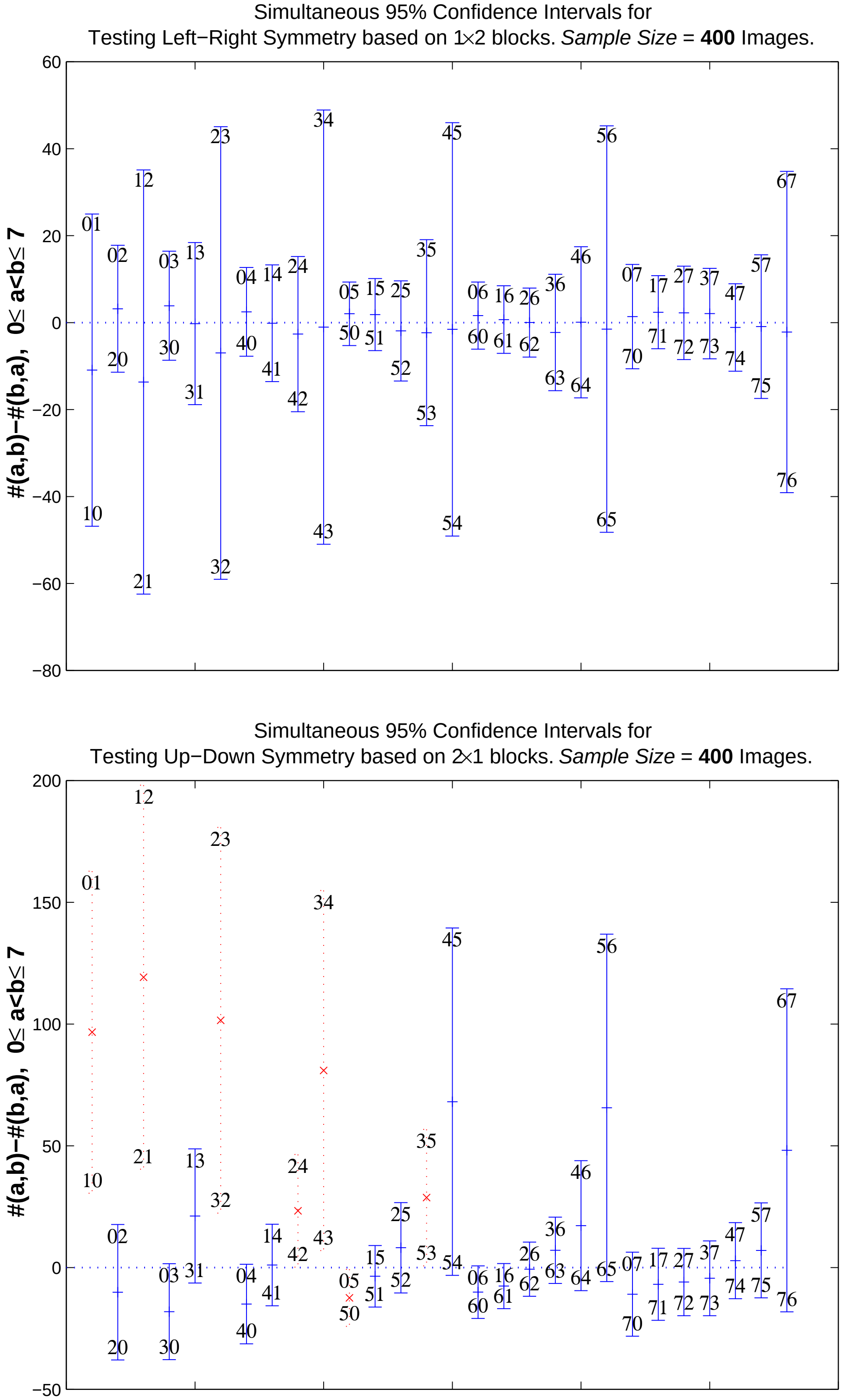


Figure 15: Alternate tests based on Z for left-right (top) and up-down (bottom) reflection symmetries

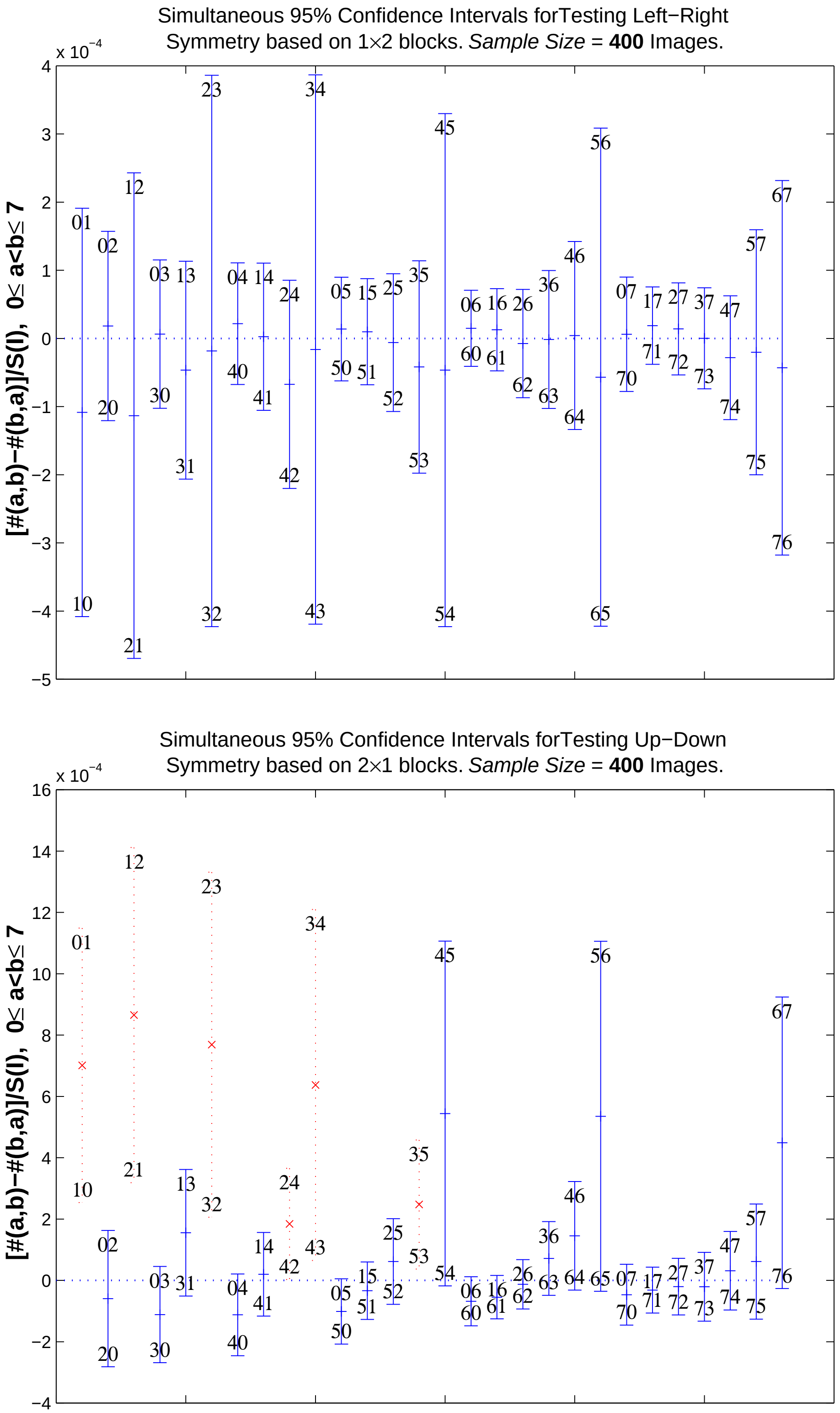


Figure 16: Alternate tests based on Z/S for left-right (top) and up-down (bottom) reflection symmetries

C.5 Other Model Testing Results

Table 16 provides adjusted model complexities corresponding to the aggregation of 184 most rare states in one state (§5.3). Thus, since this aggregation reduces dimension of the unconstrained model space from 256 to 73, the number of degrees of freedom is simply 72− adjusted complexity, the difference between the adjusted number of free parameters of the ground model and the adjusted model complexity.

Models	Original complexity	Adjusted complexity	Degrees of Freedom
p_{dom}	43	43	29
p^{G}_{potts}	2	2	70
p^{vhn}_{potts}	3	3	69
p^{vhnC}_{potts}	8	6	66
p^{g}_{pair}	42	19	53
p^{h}_{symm}	135	43	29
p^{v}_{symm}	135	43	29
p^{vh}_{symm}	75	26	46
p^{n}_{symm}	127	36	36
p^{vhn}_{symm}	43	14	58
p^{g}_{symm}	54	19	53
p^{G}_{symm}	30	10	62

Table 16: Model testing parameters for the 2×2 case.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	921.9	1187.6	1315.8	1372	1351.4	1333.1	1289.6	1151.6
$T(\lambda)$	4801	3448	3162	3073	3196	3298	3619	6785

Table 17: Strong rejection of primitive models. Top: p_{dom}, $\chi^2_{0.99,29} = 49.588$. Bottom: p^{vhn}_{potts}, $\chi^2_{0.99,69} = 99.228$.

Remark C.1 Recall (§5.3) that under present sample size limitations, the difference between p^{g}_{pair} and p^{g}_{symm} is cancelled due to the aggregation. The difference

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	64.55	65.76	65.69	65.12	64.25	63.92	63.25	61.17
$P - val(\lambda)$	0.002	0.002	0.002	0.002	0.003	0.003	0.003	0.006
$T(\lambda)$	91.66	88.76	87.24	86.01	85.01	84.7	84.13	82.48
$P - val(\lambda)$	0.003	0.006	0.008	0.01	0.012	0.013	0.014	0.019

Table 18: Models close to being not rejected. Top: p^{n}_{symm}, $\chi^2_{0.99,36} = 58.619$. **Bottom:** p^{vhn}_{symm}, $\chi^2_{0.99,58} = 85.95$.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	32.24	34.94	34.92	33.84	32.26	31.69	30.6	27.65
$P - val(\lambda)$	0.31	0.207	0.207	0.245	0.309	0.334	0.384	0.537
$T(\lambda)$	19.05	19.82	19.82	19.52	19.05	18.87	18.51	17.47
$P - val(\lambda)$	0.92	0.898	0.899	0.907	0.92	0.925	0.933	0.954

Table 19: Non-rejected models. Top: p^{v}_{symm}, $\chi^2_{0.95,29} = 42.557$. **Bottom:** p^{h}_{symm}, $\chi^2_{0.95,29} = 42.557$.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	41.6	41.43	40.08	38.47	36.86	36.34	35.38	32.85
$P - val(\lambda)$	0.657	0.664	0.717	0.777	0.83	0.845	0.872	0.928
$T(\lambda)$	47.04	44	42.33	40.95	39.84	39.51	38.92	37.33
$P - val(\lambda)$	0.704	0.806	0.853	0.886	0.909	0.916	0.926	0.949

Table 20: Non-rejected models. Top: p^{vh}_{symm}, $\chi^2_{0.9,46} = 58.641$. **Bottom:** p^{g}_{symm}, $\chi^2_{0.9,53} = 66.548$.

between the values in Table 11 and Table 20 (bottom) is only due to using different estimators (Remark 5.2 from §5.3): In the former case we used the constrained MLE (minimum power-divergence estimator with $\lambda = 0$) with all eight tests, whereas in the latter case the genuine minimum power-divergence estimators are used (i.e., the values of the parameter λ in the test statistic $T(\lambda)$ and that used for estimation are the same). Naturally, we attribute the overall close match between the corresponding table entries to the property of asymptotic equivalence of the minimum power divergence estimators (Appendix B).

C.6 More on Stability

In §§3.6,3.6.1,3.7 we analyzed the stability of p on 2×2 patches across scenes and with respect to scale invariance. It was necessary to aggregate patches into either one large, "rare" class, or to do the analysis on the G-invariant classes. In the first case (§3.6.1), we had to collapse at least 136 patches into one to minimize the effect of low bin counts, thus leaving 120 free parameters to estimate. Since even more extreme aggregation would have been necessary in the scaling experiments, we chose alternatively to work with an integrated version of p, namely the one induced on the space of G-invariant classes. We now can complete this discussion by demonstrating a gain in test reliability (as measured, for example, by low dependence on λ) when raw, unconstrained estimates of p are replaced by those constrained by the G-symmetries. To make the comparison fair, we first aggregate the 17 rare classes exactly as we did in the scaling experiments (§3.7), leaving only 14 free parameters to estimate. (Recall (§6.4.1) that there are exactly 31 G-symmetric orbits.) In order to return from the quotient space of G-classes to the original patch space Ω

(except for the "rare" super-state that contains 136 patches and is not affected), the 14 non-collapsed probabilities must be uniformly divided among the class members (§6.1 and Appendix B). Although the composition of the present compound class slightly differs from that of its sibling in §3.6.1 (where we aggregated rare patches intuitively and independently of the G partition), the two have exactly the same size (136) and nearly identical masses (≈ 0.025).

Tables 21 and 22 demonstrate the point: Under the G-invariant model, the testing for cross scenes stability and scale invariance becomes more reliable as the contribution of all the states to the divergence measures (equivalently, the statistics $T(\lambda)$) is more uniform [54]. This can be deduced from similarity in the statistics and P-values for different values of λ. We emphasize that all the figures in each table are based on the same patch sample. Blank entries indicate the presence

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	-	142.78	141.43	129.291	121.188	118.93	115.01	105.24
$P - val(\lambda)$	-	0.077	0.088	0.265	0.453	0.51	0.612	0.829
$T(\lambda)$	10.63	10.7	10.73	10.75	10.76	10.76	10.76	10.75
$P - val(\lambda)$	0.715	0.709	0.707	0.706	0.705	0.705	0.705	0.706

Table 21: Comparison of inter-scene stability testing based on raw empirical (top) and on the G-symmetric model.

λ	-2	-1	-0.5	0	0.5	0.67	1	2
$T(\lambda)$	-	159.95	158.14	132.29	116.48	112.42	105.73	91.41
$P - val(\lambda)$	-	0.009	0.011	0.209	0.574	0.676	0.821	0.976
$T(\lambda)$	22.32	22.31	22.28	22.22	22.13	22.1	22.03	21.78
$P - val(\lambda)$	0.072	0.072	0.073	0.074	0.076	0.077	0.078	0.083

Table 22: Comparison of scale invariance testing based on raw empirical (top) and on the G-symmetric model.

(despite the aggregation) of zero counts, rendering $T(\lambda)$ undefined for $\lambda < -1$ (Appendix B). The uniform redistribution of mass within the G-classes eliminates this problem (the bottom rows) as we collapse into one all the classes consistently showing counts less than five.

APPENDIX D

ALGEBRAIC PRELIMINARIES AND COMPUTATIONS

In this work G is always a finite group; we also write 1 for its identity element and the relation "subgroup of" is denoted by "$\leq$". The following definition of *group action* appears in [18]:

Definition D.1 A *group action* of a group G on a set A is a map from $G \times A$ to A (written as $g \cdot a$, for all $g \in G$ and $a \in A$) satisfying the following properties:

(1) $g_1 \cdot (g_2 \cdot a) = (g_1 g_2) \cdot a$, for all $g_1, g_2 \in G$, $a \in A$, and

(2) $1 \cdot a = a$, for all $a \in A$.

If G acts on A, then each $g \in G$ defines a function σ_g $A \to A$ by $\sigma_g(a) = g \cdot a$. Based on this observation and in order to avoid confusion with the group operation, we will write "$g(a)$" for $g \cdot a$.

In the following definitions, suppose G acts on A.

Definition D.2 Let $\mathcal{O} \subset A$. $\mathcal{O}$ is said to be an *orbit* under the action of G on A if for any $a, b \in \mathcal{O}$, $\exists g \in G$, such that $g(a) = b$.

Also, it is easy to see that the set of orbits partitions A.

Definition D.3 For each $a \in A$, the set $\{g \in G : g(a) = a\}$ is called the stabilizer of a and is denoted by G_a.

In the context of this work, the central example of group action is the action of the group G (§4.4) on a microimage space $\Omega_L = M_{2\times 2}(\mathcal{C}_L)$ (§3.1). In order to simplify certain computations and assuming L is even, we rescaled the intensity range as follows: $\mathcal{C}_L \cong \{-\frac{L-1}{2}, \ldots, -\frac{1}{2}, \frac{1}{2}, \ldots, \frac{L-1}{2}\}$ (§6.4.3). We also observed that this action is the restriction of a "larger" action of G on the vector space $\mathbb{R}^4 \supset \Omega_L$. The embedding of Ω_L in $\mathbb{R}^4$ was carried out by ordering the components of $\omega \in \Omega_L$ as $\{\omega_{2,1}, \omega_{2,2}, \omega_{1,2}, \omega_{1,1}\}$. Ω_L/G, the partition of Ω_L into G-orbits is denoted by $\mathcal{S}_L$, whereas the embedding partition $\mathbb{R}^4/G$ is written as $\mathcal{S}_{\mathbb{R}}$.

Next, we prove Proposition 6.8 which states:

The size of the partition $\mathcal{S}$ of $\Omega = M_{2\times 2}(\mathcal{C}_L)$ $(L = 2n)$ into the G-invariant classes (orbits) is $|\mathcal{S}_L| = \frac{L^4+2L^3+6L^2+4L}{16} = n^4 + n^3 + \frac{n(1+3n)}{2}$. Among them, there are L orbits of size two, $\frac{L^2}{4}$ orbits of size four, $\frac{2L^3+3L^2-10L}{8}$ orbits of size eight, and $\frac{L^4-2L^3-4L^2+8L}{16}$ orbits of size 16.

Proof. The $n = 1$ case is special but trivial. There are two orbits of size two:

$$\left\{\begin{matrix}-\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & -\frac{1}{2}\end{matrix}, \begin{matrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\end{matrix}\right\}, \left\{\begin{matrix}-\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{matrix}, \begin{matrix}\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{matrix}\right\},$$

one orbit of size four:

$$\left\{\begin{matrix}-\frac{1}{2} & -\frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\end{matrix}, \begin{matrix}\frac{1}{2} & -\frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{matrix}, \begin{matrix}\frac{1}{2} & \frac{1}{2}\\ -\frac{1}{2} & -\frac{1}{2}\end{matrix}, \begin{matrix}-\frac{1}{2} & \frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{matrix}\right\},$$

and one orbit of size eight:

$$\left\{\begin{matrix}\frac{1}{2} & -\frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\end{matrix}, \begin{matrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{matrix}, \begin{matrix}\frac{1}{2} & \frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{matrix}, \begin{matrix}-\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\end{matrix}, \begin{matrix}-\frac{1}{2} & \frac{1}{2}\\ -\frac{1}{2} & -\frac{1}{2}\end{matrix}, \begin{matrix}-\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{matrix}, \begin{matrix}-\frac{1}{2} & -\frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{matrix}, \begin{matrix}\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & -\frac{1}{2}\end{matrix}\right\}$$

To prove the general case, one first recalls that $\forall \mathcal{O}, \forall \omega \in \mathcal{O}$, $|\mathcal{O}| = |G : G_\omega|$, the size of the orbit $\mathcal{O}$ equals the index of the stabilizer G_ω.

Since $|G| = 16$, $|\mathcal{O}|$ can only be $1, 2, 4, 8, 16$. Clearly, there is no ω with $G_\omega = G$ because $i(\omega) = \omega$ has no solution. For the same reason G_ω can not contain i, si, or r^2si among its generators. This leaves only two copies of D_8 (i.e. $\langle r, s | r^4 = s^2 =$

$1, rs = sr^3\rangle$ and $\langle ri, s | (ri)^4 = s^2 = 1, (ri)s = s(ri)^3\rangle)$ as possible stabilizers of index two. The first group gives rise to the two equations $r(\omega) = \omega$ and $s(\omega) = \omega$ with L solutions of the form $\left(\begin{smallmatrix} \lambda & \lambda \\ \lambda & \lambda \end{smallmatrix}\right), \lambda \in \mathcal{C}_L$, thus yielding $L/2$ orbits of size two. The second choice implies that $(ri)(\omega) = \omega$ and $s(\omega) = \omega$, resulting in the 2^n patches of the form $\left(\begin{smallmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{smallmatrix}\right), \lambda \in \mathcal{C}_L$ that are partitioned into $L/2$ size-two orbits. Hence, the total number of size-two orbits becomes L.

We now count orbits of size four. The following subgroups are the only subgroups of G of index four not containing i, si, or r^2si: $\langle r\rangle$, $\langle ri\rangle$, $\langle r^3i\rangle$, $\langle r^2, s\rangle$, $\langle r^2, rs\rangle$, $\langle r^2, rsi\rangle$, $\langle r^2i, rs\rangle$, $\langle r^2i, rsi\rangle$. Since all the ω's fixed by the rotation group are necessarily fixed by the entire $\langle r, s | r^4 = s^2 = 1, rs = sr^3\rangle$ group, the rotation group can not be a proper stabilizer itself. Similarly, $(ri)(\omega) = \omega \Rightarrow s(\omega) = \omega$ implies that $\langle ri\rangle$ is a proper subgroup of a larger stabilizer, and for the same reason $(r^3i)(\omega) = \omega \Rightarrow s(\omega) = \omega$ makes it impossible for $\langle r^3i\rangle$ to be a stabilizer. Now notice, $\langle r^2, rs\rangle$ can not be a proper stabilizer since $[(r^2)(\omega) = \omega] \wedge [(rs)(\omega) = \omega] \Rightarrow r(\omega) = \omega$[1]; $\langle r^2, rsi\rangle$ can not be a proper stabilizer because $[(r^2)(\omega) = \omega] \wedge [(rsi)(\omega) = \omega] \Rightarrow (ri)(\omega) = \omega$. Finally, $\langle rs, r^3s\rangle$ fails to be a stabilizer since $[(rs)(\omega) = \omega] \wedge [r^2(\omega) = \omega] \Rightarrow r(\omega) = \omega$.

Next, $\langle r^2, s\rangle$ is a stabilizer for all elements of the form: $\left(\begin{smallmatrix} \lambda & \gamma \\ \gamma & \lambda \end{smallmatrix}\right)$, where $\gamma, \lambda \in \mathcal{C}_L, \gamma \neq \lambda,\ \gamma \neq -\lambda$. Since there are $L(L-2)$ such matrices, and the orbit of each of them consists of matrices of the same form (up to renaming of λ and γ), they must form exactly $L(L-2)/4$ size-four orbits.

Matrices of the form $\left(\begin{smallmatrix} -\lambda & -\lambda \\ \lambda & \lambda \end{smallmatrix}\right)$, with $\lambda \in \mathcal{C}_L$ are stabilized by $\langle r^2i, rs\rangle$. In fact, these will represent only $L/2$ distinct matrices as λ runs effectively only through half of the range $\mathcal{C}_L$. Since no two distinct such matrices fall into the same orbit, we obtain $L^2/4$ as the total number of size-four orbits. We also notice that the

[1] We use "$\wedge$" to denote the logical *and*.

subgroup $\langle r^2 i, rsi\rangle$ is a stabilizer for the elements of the form $\left(\begin{smallmatrix} \lambda & -\lambda \\ \lambda & -\lambda \end{smallmatrix}\right)$, which are rotationally equivalent to the previous matrices, hence adding no new orbits.

The last task is to compute the number of orbits of size eight. First, we list all the subgroups of index eight (thus, order two) not containing i, si, or r^2si. These are: $\langle r^2\rangle$, $\langle r^2\rangle$, $\langle r^2 i\rangle$, $\langle s\rangle$, $\langle r^2 s\rangle$, $\langle rs\rangle$, $\langle r^3 s\rangle$, $\langle rsi\rangle$, and $\langle r^3 si\rangle$. $\langle r^2\rangle$ immediately leaves the list since it is a proper subgroup of a larger stabilizer ($r^2(\omega) = \omega \Rightarrow s(\omega) = \omega$). Matrices of the form $\left(\begin{smallmatrix} \lambda & \delta \\ \gamma & \lambda \end{smallmatrix}\right)$, where $\delta, \gamma, \lambda \in \mathcal{C}_L,\ \gamma \neq \delta$, are stabilized by $\langle s\rangle$, whereas rotationally equivalent to them matrices of the form $\left(\begin{smallmatrix} \delta & \lambda \\ \lambda & \gamma \end{smallmatrix}\right)$ are stabilized by $\langle r^2 s\rangle$. Since size-eight orbits generated by these $2L^2(L-1)$ matrices are composed of these matrices only, we arrive at $L^2(L-1)/4$ distinct orbits of size eight. Next, observe that $\langle rs\rangle$ fixes $L(L-2)$ matrices of the form $\left(\begin{smallmatrix} \gamma & \gamma \\ \lambda & \lambda \end{smallmatrix}\right)$, with $\gamma, \lambda \in \mathcal{C}_L,\ \gamma \neq \lambda,\ \gamma \neq -\lambda$, whereas their $L(L-2)$ rotational equivalents $\left(\begin{smallmatrix} \lambda & \gamma \\ \lambda & \gamma \end{smallmatrix}\right)$ are fixed by $\langle r^3 s\rangle$. Since all the matrices inside the corresponding orbits of size eight are of either of the two forms, we add $L(L-2)/4$ orbits of size eight. The same number of $L(L-2)/4$ size-eight orbits come from $L(L-2)$ matrices of the form $\left(\begin{smallmatrix} -\lambda & -\gamma \\ \lambda & \gamma \end{smallmatrix}\right)$ fixed by $\langle r^3 si\rangle$, with $\gamma, \lambda \in \mathcal{C}_L,\ \gamma \neq \lambda,\ \gamma \neq -\lambda$, and from their $L(L-2)$ rotational equivalents of the form $\left(\begin{smallmatrix} \gamma & -\gamma \\ \lambda & -\lambda \end{smallmatrix}\right)$ fixed by $\langle rsi\rangle$. The last source of size-eight orbits is matrices stabilized by $\langle r^2 i\rangle$. They are represented by $\left(\begin{smallmatrix} -\gamma & -\lambda \\ \lambda & \gamma \end{smallmatrix}\right)$, where $\gamma \neq \lambda,\ \gamma \neq -\lambda$. There are exactly $L(L-2)$ such matrices, producing the last $L(L-2)/8$ orbits of size eight.

Summing over orbits of sizes less than 16, we get $2 \times L + 4 \times L^2/4 + 8 \times (L^3/4 + 3L^2/8 - 5L/4)$ as the total number of elements in these orbits. Hence, the number of orbits of size 16 is $(L^4 - 2L^3 - 4L^2 + 8L)/16 = n^4 - n^3 - n^2 + n$. Finally, the total number of orbits is $\frac{L^4+2L^3+6L^2+4L}{16} = n^4 + n^3 + \frac{n(3n+1)}{2}$. ◇

D.1 Fundamental Invariants of $\mathbb{R}[x_1, x_2, x_3, x_4]^G$

We now provide an inductive proof of Lemma 6.13, which states:

The following set of polynomials generates $\mathbb{R}[x_1, x_2, x_3, x_4]^G$*:*

$$\begin{aligned}
f_1(x) &= (x_1 + x_3)(x_2 + x_4), \\
f_2(x) &= x_1x_3 + x_2x_4, \\
f_3(x) &= x_1^2 + x_2^2 + x_3^2 + x_4^2, \\
f_4(x) &= x_1x_2x_3x_4, \\
f_5(x) &= (x_1^2 + x_3^2)(x_2^2 + x_4^2).
\end{aligned} \tag{D.1}$$

Proof. It is immediate to see that $f_1, \ldots, f_5$ respect the action of r, s, i, generators of G. Therefore, they $f_1, \ldots, f_5 \in \mathbb{R}[x_1, x_2, x_3, x_4]^G$. We base our computations on a sequence of decompositions of the original G action, first step of which is given by:

$$\begin{gathered}
\mathcal{S}_{\mathbb{R}} \cong (\mathbb{R}^4/G_1) \,/\, (G/G_1)\,, \\
\text{where } G_1 = \langle s, r^2 | s^2 = (r^2)^2 = 1, r^2s = sr^2 \rangle \trianglelefteq G
\end{gathered} \tag{D.2}$$

The equation above simply says that the original action of G on $\mathbb{R}^4$ decomposes into two actions as follows: First, G_1, a *normal subgroup* of G, acts on $\mathbb{R}^4$, producing the orbit set $\mathbb{R}^4/G_1$, and then the *quotient group* G/G_1 acts on $\mathbb{R}^4/G_1$, resulting in "the same" orbits $\mathcal{S}_{\mathbb{R}}$, just as if G acted on $\mathbb{R}^4$ directly. Thus, we first aim to find $y_1(x), \ldots, y_k(x)$ for some k, generators for $\mathbb{R}[x]^{G_1}$, and then will focus on the polynomials (in those generators) that are invariant under G/G_1.

Claim D.4 $\mathbb{R}[x]^{G_1} = \mathbb{R}[x_1 + x_3, x_2 + x_4, x_1x_3, x_2x_4]$.

Proof. It suffices to prove that $\mathbb{R}[x]^{\langle r^2 s\rangle} = \mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$ and $\mathbb{R}[x]^{\langle s\rangle} = \mathbb{R}[x_1, x_2+x_4, x_3, x_2x_4]$, since $\mathbb{R}[x_1+x_3, x_2+x_4, x_1x_3, x_2x_4] = \mathbb{R}[x_1+x_3, x_2, x_1x_3, x_4] \cap$

$\mathbb{R}[x_1, x_2+x_4, x_3, x_2x_4]$. In fact, we only prove the first of these statements since the second one proves along the same lines interchanging x_1 with x_2 and x_3 with x_4. We argue by induction on the *degree* function, $\deg = \deg_1 + \deg_2 + \deg_3 + \deg_4$, where $\deg_k$ is the highest power of x_k $(k = 1, 2, 3, 4)$ in a given polynomial. Let us begin by noticing that the result holds for all polynomials of $\deg = 0$ (i.e. constants.) Assume now that the result is true for $\deg \leq N$, $N \geq 0$ and show that it also holds for $\deg = N + 1$. A generic polynomial $r(x_1, x_2, x_3, x_4) \in \mathbb{R}^{\langle r^2 s\rangle}[x]$ such that $\deg(r) \leq N + 1$ has the form:

$$\sum_{\substack{i,j,k,l\geq 0\\ i+j+k+l\leq N+1}} a_{i,j,k,l}x_1^i x_2^j x_3^k x_4^l = \overbrace{\sum_{\substack{i,k\geq 0\\ i+k\leq N}} a_{i,0,k,0}x_1^i x_3^k}^{1} + \overbrace{a_{N+1,0,0,0}x_1^{N+1} + a_{0,0,N+1,0}x_3^{N+1}}^{2} + \tag{D.3}$$

$$\overbrace{x_1x_3 \sum_{\substack{i,k>0\\ i+k=N+1}} a_{i,0,k,0}x_1^{i-1}x_3^{k-1}}^{3} + \sum_{\substack{j,l\geq 0\\ 0<j+l\leq N+1}} \left(\overbrace{\sum_{\substack{i,k\geq 0\\ 0\leq i+k\leq N+1-j-l}} a_{i,j,k,l}x_1^i x_3^k}^{4} \right) x_2^j x_4^l \tag{D.4}$$

In order for the left hand side to be invariant under $x_1 \leftrightarrow x_3$, each of the terms $1 - 4$ in (D.3)-D.4 must be invariant under the same action. By the induction argument, terms of degree N and below are already in the desired form. Thus, the first sum and all the sums labeled 4 belong to $\mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$. This implies that the entire double sum of (D.4) is in $\mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$. The cofactor of x_1x_3 in the third term of (D.4) is also invariant and has degree N, hence lies in $\mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$ as well. The invariance of the second term of (D.3) forces $a_{N+1,0,0,0} = a_{0,0,N+1,0}$. We now notice that if $N = 0$, then

$$a_{N+1,0,0,0}x_1^{N+1} + a_{0,0,N+1,0}x_3^{N+1} = a_{1,0,0,0}(x_1 + x_3) \in R[x_1 + x_3, x_2, x_1x_3, x_4]$$

For $N \geq 1$, on the other hand,

$$x_1^{N+1} + x_3^{N+1} = (x_1 + x_3)(x_1^N + x_3^N) - x_1x_3(x_1^{N-1} + x_3^{N-1}) \in R[x_1 + x_3, x_2, x_1x_3, x_4]$$

by the induction argument. This shows that the left hand side of (D.3),(D.4) belongs to $\mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$. $\diamond$

Thus, we have obtained a set of generators for $\mathbb{R}[x]^{G_1}$:

$$y_1 = x_1 + x_3,\ y_2 = x_2 + x_4,\ y_3 = x_3x_4,\ y_4 = x_2x_4, \tag{D.5}$$

which are algebraically independent. We now want to find $\mathbb{R}^{G/G_1}[y_1, y_2, y_3, y_4]$. Recall that

$$G/G_1 = \{1, \bar{r}, \bar{\imath}, \overline{\imath r}\}$$

and that its action on the orbit set $\mathbb{R}^4/G_1$ translates into

$$\bar{r} : y_1 \leftrightarrow y_2, \qquad y_3 \leftrightarrow y_4$$

$$\bar{\imath} : y_1 \mapsto -y_1,\ y_2 \mapsto -y_2, \qquad y_3 \leftrightarrow y_3,\ y_4 \leftrightarrow y_4$$

Continuing (D.2) to decompose the original G action, we write:

$$(\mathbb{R}^4/G_1)/(G/G_1) \cong \big((\mathbb{R}^4/G_1)/G_2\big)\Big/\big(G/G_1/G_2\big),\ \text{where } G_2 = \langle\bar{\imath}\rangle \trianglelefteq G/G_1 \tag{D.6}$$

Claim D.5 $\mathbb{R}[y_1, y_2, y_3, y_4]^{G_2} = \mathbb{R}[y_1^2, y_2^2, y_1y_2, y_3, y_4]$

Proof. Using induction just as in the proof of Claim D.4, we can simply imagine replacing x_1 with y_1, x_3 with y_2, x_2 with y_3, and x_4 with y_4, which yields equations essentially identical to (D.3),(D.4):

$$\begin{aligned}\sum_{\substack{i,j,k,l\geq 0\\ i+j+k+l\leq N+1}} a_{i,j,k,l}y_1^iy_2^jy_3^ky_4^l = \sum_{\substack{i,j\geq 0\\ i+j\leq N}} a_{i,j,0,0}y_1^iy_2^j + \overbrace{a_{N+1,0,0,0}y_1^{N+1} + a_{0,N+1,0,0}y_2^{N+1}}^{2} + \\ y_1y_2\sum_{\substack{i,j>0\\ i+j=N+1}} a_{i,j,0,0}y_1^{i-1}y_2^{j-1} + \sum_{\substack{k,l\geq 0\\ 0<k+l\leq N+1}}\left(\sum_{\substack{i,j\geq 0\\ 0\leq i+j\leq N+1-k-l}} a_{i,j,k,l}y_1^iy_2^j\right)y_3^ky_4^l\end{aligned} \tag{D.7}$$

The only other difference from the previous proof is as follows: The new second term D.7 disappears if $N+1$ is odd, whereas even $N+1$ immediately yields the needed form, i.e. $y_{1,2}^{N+1} = (y_{1,2}^2)^{(N+1)/2}$. ◇

Next, notice:

$$\mathbb{R}[y_1^2, y_2^2, y_1y_2, y_3, y_4] \cong \mathbb{R}[z_1, z_2, z_3, z_4, z_5]\big/\langle z_1z_2 - z_5^2\rangle,$$

under:

$$y_1^2 \rightarrow z_1,\ y_2^2 \rightarrow z_2,\quad y_3 \rightarrow z_3,\quad y_4 \rightarrow z_4,\ y_1y_2 \rightarrow z_5.$$

We now show by induction that

$$\left(\mathbb{R}[z_1, z_2, z_3, z_4, z_5]\big/\langle z_1z_2 - z_5^2\rangle\right)^{(G/G_1)\big/G_2} = \mathbb{R}[z_1+z_2, z_3+z_4, z_3z_4, z_1z_3+z_2z_4, z_5], \tag{D.8}$$

where $(G/G_1)\big/G_2 = \langle\overline{\overline{r}}\rangle$, and its action results in exchanging z_1 with z_2 and z_3 with z_4. First, denote the right hand side of (D.8) by R and focus on the inductive transition from deg $\leq N$ to deg $= N+1$. A generic polynomial of interest splits into two sums, one with deg $\leq N$ and the other - with deg $= N+1$, each of which is separately invariant under the action of $\overline{\overline{r}}$. Since the first sum is in R by the induction assumption, we continue on to decompose the second one as follows:

$$\sum_{\substack{i,j,k,l\geq 0\\ i+j+k+l=N+1}} a_{i,j,k,l} z_1^i z_2^j z_3^k z_4^l = \overbrace{z_1z_2z_3z_4 \sum_{\substack{i,j,k,l>0\\ i+j+k+l=N+1}} a_{i,j,k,l} z_1^{i-1} z_2^{j-1} z_3^{k-1} z_4^{l-1}}^{1} + \tag{D.9}$$

$$\overbrace{z_1z_2\left(\sum_{\substack{i,j,k>0\\ i+j+k=N+1}} a_{i,j,k,0} z_1^{i-1} z_2^{j-1} z_3^k + \sum_{\substack{i,j,l>0\\ i+j+l=N+1}} a_{i,j,0,l} z_1^{i-1} z_2^{j-1} z_4^l\right)}^{2} +$$

$$\overbrace{z_3z_4\left(\sum_{\substack{i,k,l>0\\ i+k+l=N+1}} a_{i,0,k,l} z_1^i z_3^{k-1} z_4^{l-1} + \sum_{\substack{j,k,l>0\\ j+k+l=N+1}} a_{0,j,k,l} z_2^j z_3^{k-1} z_4^{l-1}\right)}^{3} +$$

$$\overbrace{z_1 z_2 \sum_{\substack{i,j>0\\ i+j=N+1}} a_{i,j,0,0} z_1^{i-1} z_2^{j-1}}^{4} + \overbrace{z_3 z_4 \sum_{\substack{k,l>0\\ k+l=N+1}} a_{0,0,k,l} z_3^{k-1} z_4^{l-1}}^{5} +$$

$$\overbrace{\sum_{\substack{i,k>0\\ i+k=N+1}} a_{i,0,k,0} z_1^i z_3^k + \sum_{\substack{j,l>0\\ j+l=N+1}} a_{0,j,0,l} z_2^j z_4^l}^{6} + \overbrace{\sum_{\substack{i,l>0\\ i+l=N+1}} a_{i,0,0,l} z_1^i z_4^l + \sum_{\substack{j,k>0\\ j+k=N+1}} a_{0,j,k,0} z_2^j z_3^k}^{7} +$$

$$\overbrace{a_{N+1,0,0,0} z_1^{N+1} + a_{0,N+1,0,0} z_2^{N+1}}^{8} + \overbrace{a_{0,0,N+1,0} z_3^{N+1} + a_{0,0,0,N+1} z_4^{N+1}}^{9}$$

An immediate inspection of (D.9) combined with the symmetry of the coefficients $a_{i,j,k,l} = a_{j,i,l,k}$ reveals that each of the terms numbered one through nine is individually invariant under the the given action. By the inductive argument, terms one through five are already in R, and following the pattern of the second term of (D.3) eventually shows that terms eight and nine are also in R. We now rewrite the sum of terms six and seven as follows:

$$\sum_{\substack{i,k>0\\ i+k=N+1}} a_{i,0,k,0} \left(z_1^i z_3^k + z_2^i z_4^k\right) + \sum_{\substack{i,k>0\\ i+k=N+1}} a_{i,0,0,k} \left(z_1^i z_4^k + z_2^i z_3^k\right)$$

Observe that for $i, k > 0$:

$$z_1^i z_3^k + z_2^i z_4^k = (z_1 z_3 + z_2 z_4)(z_1^{i-1} z_3^{k-1} + z_2^{i-1} z_4^{k-1}) - z_1^{i-1} z_2 z_3^{k-1} z_4 - z_1 z_2^{i-1} z_3 z_4^{k-1} \quad \text{(D.10)}$$

$$z_1^i z_4^k + z_2^i z_3^k = (z_1 z_4 + z_2 z_3)(z_1^{i-1} z_4^{k-1} + z_2^{i-1} z_3^{k-1}) - z_1 z_2^{i-1} z_3^{k-1} z_4 - z_1^{i-1} z_2 z_3 z_4^{k-1}$$

We conclude by considering the first of the two equations above and noticing that the second equation can be treated similarly due to that $z_1 z_4 + z_2 z_3$ equals $(z_1 + z_2)(z_3 + z_4) - (z_1 z_3 + z_2 z_4)$, and thus lies in R. The following expression in conjunction with the induction argument helps to see why the left hand side of (D.10) belongs to R:

$$z_1^{i-1}z_2z_3^{k-1}z_4+z_1z_2^{i-1}z_3z_4^{k-1} = \begin{cases} z_1z_3+z_2z_4, \text{ if } \ i-1=k-1=0 \\ z_3z_4(z_2z_3^{k-2}+z_1z_4^{k-2}), \text{ if } \ i-1=0, \ k-1>0 \\ z_1z_2(z_1^{i-2}z_4+z_2^{i-2}z_3), \text{ if } \ i-1>0, \ k-1=0 \\ z_1z_2z_3z_4(z_1^{i-2}z_3^{k-2}+z_2^{i-2}z_4^{k-2}), \text{ if } \ i-1,k-1>0. \end{cases}$$

Summarizing the results proved to this point, we return to the initial x indeterminates:

$$\mathbb{R}[x_1,x_2,x_3,x_4]^G = \mathbb{R}[(x_1+x_3)^2+(x_2+x_4)^2, x_1x_3+x_2x_4, \tag{D.11}$$
$$x_1x_2x_3x_4, (x_1+x_3)^2x_1x_3+(x_2+x_4)^2x_2x_4, (x_1+x_3)(x_2+x_4)]$$

These generators are not unique, and recognizing that

$$(x_1+x_3)^2+(x_2+x_4)^2 = f_3(x)+2f_2(x),$$
$$(x_1+x_3)^2x_1x_3+(x_2+x_4)^2x_2x_4 = \tfrac{1}{2}[f_5(x)-f_1^2(x)]+$$
$$f_2(x)f_3(x)+2f_2^2(x)-2f_4(x),$$

with f_1, f_2, f_3, f_4, f_5 as in (6.16), makes it clear that

$$\mathbb{R}[x_1,x_2,x_3,x_4]^G = \mathbb{R}[f_1(x), f_2(x), f_3(x), f_4(x), f_5(x)].$$

A straightforward computation verifies that none of the above five generators can be expressed as a real polynomial in the remaining four. We conclude by instantiating a well-known fact (see, for example, [12]):

$$\mathbb{R}[x_1,x_2,x_3,x_4]^G \cong \mathbb{R}[w_1,w_2,w_3,w_4,w_5]/J_F, \text{ where} \tag{D.12}$$
$$J_F = \{h \in \mathbb{R}[w_1,w_2,w_3,w_4,w_5] : h(f_1,f_2,f_3,f_4,f_5) = 0 \in \mathbb{R}[x_1,x_2,x_3,x_4]\} =$$
$$\langle q \rangle, \text{ and } q(w_1,w_2,w_3,w_4,w_5) = 4w_1^2w_3+8w_1w_2w_5+2w_1w_3w_5-2w_1w_4^2w_5+$$

$$16w_2^2 - 8w_2w_3 - 8w_2w_4^2 + 4w_2w_5^2 + w_3^2 - 2w_3w_4^2 + w_4^4 \tag{D.13}$$

In order to compute J_F, the *syzygy* ideal, one can use, for example, the *elimination method* based on computation of a *Gröbner basis* for the ideal $J_F = \langle f_2 - w_1, f_4 - w_2, f_5 - w_3, f_1 - w_4, f_3 - w_5 \rangle \subset \mathbb{R}[x_1, x_2, x_3, x_4, w_1, w_2, w_3, w_4, w_5]$ [12]. The above generator for J_F was computed analytically and later verified using *Macaulay2* [16]. $\diamond$

D.2 Comments on Proposition 6.14

In Proposition 6.14 we stated that f_1, f_2, f_3, f_4 (6.16) suffice to enumerate $\mathcal{S}_L = M_{2\times 2}(\mathcal{C}_L)/G$ for $L \leq 8$ and $L = 10, 12$, and that f_1, f_2, f_4 fully enumerate $\mathcal{S}_{2n}$ for $n \leq 4$. These statements were observed by means of a straightforward range counting: Since f_1, f_2, f_3, f_4, f_5 are constant on the orbits of $\mathcal{S}_L$, we defined, for example, the map $\bar{F} = (\bar{f}_1, \bar{f}_2, \bar{f}_4) : \mathcal{S}_L \to \mathbb{R}^3$ by setting each component of $\bar{F}(\mathcal{O})$ to equal the common value of the respective generator f on $\mathcal{O}$. The size of the range of $\bar{F}$ was calculated to be equal to $|\mathcal{S}_{2n}|$ for $n \leq 4$. The other statement was verified in the same manner.

These and other properties of the generators f_1, f_2, f_3, f_4, f_5 in the context of restricting $\mathbb{R}[x_1, x_2, x_3, x_4]^G$ to the G-symmetric functions on Ω_L can also be analyzed using symbolic computations provided by algebraic geometry. We conclude this supplement by briefly outlining the framework of such computations.

Observe that we can identify all G-invariant functions on Ω_L with the elements of the quotient ring (algebra) $\mathbb{R}[x_1, x_2, x_3, x_4]^G/J_L$, where $J_L = \{f \in \mathbb{R}[x_1, x_2, x_3, x_4]^G : f(\omega) = 0 \; \forall \omega \in \Omega_L\}$ is the ideal of G-invariant polynomials vanishing on all the orbits. (J_L is also the *ideal of the real affine variety* Ω_L.) Since the number of orbits in $\mathcal{S}_L$ is finite, $J_L \supsetneqq \{0\}$ is no longer trivial as in the case of $\mathcal{S}_{\mathbb{R}}$.

Now, let ϕ be the ring isomorphism of (D.12). We then have an isomorphism of the corresponding quotients:

$$\mathbb{R}[x_1, x_2, x_3, x_4]^G / J_L \cong \mathbb{R}[w_1, w_2, w_3, w_4, w_5] / J_F / \phi(J_L) \tag{D.14}$$

For the value of L of interest, computations to support the respective claim of Proposition 6.14 (or the conjecture) are likely to involve finding a *Gröbner basis* for the right hand side of (D.14).

APPENDIX E

SOME CORRECTIONS

After the thesis had already been processed by the University, I discovered a mistake in one of the models by an analytic verification of the previously used numerical computations: In §4.7, p^{g}_{pair} is introduced as the "pair-potential geometrically symmetric" model, but the computation of the corresponding parameter space was wrong and produced a more complex model of dimension 42 (see Table 16) instead of 15, the true number of independent parameters of p^{g}_{pair}. Consequently, the results reported in §5.3 (Table 11) do not correspond to p^{g}_{pair}.

BIBLIOGRAPHY

[1] Y. Amit, D. Geman, and B. Jedynak. Efficient focusing and face detection. In H. Wechsler and J. Phillips, editors, *Face Recognition: From Theory to Applications, NATO ASI Series F*. Springer-Verlag, Berlin, 1998.

[2] Y. Amit, D. Geman, and K. Wilder. Joint induction of shape features and tree classifiers. *IEEE Trans. PAMI*, 19:1300–1306, 1997.

[3] L. Bain and M. Englehardt. *Introduction to Probability and Mathematical Statistics*. Wadsworth Publishing Company, 2nd edition edition, 1991.

[4] H. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

[5] A. J. Bell and T. J. Sejnowski. Edges are the "independent components" of natural scenes. Technical report, Computational Neurobiology Laboratory, The Salk Institute, La Jolla, CA 92037, 1996.

[6] K. Binder and D.W. Heerman. *Monte Carlo Simulation in Statistical Physics*. Springer-Verlag Berlin Heidelberg, 1992.

[7] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press Inc., New York, 1995.

[8] R. D. Boss and E. W. Jacobs. Archetype classification in an iterated transformation image compression algorithm. In Y. Fisher, editor, *Fractal Image Compression - theory and Application*, pages 79–90. Springer-Verlag, New York, 1994.

[9] Z. Chi. *Probability Models for Complex Systems*. PhD thesis, Brown University, 1998.

[10] W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, Inc., 1964.

[11] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, New York, 1991.

[12] D. Cox, J. Little, and O'Shea D. *Ideals, Variety, and Algorithms*. Springer, 1996.

[13] D. Cox and B. Sturmfels, editors. *Applications of Computational Algebraic Geometry.* American Mathematical Society, 1997.

[14] N. Cressie. *Statistics for Spatial Data.* John Wiley & Sons, Inc., 1991.

[15] W. E. Deming. *Some Theory of Sampling.* John Wiley & Sons, Inc., 1950.

[16] Department of Mathematics, The University of Illinois, http://www.math.uiuc.edu/Macaulay2/Manual. *Macaulay 2.*

[17] R. Dubes and A. Jain. Random field models in image analysis. *Journal of Applied Statistics*, 1989.

[18] D. S. Dummit and R. M. Foote. *Abstract Algebra.* Prentice Hall, Inc., 1991.

[19] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Optical Soc. America*, 4, 1987.

[20] J. Fogarty. On Noether's bound for the degrees of generating invariants. Technical report, Department of Mathematics and Statistics, University of Massachusetts Amherst, 1999.

[21] D. Geman and B. Jedynak. Model-based classification trees. Technical report, Department of Mathematics and Statistics, University of Massachusetts at Amherst, November, 1998.

[22] D. Geman and A. Koloydenko. Invariant statistics and coding of natural microimages. In S.C. Zhu, editor, *IEEE Workshop on Statistical and Computational Theories of Vision*, http://www.cis.ohio-state.edu/∼szhu/sctv99/Geman1.html, 1999.

[23] S. Geman. Invariant binding in composition systems. Technical report, Division of Applied Mathematics, Brown University, 1998.

[24] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[25] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 1984.

[26] J. Goutsias. Mutually compatible Gibbs random fields. *IEEE Transactions on Information Theory*, 35(6), 1989.

[27] J. Goutsias and H. J. A. M. Heijmans. Fundamenta morphologicae mathematicae. *Fundamenta Informaticae*, 2000.

[28] P. Greenwood and M. Nikulin. *A Guide to Chi-Squared Testing.* John Wiley & Sons, Inc., 1996.

[29] D. Griffeath. Introduction to random fields. In Knap Kemeny and Snell, editors, *Denumerable Markov Chains.* Springer-Verlag New York Inc., 1976.

[30] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc.R.Soc.Lond.*, B 265:359–366, 1998.

[31] W. W. Hauck. A note on confidence bands for the logistic response curve. *The American Statistician*, 37(2), 1983.

[32] H. J. A. M. Heijmans and C. Ronse. The algebraic basis of mathematical morphology - part I: Dilations and erosions. *Computer Vision, Graphics and Image Processing*, 1990.

[33] J. Huang. *Statistics of Natural Images and Models.* PhD thesis, Brown University, 2000.

[34] J. Huang and D. Mumford. Statistics of natural images and models. In *Computer Vision and Pattern Recognition*, 1999.

[35] J. Huang and D. Mumford. Image statistics for the British aerospace segmented database. Technical report, Division of Applied Mathematics, Brown University, 2000.

[36] M. F. Janowitz. A model for ordinal filtering of digital images. In D. J. DePriest and E. J. Wegman, editors, *Statistical Image Processing and Graphics*, pages 25–41. Marcel Dekker, Inc., 1986.

[37] G. Kanji. *100 Statistical Tests.* SAGE Publications Ltd, 1993.

[38] R. Kinderman and J. L. Snell. *Markov Random Fields and their Applications.* American Mathematical Society, 1980.

[39] S. A. Kovalenko. Photographic gallery of Sergei Kovalenko. Web published at http://www.umass.edu/tei/ogia/gallery, 1999-2000.

[40] S. Kullback. *Information Theory and Statistics.* Dover Publications, Inc., 1997.

[41] S. L. Lauritzen. *Graphical Models.* Oxford University Press Inc., New York, 1996.

[42] A. Lee and D. Mumford. An occlusion model generating scale invariant images. In S.C. Zhu, editor, *IEEE Workshop on Statistical and Computational Theories of Vision*, http://www.cis.ohio-state.edu/∼szhu/workshop/Lee.html, 1999.

[43] T-W. Lee, M. Girolami, Bell A. J., and Sejnowski T. J. A unifying information-theoretic framework for independent component analysis. *International Journal on Mathematical and Computer Modeling*, 1998.

[44] M. S. Lewicki and B. A. Olshausen. A probabilistic framework for adaptation and comparison of image codes. Technical report, Computational Neurobiology Lab, The Salk Institute and Center for Neuroscience, University of California, Davis, 1998.

[45] Chunming Li. *Classification by active testing with applications to imaging and change detection.* PhD thesis, University of Massachusetts, Amherst, 1999.

[46] J. K. Lindsey. *Parametric Statistical Inference.* Oxford University Press Inc., 1996.

[47] A. Lippman. *A maximum entropy method for expert system construction.* PhD thesis, Brown University, 1986.

[48] J. Malik, S. Belongie, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *International Conference on Computer Vision*, 1999.

[49] The MathWorks, Inc., http://www.mathworks.com. *Getting Started with Matlab.*

[50] D. Mumford and B. Gidas. Stochastic models for generic images. Technical report, Division of Applied Mathematics, Brown University, January, 2000.

[51] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. In *Proc. Workshop on Information Theory and the Brain.* 1995.

[52] P. J. Olver. *Classical Invariant Theory.* Cambridge University Press, 1999.

[53] S. Port. *Theoretical Probability for Application.* John Wiley & Sons, Inc., 1994.

[54] T. Read and N. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer-Verlag New York, Inc., 1988.

[55] B. D. Ripley. *Stochastic simulation.* John Wiley & Sons, Inc., 1987.

[56] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[57] C. Ronse. Fourier analysis, mathematical morphology, and vision. Technical Report WD54, Philips Research Laboratory Brussels, 1989.

[58] D. L. Ruderman. Origins of scaling in natural images. *Vision Research*, pages 814–817, 1996.

[59] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Phys. Rev. Letters*, 73:814–817, 1994.

[60] D. Saupe, R. Hamzaoui, and H. Hartenstein. Fractal image compression - an introductory overview. In J. Hart, editor, *Fractal Models for Image Synthesis, Compression, and Analysis*, ACM SIGGRAPH'96 Course Notes. 1996.

[61] M. Schervish. *Theory of Statistics*. Springer-Verlag New York, Inc., 1997.

[62] J. Shao. *Mathematical Statistics*. Springer-Verlag New York, Inc., 1999.

[63] A. N. Shiryaev. *Probability*. Springer-Verlag New York Inc., second edition, 1996.

[64] L. Smith. *Polynomial Invariants of Finite Groups*. A K Peters, Ltd., 1995.

[65] B. Sturmfels. *Algorithms in invariant theory*. Springer-Verlag/Wien, 1993.

[66] K. Wilder. *Decision tree algorithms for handwritten digit recognition*. PhD thesis, University of Massachusetts, Amherst, 1998.

[67] E. Zeidler. *Applied Functional Analysis: Applications to Mathematical Physics*, volume 108 of *Applied Mathematical Sciences*. Springer-Verlag New York, Inc., 1995.

[68] S. C. Zhu. Embedding Gestalt laws in Markov random fields. *IEEE Trans. PAMI*, 21, November 1999.

[69] S. C. Zhu, A. Lanterman, and M. Miller. Clutter modeling and performance and analysis in automatic target recognition. In *Workshop on Detection and Classification of Difficult Targets*. Redstone Arsenal, 1998.

[70] S. C. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Trans. PAMI*, 19:1236–1250, 1997.