Quantification of Order in Point Patterns

Emmanouil D. Protonotarios

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of University College London.

Department of Computer Science
Centre for Mathematics & Physics in the Life Sciences & Experimental Biology
University College London

July 2016
I, Emmanouil D. Protonotarios, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: "Date: 1/7/2016"
Abstract

Pattern attributes are important in many disciplines, e.g. developmental biology, but there are few objective measures of them. Here we concentrate on the attribute of order in point patterns and its objective measurement.

We examine perception of order and develop analysis algorithms that quantify the attribute in accordance with perception of it. Based on pairwise ranking of point patterns by degree of order, we show that judgements are highly consistent across individuals and that the perceptual dimension has an interval scale structure, spanning roughly 10 just-noticeable differences (jnds) between disorder and order. We designed a geometric algorithm that estimates order to an accuracy of half a jnd by quantifying the variability of the spaces between points. By anchoring the output of the algorithm so that Poisson point processes score on average 0, and perfect lattices score 10, we constructed an absolute interval scale of order. We demonstrated its utility in biology by quantifying the order of the Drosophila dorsal thorax epithelium during development.

The psychophysical scaling method used relies on the comparison of stimuli with similar levels of order yielding a discrimination-based scale. As with other perceptual dimensions, an interesting question is whether supra-threshold perceptual differences are consistent with this scale. To test that we collected discrimination data, and data based on comparison of perceptual differences. Although the judgements of perceptual differences were found to be consistent with an interval scale, like the discrimination judgements, no common interval scale that could predict both sets of data was possible.

Point patterns are commonly displayed as arrangements of dots. To examine how presentation parameters (dot size, dot numbers, and pattern area) affect discrimination, we collected discrimination data for ten presentation conditions. We found that discrimination performance depends on the ratio ‘dot diameter / average dot spacing’.
Acknowledgements

I am grateful to my supervisors, Lewis D. Griffin and Alan Johnston, for their guidance and support throughout my PhD, but also before that, during my MRes studies. They have always been available, helpful and encouraging far beyond the strict description of their academic roles. This work would not have been completed without their assistance in overcoming my mental blocks. I also express my gratitude to Professor Buzz Baum for his help and enthusiasm for the project, and to Dr Ginger Hunter and Scott Curran for working hard to provide me with the biological data.

The academic and administrative staff of UCL, particularly in the departments of CoMPLEX and Computer Science have been of incredible support as well. CoMPLEX has not offered simply a Doctoral Training Programme but a complete and unique life experience that I will always remember. During these years I have had the opportunity to interact with very interesting people and share this adventure with amazing colleagues. I have learnt from them and I have enjoyed life in London because of their presence. Being a member of the College Hall Wardenial Team has been a big part of this experience. I am grateful to the Warden, Sangeeta Banerjee, for giving me this opportunity and to all my colleagues and students for the wonderful time we had. Many of them had to suffer participating in my experiments.

I would like to thank the Greek State Scholarships Foundation and the Greek Ministry of Education, which supported my studies during this difficult period for Greece. My advisor Professor Zanni-Vlastou Roza helped so that the whole process ran smoothly and without any problems.

My family and friends have consistently supported me, even if I have not been able to see them as much as I wanted. I owe an apology to all for being so absorbed in my work, and I hope I will soon compensate for that.
# Contents

1 Introduction

1.1 Motivation .................................................. 14
1.2 Problem Statement and Approach ............................... 16
1.3 Thesis Outline ............................................... 19
1.4 Contributions .............................................. 20
1.5 Publications ................................................. 21

2 Background

2.1 Introduction .................................................. 24
2.2 Measurement Theory ......................................... 24
   2.2.1 The Representational Aspect ............................. 24
   2.2.2 The Pragmatic Aspect ..................................... 26
   2.2.3 Quantification of Order as a Measurement Process .... 27
2.3 Perception of Order .......................................... 29
2.4 Psychophysics ................................................ 36
   2.4.1 Psychophysical Scaling Methods ......................... 37
   2.4.2 Model Fitting ............................................. 40
   2.4.3 Linear Models ............................................ 41
   2.4.4 Correction of a Model for Observers’ Lapses ........... 42
2.5 Measures of Order ........................................... 42
   2.5.1 Measures of Order Used in Biology ...................... 43
   2.5.2 Measures of Order from Diverse Scientific Areas ....... 46
2.6 Conclusion ................................................... 56
5.2.1 Stimuli .................................................. 120
5.2.2 Observers .................................................. 123
5.2.3 Procedure .................................................. 124
5.2.4 Magnitude Comparison Experiment .................. 124
5.2.5 Magnitude Difference Comparison Experiment .... 124
5.3 Results ...................................................... 127
5.3.1 Agreement Rates ........................................ 127
5.3.2 Magnitude Comparison Scaling ....................... 128
5.3.3 Magnitude Difference Comparison Scaling .......... 132
5.3.4 Common Interval Scale Model for Both Experiments 134
5.4 Discussion .................................................. 138
5.5 Conclusion .................................................. 142

6 Discrimination Performance for Order in Point Patterns 143
  6.1 Introduction .............................................. 143
  6.2 Methods .................................................. 144
    6.2.1 Experiments .......................................... 144
    6.2.2 Stimuli Generation .................................... 145
    6.2.3 Experimental Design Matrix ......................... 146
    6.2.4 Stimuli Presentation and Task ....................... 146
  6.3 Methods for First Group of Experiments ............... 147
    6.3.1 Presentation Conditions ............................ 147
    6.3.2 Observers ........................................... 148
  6.4 Results for First Group of Experiments ............... 149
    6.4.1 Agreement Rates .................................... 149
    6.4.2 Discrimination Scales ................................ 149
  6.5 Discussion on the First Group of Experiments ........ 150
  6.6 Methods for Second Group of Experiments ............. 153
    6.6.1 Presentation Conditions ............................ 153
    6.6.2 Observers ........................................... 154
  6.7 Results for Second Group of Experiments .............. 155
List of Figures

1.1 Spatial patterns with regularity ........................................ 14
1.2 Changes in the *Drosophila* notum during development ........ 18

2.1 Man-made regular patterns from ancient civilisations .......... 30
2.2 Perception of order in point and polygonal patterns ............ 34
2.3 Diagrammatic presentation of the Thurstone *Case V* model .... 39
2.4 Demonstration of the Voronoi diagram and the Delaunay triangulation based on a point pattern .................. 44

3.1 Depiction of point patterns from set *A* ............................ 61
3.2 Predicted and experimental preference rates ....................... 65
3.3 Visual representation of the *p*-scale of perceptually-based order ... 66
3.4 Comparison between the estimated attribute values derived with the error function and the logistic function models for set *A* ........ 66
3.5 Diagrammatic summary of the deviance GoF test and the estimation of the uncertainty of the estimated order values .......... 68
3.6 Schematic of the *pairwise local symmetry* measure ............. 71
3.7 Schematic of the *centroidal symmetry* measure ................. 72
3.8 Schematic of the *Delaunay sides entropy* measure .............. 74
3.9 Schematic of the *Delaunay triangles entropy* measure .......... 75
3.10 Exact binomial test for the *Delaunay triangles entropy* *g*-scale predictions for set *B* ........................................... 81
3.11 First example of poor *g*-scale performance for set *B* .......... 81
3.12 Second example of poor *g*-scale performance for set *B* ....... 82
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.13</td>
<td>The proposed absolute scale (a-scale) of order</td>
<td>83</td>
</tr>
<tr>
<td>3.14</td>
<td>Sensitivity of the entropy-based measures on the bin size of the histograms</td>
<td>85</td>
</tr>
<tr>
<td>3.15</td>
<td>Histogram of a-scale values for simulated Poisson patterns</td>
<td>87</td>
</tr>
<tr>
<td>4.1</td>
<td>Automatic classification of potential precursor bristle cells</td>
<td>94</td>
</tr>
<tr>
<td>4.2</td>
<td>Manual classification of potential precursor bristle cells</td>
<td>95</td>
</tr>
<tr>
<td>4.3</td>
<td>Border and interior points of a point pattern</td>
<td>97</td>
</tr>
<tr>
<td>4.4</td>
<td>Timecourses of cell numbers and proportion of bright cells</td>
<td>99</td>
</tr>
<tr>
<td>4.5</td>
<td>Data points and linear fits for order versus time for the four movies</td>
<td>100</td>
</tr>
<tr>
<td>4.6</td>
<td>Linear trends of order when all movies are taken into account</td>
<td>101</td>
</tr>
<tr>
<td>4.7</td>
<td>Location of cells in movie A at specific times during development</td>
<td>102</td>
</tr>
<tr>
<td>4.8</td>
<td>Epithelial tissue. Number of cells during development</td>
<td>104</td>
</tr>
<tr>
<td>4.9</td>
<td>Epithelial tissue. Order during development</td>
<td>104</td>
</tr>
<tr>
<td>4.10</td>
<td>Spatial representation of local symmetry for the fly’s notum epithelial tissue</td>
<td>106</td>
</tr>
<tr>
<td>4.11</td>
<td>Epithelial tissue spatial analysis (vertical partitioning). Number of cells during development</td>
<td>108</td>
</tr>
<tr>
<td>4.12</td>
<td>Epithelial tissue spatial analysis (vertical partitioning). Order during development</td>
<td>108</td>
</tr>
<tr>
<td>4.13</td>
<td>Epithelial tissue spatial analysis (horizontal partitioning). Number of cells during development</td>
<td>110</td>
</tr>
<tr>
<td>4.14</td>
<td>Epithelial tissue spatial analysis (horizontal partitioning). Order during development</td>
<td>110</td>
</tr>
<tr>
<td>4.15</td>
<td>Quantification of order for simulation case I</td>
<td>112</td>
</tr>
<tr>
<td>4.16</td>
<td>Quantification of order for simulation case II</td>
<td>113</td>
</tr>
<tr>
<td>4.17</td>
<td>Quantification of order for simulation case III</td>
<td>114</td>
</tr>
<tr>
<td>4.18</td>
<td>Quantification of order for simulation case IV</td>
<td>115</td>
</tr>
<tr>
<td>5.1</td>
<td>Diagrammatic presentation of the Thurstone Case V model for two stimuli placed far apart on the perceptual scale</td>
<td>118</td>
</tr>
</tbody>
</table>
List of Figures

5.2 Boxplot of $a$-scale order values for generated point patterns for varying jitter levels ........................................ 121
5.3 $a$-scale order values for selected point patterns for the magnitude and difference comparison tasks against the corresponding jitter levels 123
5.4 Magnitude comparison task ........................................... 125
5.5 Magnitude difference comparison task ............................... 126
5.6 Fitted values of order on the magnitude comparison scale ........ 130
5.7 Confidence intervals for the fitted values of order on the magnitude comparison scale ........................................ 131
5.8 Fitted values of order on the difference comparison scale ........ 134
5.9 Scatter plot of the normalised sensitivity per observer for the first and second session .............................................. 135
5.10 Simulation of the $W$ statistic distribution ............................ 137
5.11 Scatter plot of estimated values of order derived from the discrimina
tion task and from the difference comparison task .................... 138
5.12 Mismatch between discrimination- and difference-based scales by focussing on end- and midpoints of the scales ...................... 139

6.1 Experimental design matrix for the pairwise comparisons .......... 147
6.2 Discrimination scales for experiments in Subgroup A ................ 151
6.3 Discrimination scales for experiments in Subgroup B ................ 151
6.4 Discrimination performance for experiments of Group 1 vs. aspect ratio .............................................................. 153
6.5 Discrimination scales for experiments in Group 2 .................... 156
6.6 Discrimination performance for all experiments vs. aspect ratio .. 157
6.7 Discrimination scales for experiments of high and low discrimina
tion performance .................................................................. 159
6.8 Discrimination performance vs. average experiment duration .... 160

7.1 Point patterns forming textures or images ............................ 163
List of Tables

3.1 Agreement rates for the psychophysical experiment 62
3.2 Performance of candidate g-scales 78
3.3 Preference rates for selected pairs of patterns of set A 79

5.1 Agreement rates for discrimination and difference tasks 127

6.1 Presentation parameters for Group 1 148
6.2 Agreement rates for Group 1 150
6.3 Presentation parameters for Group 2 154
6.4 Agreement rates for Group 2 155
**Abbreviations**

**APF:** After Pupa Formation

**a-scale:** absolute interval scale of order

**CDF:** Cumulative Distribution Function

**CI:** Confidence Interval

**CVT:** Centroidal Voronoi Tessellation

**fcc:** face-centred-cubic

**GoF:** Goodness-of-Fit

**g-scale:** geometric scale of order

**jnd:** just-noticeable difference

**LCD:** Liquid Crystal Display

**ML:** Maximum Likelihood

**MLDS:** Maximum Likelihood Difference Scaling

**PDI:** polydispersity index

**p-scale:** perceptual scale of order

**RMS:** Root Mean Square

**RMSE:** Root Mean Square Error

**std:** standard deviation

**2AFC:** two-alternative forced-choice
Chapter 1

Introduction

1.1 Motivation

Spatial patterns with regularity are found in living, natural non-living, and man-made systems at a wide variety of scales. Their presence can be informative about mechanisms of formation (e.g. Liesegang rings), properties (e.g. glasses), and functions (e.g. sensory receptors) (Figure 1.1). Despite the diversity of domains, because patterns are an abstraction of structure, generic methods for their analysis are possible.

Figure 1.1: Spatial patterns with regularity. Left: Geological formations (Liesegang rings) in Paria Canyon/Vermilion Cliffs, Northern Arizona (photograph taken by Gabe Farnsworth). Right top: SiO$_2$ having different optical properties depending on its structure, crystalline of high order, or amorphous, of less ordered structure (image from www.nde-ed.org). Right bottom: Scanning electron microscopy of the basal region of the mammalian cochlea showing the organised state of hair cell bundles (reproduced/adapted with permission from Kiernan et al. (2005)).
1.1. Motivation

There are many types of spatial patterns (e.g. grey-level textures, binary-valued images, polygonal meshes) but in this work we consider only arrangements of points on the plane (point patterns). Although only a fraction of pattern space, point patterns are widely applicable. They represent systems when the focus is on the relative location of the elements, and are commonly used in science for the analysis of biological tissue, material structure, foams, hard-sphere packing, ecological models, etc.

Point patterns can be analysed geometrically and examined by eye. Vision strongly engages with point patterns, evoking a complex bouquet of dependent qualities which seem to vary along several dimensions (e.g. density, degree of anisotropy, complexity, clustering, frequency of linear structures, etc.) not all of which correspond to agreed-upon geometric measures. This work is concerned with one such unmodelled dimension: the degree of order/disorder which we will refer to simply as order. The notion of order appears in Gestalt psychology (Koffka, 1935), and is related to arrangements of objects where their relative placement has a significance. Depending on the type of visual texture or pattern, order may be related to different regularities (laws) that govern this arrangement, e.g. translational symmetry in periodic structures, points forming pairs due to proximity in Glass patterns (Glass, 1969), self-similarity at a range of scales in fractals, etc. Consequently, depending on the analysed system, it appears at a variety of forms in scientific literature.

Order is particularly relevant to biology and biomedicine, because living systems tend to be well-, but not perfectly-, ordered, while the generation of disorder is associated with ageing and disease (Guillaud et al., 2004; Sudbø et al., 2000). In biological systems it is observed at all organisational scales. For example, at a fine scale, order is critical to the regulation of molecular machines (Goodsell and Olson, 2000), which often include juxtaposed modules of different symmetries (Nickell et al., 2009). At an intermediate scale, the cellular cytoskeleton can take on a more or less ordered state depending on its function (Jacinto and Baum, 2003), and in neoplasia has been seen to undergo a change from a well-organized state to a
convoluted mesh (Kellie, 1996). At a coarse scale, the formation of orderly skin patterns in zebra fish arises from a self-organising system of cell-cell interactions (Nakamasu et al., 2009).

Within biology, order is especially pertinent in development where disorder is continually being generated and corrected. Moreover, changes from disordered to ordered states are frequently seen (Cohen et al., 2010; Marinari et al., 2012; Classen et al., 2005; Cagan and Ready, 1989) and can be altered by mutation and perturbation of environmental conditions such as temperature and nutrition (French et al., 1998). Post-development, the final degree of order impacts on the effective functioning of the organ (Classen et al., 2005; Tilney and Saunders, 1983); and the fitness of the organism (Petrie et al., 1991; Dakin and Montgomerie, 2011), in some cases through effects on sexual attractiveness (Watson and Thornhill, 1994; Oliver et al., 2009; Chen et al., 2012) and mating success (Petrie et al., 1991). Not all developmental processes aspire to exact order. For example, in the mammalian eye, spacings of parafoveal receptors are less than perfectly regular which prevents Moiré-like aliasing (Wässle and Boycott, 1991). Post-development, the ideal level of order is maintained in healthy tissues by homeostatic processes that combine the disordering effects of cell proliferation with active processes to restore the resting level of intermediate order (Gibson et al., 2006)—processes that fail in diseases such as cancer (Guillaud et al., 2004; Csikász-Nagy et al., 2013).

1.2 Problem Statement and Approach

As in colour science, we often need to provide an agreed measure of a perceptible quality. As perceptual scales may exhibit individual differences, this agreement needs to be with the consensus. A common approach to eliminating the need for an observer is to devise a physical scale which is in good agreement with a perceptual one. When sufficient consensus exists, and good agreement with it is achieved, the perceptual scale may be abandoned and replaced with its objective surrogate. For example, the Mohs (Tabor, 1954) and Vickers (R L Smith, 1922) scales for material hardness are used without reference to tactile hardness.
1.2. Problem Statement and Approach

Finding objective proxies for the extremes of perceptual dimensions is often easier than for the middle range of the dimensions. For order, clear mathematical characterizations exist for patterns at the extremes: at the perfectly \emph{ordered} end, the Group Theoretical analysis of spatial isometries, results in a definitive characterization of exact order (Griffin, 2009; Kittel, 2004); at the \emph{disordered} end, statistical concepts of randomness allow perfect disorder to be uniquely characterized as arising from Poisson point processes (Illian et al., 2007). By contrast, there are only piecemeal mathematical theories concerning the middle range of intermediate levels of order (Cliffe and Goodwin, 2013; Dunleavy et al., 2012; Sausset and Levine, 2011; Truskett et al., 2000; Steinhardt et al., 1983). Yet, perceptually, the middle range seems to have well-defined structure: some patterns appear more ordered than others, suggesting an ordinal scale structure; further, some differences in order are larger than other differences (Morgan et al., 2012), suggesting an interval scale structure. So, our hypothesis is that the \emph{consensus of order has the structure of an interval scale}. In addition to testing this, the further aim is to define an objective interval scale closely aligned with this consensus. Importantly, such a scale would provide a wide range of natural scientists with a tool by which to quantify order within images to study the evolution of order/disorder in natural systems and in systems responding to experimental perturbation.

A specific example, which illustrates the complexities of order in a biological system, is provided by the notum (dorsal thorax), of \textit{Drosophila melanogaster} (the fruit fly). This tissue has long been used for the study of patterning during development and across evolution (Simpson et al., 1999). By 24 h after pupa formation (APF), the notum exhibits a well-spaced array of precursor mechanosensory bristle cells (\textit{microchaetes}), despite being relatively disordered during the proliferative phase 12 h earlier (Cohen et al., 2011). This rapid and pronounced change in order makes the tissue convenient for laboratory study which can be performed using live imaging techniques which do not interfere with the developmental process.

When notum development is imaged to reveal both cell shapes and cell state during this period of 12-24 h APF, two types of order become apparent (Figure 1.2).
First, epithelial cells form an approximate hexagonal lattice through cycles of division, shape change and neighbour exchange events. Second, a subset of these cells change their state (as measured by levels of proneural gene expression) to form a self-organizing (Cohen et al., 2011; Koto et al., 2011) patterned array of cells, which later give rise to well-spaced mechanosensory bristle organs.

Viewing movies of notum development (Figure 1.2), it is clear that the level of order increases in the period from 12 to 24 h APF. While convenient, such subjective assessments of order are imprecise, difficult to quantify, and are likely to be affected by the history of previous observations (Ouhnana et al., 2013). Additionally, irrelevant factors (e.g. element density, size, shape, orientation or brightness) may be difficult for observers to ignore—confounding the analysis.

As an alternative to subjective assessment, geometric algorithms have been proposed that aim to quantify order. For example, the variance of distances from each point to its nearest neighbour (Cohen et al., 2010). These approaches will be misleading unless they accurately correspond with subjective assessment across the full range of levels of order. Moreover, if a geometric algorithm is to replace subjective assessment, its agreement with interval scale structure, not just ordinal, needs to be demonstrated. We would be entitled to call a geometric assessment of
order that has been validated this way a measurement of order giving values on a well-defined interval scale. Using such a measure we would, for example, be able to quantify the changing order of the developing fly notum and plot its detailed timecourse using the interval scale as a meaningful ordinate axis.

1.3 Thesis Outline

This thesis is divided into seven chapters:

In Chapter 2 we present the framework for our approach for the quantification of perceptually-based order of point patterns (Section 2.2), we review previous work that is related to the perception of order in point patterns (Section 2.3), existing psychophysical methods that are suitable for our analysis (Section 2.4), and approaches from various scientific fields that aim to quantify order, or provide useful insights for novel measures (Section 2.5). Based on the examined literature we conclude that there is a variety of quantifications but none is validated as a measure of order.

In Chapter 3 we report experimental results and analyses that assess whether the perception of point patterns varies along a dimension of order with an interval scale structure (Section 3.2). The analysis tests for this by attempting to place test patterns into a perceptual interval scale ($p$-scale, for perceptual) so that their relative positions predict judgements of relative order made by multiple observers. Our results confirm the reality of the perceptual dimension and establish that it is quite well extended in terms of its number of distinguishable levels. We then proceed (Section 3.3) to establish a pragmatic interval scale of objective order, as assessed by a novel geometric algorithm ($g$-scale, for geometric), which accurately corresponds with the $p$-scale. We demonstrate that previously proposed quantifications of order correspond less accurately. Finally (Section 3.4), by scaling the $g$-scale, so that absolute disorder maps to 0 and perfect order to 10, we define a convenient absolute scale ($a$-scale) for the measurement of order.

In Chapter 4 we demonstrate the applicability of the constructed $a$-scale in the quantification of order for the developing fly notum. We do that for the two different
levels of organisation, (i) the bristle cell array refinement (Section 4.3), and (ii) the background epithelial cell packing (Section 4.4). Finally (Section 4.5), with the use of simulations of evolving point patterns we reveal inconsistencies between different geometric measures of order, so as to emphasize problematic interpretations of such quantifications when using measures that are not properly validated.

In Chapter 5 we examine whether sub- and supra-threshold judgements of order are consistent with a common interval scale. While we confirm the interval scale structure of the derived scales for each type of judgement separately, we reject the existence of a common interval scale that accounts for both. This chapter presents a neat example where the assumption of a common equal-variance signal detection model for both discrimination and difference comparison data is violated, and further suggests the combination of the two different scaling approaches that we employed, as a convenient method for rejecting such hypotheses.

In Chapter 6 we examine the effect of the presentation parameters (dot size and dot numbers) on the discrimination performance with respect to order for a specific class of point patterns (jittered square lattice of points) (Sections 6.4 and 6.7). This analysis led to the identification of the ratio of dot size to average inter-dot distance as the main parameter that affects the discrimination performance (Section 6.8).

Finally, in Chapter 7, we summarise our conclusions, present the limitations of our work, and suggest directions for further research.

1.4 Contributions

In this work, we have made five main contributions:

1. We have validated a high inter-observer agreement in judgements of order for a diverse set of point patterns.

2. We have validated that the consensus of judgements of order is consistent with an interval scale structure.
3. We have constructed an absolute interval scale for the quantification of order, based on a geometric algorithm, that predicts order with an accuracy of 0.5 jnds (just-noticeable differences), and have demonstrated its applicability by analysing the progress of organisation of the developing fly notum.

4. We have rejected a common perceptual scale for both sub-threshold and supra-threshold judgements of order.

5. We have identified, for a common class of point patterns, the ratio of dot size to average inter-dot distance as the main parameter that affects discrimination performance.

1.5 Publications

Work done during this PhD contributed to the following publications:


Parts of mentioned publications have been reused in current thesis. My contribution in publications 1 and 3 concerns parts of the methods and is not contained in the thesis. An additional article has been submitted for publication to the *Journal of Vision*:

- Protonotarios, E. D., Johnston A. & Griffin, L. D. Difference magnitude is not measured by discrimination steps for order of point patterns. (revision submitted in January 2016, minor revision invited in February 2016)

Another article concerning the presentation parameters that affect discrimination performance of order in point patterns (Chapter 6), is under preparation (September 2015).

I have presented parts of the work described in the thesis at the following workshops:

- The Institute for the Physics of Living Systems, LMCB & MRC Lab, UCL, March 2015

- Visual Statistics in Humans and Machines Symposium, City University London, January 2015, and

- Theory in Biology Series, MRC Lab, UCL, June 2013.
1.5. Publications

Selected parts have been presented by Lewis D. Griffin at:

- Quantitative Biology Conference, National Institute for Medical Research, The Francis Crick Institute, London, June 2014,

and by Professor Buzz Baum at:

- Santa Barbara Advanced School of Quantitative Biology 2013, Summer Research Course “New Approaches to Morphogenesis: Live Imaging and Quantitative Modelling”, July 22 - August 24, 2013, and

Chapter 2

Background

2.1 Introduction
In this chapter we present the framework for our approach for the quantification of perceptually-based order of point patterns (Section 2.2), we review previous work that is related to the perception of order in point patterns (Section 2.3), existing psychophysical methods that are suitable for our analysis (Section 2.4), and approaches from various scientific fields that aim to quantify order, or provide useful insights for novel measures (Section 2.5). Based on the examined literature we conclude that there is a variety of quantifications but none is validated as a measure of order.

2.2 Measurement Theory
Measurement theory provides the framework for a solid justification of our approach as an objective way to quantify perceptually-based order. Two principal aspects can be recognised in any measurement process: the representational and the pragmatic (Townsend and Ashby, 1984; Hand, 1996, 2004). These are presented in the following sections.

2.2.1 The Representational Aspect
According to the representational aspect, measurement is defined fundamentally as a process of assigning numbers to objects in a way that essential qualitative relations among these objects (in terms of the attribute in question) are reflected in the
numbers and in the properties of the number system used. These relations are validated by experience. Basic empirical relations are: the determination of equality and rank ordering; and these can be with respect to raw attributes or their differences or ratios. (Stevens, 1951; Townsend and Ashby, 1984; Narens and Mausfeld, 1992; Hand, 2004).

An example of such measurement process is the assignment of specific numbers to objects to represent their length. This representation can be simple (direct measurement), like the assignment of length, or a more complicated representation (indirect measurement) as in the case of an overall score resulting from a complex process, e.g. the percentage of correct answers to a test. As our observations of the empirical relationships between the objects are based on a finite sample of objects, representational measurement necessarily involves inductive reasoning (Hand, 2004; Stevens, 1951).

Based on the attribute under study we can define (mutually exclusive and exhaustive) equivalence classes for the objects, with two objects belonging in the same class when, in terms of this specific attribute, they are indistinguishable. For example, for the length of rigid rods, two objects belong in the same class when, if put side-by-side with one edge coinciding, their other edges also coincide. For this comparison process, given that for two rods \(a\) and \(b\) the other edges do not always coincide, we can define the ‘\(a\) is longer than \(b\)’ relationship, and require for the numbers \(M(a)\) and \(M(b)\) assigned to the objects \(a\) and \(b\) that \(M(a) > M(b)\). In the particular case of rigid rod lengths, the act of concatenation (symbolised as ‘\(\circ\)’) is possible, and further empirical relationships can be validated. When two rods, \(a\) and \(b\), are put end-to-end in a straight line, a third one, \(c\), can be found that is indistinguishable from their concatenation \((a \circ b = c)\). Length assignment must be consistent with this empirical relationship (e.g. \(M(a \circ b) = M(a) + M(b)\)). The act of concatenation combined with a consistent numerical assignment, allows one particular object to serve as unit length. Partitioning of the unit length to equal segments, allows comparisons with objects of fractional lengths. Therefore, within the limited accuracy and reliability of the measurement instruments and observations,
we can establish an isomorphic mapping between the equivalence classes of the objects with their relations, and the system of numbers being used to represent their measures (Hand, 2004; Narens and Mausfeld, 1992).

An important property of this mapping is that, in general, it is not unique. This means that depending on the set of axioms that define the allowed relationships within the appropriate system of numbers, there can be more than one valid such assignments. In the case of length, which was previously mentioned, an arbitrary rescaling of the numbers, for example, results in an equally valid numerical assignment. The invariance of the measurement model under such permissible transformations, which preserve these relationships, is crucial in the distinction between different types of scale (Hand, 2004).

The main types of scale (Stevens, 1946) are the following:

- **Nominal Scale**: The numbers are used only as labels. This assignment is the least restricted.

- **Ordinal Scale**: Numbers express rank ordering. The scale remains invariant under any order-preserving transformation.

- **Interval Scale**: In this scale three of the fundamental relations can be represented (equality, rank-order, equality of intervals). Almost all usual statistical measures can be applied on this type of scale. The scale remains invariant under a linear transformation and the zero point is decided in terms of convention or convenience. While ratios between scale numbers are not meaningful, ratios of their differences are valid; for example, one difference can be twice another.

- **Ratio Scale**: Ratio scale is the most complete and all four fundamental relations can be represented (those of the interval scale plus the equality of ratios). This scale contains an absolute zero value and remains invariant under multiplication with a constant.

### 2.2.2 The Pragmatic Aspect

In the pragmatic aspect of measurement, the procedure of measurement itself defines the variable that is being measured. For example, the weight of an object is
2.2. Measurement Theory

defined as the number of unit weights that are required in order to balance the object on a scale. For pragmatic measurement, understanding the underlying relationships between the objects is not necessary, but the capacity for prediction is important. Therefore the procedure of measurement in this sense has to be well defined and the resulting numbers should have a practical value. It is possible that the same underlying attribute can be measured with different procedures, but if there is arbitrariness on the actual definition, problems can arise and the conclusions derived through different definitions will not be the same. Although precise definitions of scientific concepts are often taken for granted, the journey to that definitions has not always been straightforward. This is true not only for concepts appearing in social or behavioural sciences but also for concepts in physics (e.g. mass) or mathematics (e.g. probability) (Hand, 2004).

A pragmatic scale can be in direct correspondence to the representational scale as for example in the Mohs system of quantifying material hardness - ten levels are specified by ten named substances: 1 is the hardness of Talcum, 2 of Gypsum, etc. These Mohs values are positions on a representational scale and their numeric relations mirror empirical relations: materials with larger Mohs values can scratch materials with lower values but not vice versa (Tabor, 1954). Vicker's hardness scale on the other hand is defined without reference to the empirical relationships between the objects and measures material hardness by the force applied through a diamond tipped indenter, divided by the area of the resulting indentation in the measured material (Dugdale, 1958). The Mohs and Vickers scales are reasonably, though not perfectly, in agreement.

2.2.3 Quantification of Order as a Measurement Process

In the framework of measurement theory as presented above, the quantification of perceptually-based order as it is performed in the current study contains both the representational and the pragmatic aspects of measurement.

In the representational aspect, the first stage is selection of objects, in this case point patterns, which are representative samples of the domain of interest. As a
result we will be able to generalise our conclusions with relative confidence. The objects must of course contain the attribute of apparent order. The selection of objects can be done a priori according to a subjective method, or a posteriori by the analysis of a wider set of samples which could exceed the limits of the space of interest, with the aim of defining these limits. The second stage is the analysis of the empirical properties of these objects, with respect to the attribute in question, and identification of the appropriate numerical system to represent these relations. The collection through a psychophysical experiment of subjective judgements of order from a number of observers, is equivalent to an indirect representational measurement process. As such, specific numerical values can be assigned to each object, whilst also the structure of the appropriate numerical system is derived.

Our approach for the identification of the empirical relationships between point patterns with respect to order uses the human visual system as a measuring device and employs established psychophysical methods for the collection of indirect measurements of the attribute. Having in mind that judgements based on the visual system and the collected responses are subject to inherent noise we aim to model these relationships with explicit expression of the noise.

The pragmatic definition of order takes the form of an algorithmic procedure based on the available information present in the point pattern (e.g. x-y coordinates of points, lengths of polygonal edges, etc.). The aim is to develop an algorithmic definition with practical predictive power.

The properties of such a measure will necessarily be constrained by the properties of the appropriate type of scale. Its usefulness is linked, however, to the actual questions that need to be answered. For example, if we need to reliably discriminate between two instances of patterns, and the type of scale is interval scale, then the accuracy of the measure has to be significantly smaller than the difference in order of the patterns that are being examined. On the other hand, if there are systematic errors in this measurement process over a wide range and, we are interested in studying differences or slopes of curves of order, we have to make sure that these systematic errors do not distort in a crucial way the interval scale.
Our attempt to quantify order with a geometrical measure can be seen as an exhaustive process of disproving established or novel measures for a variety of patterns, where these measures should be expected to work reliably. This disproving can be done in terms of the agreement in rankings or in terms of the accuracy in predicting theoretical numerical values. In this sense, one can never be absolutely confident about the validity of the best performing measure, but at least one can expect it to perform better in the widest possible sense, yet still within the field of interest.

### 2.3 Perception of Order

As mentioned in Chapter 1, the notion of order appears in Gestalt psychology (Koffka, 1935), and is related to arrangements of objects where their relative placement has a significance. This is contrasted with arrangements which are the result of ‘blind mechanical forces’ lacking a purpose or not serving a function, something that we could call randomness. As such, order may consist of specific regularities (laws) that govern this arrangement, and these may interact synergistically (Wagemans et al., 2005). With order we mean the global ‘gestalt’ of all of these. In many types of arrangements, the specific regularities can be decomposed into basic principles but this is not always the case as an arrangement can appear ordered even if one cannot identify its specific regularities. Regular patterns, exhibiting translational and/or rotational symmetry appear in ancient cultures and have attracted human interest long before these principles were properly understood and formalised (Figure 2.1).

Gestalt psychology identified a number of important principles of perceptual organisation. e.g. the principle of **proximity**, **mirror symmetry**, or **good continuation**. It was suggested that the visual system seeks order (e.g. symmetry) so as to make sense of sensory signals. Attneave (1954) proposed that the visual system compresses information to reduce redundancy. Patterns with regularities suggest simple mechanisms for such efficient encoding, and thus are preferred. Gestalt principles considered under this view, are functions of the perceptual machinery
2.3. Perception of Order

that group information together and thus provide an economical description of visual reality.

The exact visual and cognitive mechanisms that analyse order are not well understood. Symmetries appear in a variety of visual stimuli and their effect is powerful on our perception. For instance, symmetry contributes significantly to the attractiveness of a face (Jones et al., 2007). For the analysis of order-related perceptual mechanisms, simple repetitive patterns have been examined. A considerable amount of work has been conducted on the detection of symmetry in dot patterns (Wagemans et al., 1991; Wagemans, 1995). The aim of this work was mainly to study symmetry detection in different configurations (skewing of symmetry using affine transformations, change of orientation) in time controlled experiments. The goodness of regularity has been studied as well in dot patterns; but only with stimuli restricted to patterns with mirror symmetries (Nucci and Wagemans, 2007).

According to the proximity principle nearby objects are grouped together. In rectangular arrangements of dots, when for example, the horizontal spacing is even slightly larger than the vertical, dots group into vertical columns (Sekuler and Blake, 2005). Kubovy and Wagemans (1995); Kubovy et al. (1998) modelled the group-
2.3. Perception of Order

ing of this type for a range of lattice conditions as a competition of attractive forces between dots, which decay exponentially with distance. The proximity grouping effect is particularly strong in Glass patterns (Glass, 1969), where the pairs of closely placed dots create a global percept. Dakin (1997) noticing that Glass patterns contain orientation structure at a narrow range of scales examined how filter size is adaptively selected for texture processing. He proposed a size criterion that minimises the local variation of feature orientation.

Newell et al. (2013) studied static and dynamic patterns to investigate the role of the Gestalt perceptual principles on aesthetic preference and on perception of ‘naturalness’. The stimuli were either static or moving point patterns. Points were displayed as small black squares, triangles, or filled disks on white background. For the static patterns, the degree of order was manipulated by varying the amount of positional jitter of the elements, and for the dynamic patterns by varying the coherence of motion of nearby (local) or all (global) elements. For the rating a Likert scale (1 to 7) was used. To quantify ‘randomness’ for the static patterns, they used the ratio $\sigma(l)/\langle l \rangle$, where $l$ for an element in the pattern is the distance to its nearest neighbour, $\sigma(l)$ its standard deviation, and $\langle l \rangle$, its mean value for all elements in a pattern. They explained the preference of the observers to the most ‘ordered’ patterns as an effect of the Gestalt principles of *proximity* and *good continuation*. Similarly for the dynamic patterns, they attributed the higher preference for coherent motion to the *common fate* principle. The patterns were based on a square lattice of points which was aligned with the square borders of the pattern; therefore in the perfectly ordered case the elements and the border were parallel. In their study they did not examine this effect of the border on the judgements. Similar studies on aesthetics would benefit from validated measures of order.

Morgan et al. (2012) studied perception of regularity for perturbed dot patterns. They investigated the effect of sensory noise on regularity discrimination and on reliable detection of regular patterns. The patterns they used consisted of white Gaussian blobs on a grey background. Elements were arranged in an $11 \times 11$ lattice, but different configurations were also used, with the elements placed around a no-
tional circle or along a notional line. Different degrees of regularity were achieved by independent jittering of the elements on the \(x\) and \(y\) dimensions, with displacements sampled from a uniform distribution of varying range. Each pattern was presented for a brief duration (200 ms). Their results were consistent with an undersampling model, according to which estimates of regularity are based on the variance computation of displacements of a small number of elements (most commonly, five or six). One interesting finding was that discrimination is facilitated when both of the patterns are not perfectly regular (‘dipper’ effect). According to this model, sensory noise is not represented in awareness, and so observers are able to discriminate patterns of different levels of regularity, while, as long as the noise is below a threshold, both appear regular to them. Another interesting result was the effect of task-irrelevant variation (e.g. contrast) on the discrimination thresholds. This was equivalent to an increase in sensory noise. Their approach supposes an ideal observer who estimates deviation from the expected position of the elements making use of an obvious reference notional pattern. It is not certain how accurate this assumption is when the pattern is highly irregular or when there is no obvious reference pattern (e.g. when order is not based simply on translational regularity).

Ouhnana et al. (2013) demonstrated that regularity is an adaptable visual dimension. They used patterns consisting of either luminance– (Gaussian blobs), or contrast–defined (difference of Gaussians and random binary patterns) elements arranged on square grids of \(7 \times 7\) elements presented in a circular aperture. Again different degrees of regularity were achieved by independent jittering of the elements on the \(x\) and \(y\) dimensions, with displacements sampled from a uniform distribution of varying range. They found that adaptation was independent of the type of elements and its effect was unidirectional, i.e. it only caused test patterns to appear less regular. They argued that, given the unidirectionality of the effect, regularity is not coded according to the conventional model of multiple narrow tuned channels (at specific regularity levels). Since adaptors with both greater or smaller regularity than the test pattern reduced the apparent regularity of the latter, they supported that coding of regularity is norm-based (Webster, 2011), with irregularity
being the norm. Their model assumes processing of the pattern by a filter-rectify-filter cascade (Graham, 2011) with regularity being encoded via the peakedness of the distribution of the energy responses across receptive field size; the peak corresponds to the spatial frequency of the regular pattern. Adaptation occurs because the adaptor flattens the response distribution for the test pattern and thus the associated peakedness is less prominent. This study presents a promising model for the encoding of regularity. There are interesting questions related to the presented filter-rectify-filter approach. First, patterns appear regular even if they do not exhibit an exact translational symmetry, for instance when the inter-element spacing increases gradually. In this case the peakedness of the distribution can be significantly affected. A mild jittering on the position of the elements of a perfect square grid may not affect the peakedness to the same degree, but still the jittered pattern will look more irregular than the one with the smooth change in inter-element spacing. Second, in the presented analysis, only one orientation has been considered. It is clear though that hexagonal arrangements appear regular, and the same is true for arrangements where the lines connecting neighbouring points do not hold the same orientation across the whole extent of the pattern. If this model is extended to include, at the second filter stage, pooling for different orientations then this would imply that even for the square lattice arrangements the oblique orientations should be included, and this would cause severe smoothing of the distribution. Therefore, if the proposed model is correct, there should be an additional mechanism responsible for adjusting the spatial extent of the filters in combination with the inter-element spacing, similar to that proposed by Dakin (1997). Interestingly, this model may be able to explain why point patterns with regular inter-point spacing do not appear very regular if the nearby orientations are not aligned. Figure 2.2 shows such an example. The point pattern on the left appears less regular than the middle one and the middle one appears less regular than the polygonal array on the right hand side. The polygonal structure is based on exactly the same point locations with specific connections between neighbouring points being highlighted. Thus a high order
structure is revealed which was not directly perceived before these connections were shown.

![Figure 2.2: Perception of order in point and polygonal patterns. The point pattern on the left appears less ordered than the one in the middle, and the one in the middle less ordered than the polygonal array on the right. However, the polygonal structure is based on the points of the middle pattern. The edges reveal a hidden law which although not regular in a translational way, when established transforms the perceptual impression of order (modified from Ossi (2003)).](image)

Yamada et al. (2013), contrary to Ouhnana et al. (2013), found that adaptation to regularity is bidirectional, which they called the randomness aftereffect. They used a similar experimental methodology to Ouhnana et al. (2013) (an array of 16 × 16, instead of 7 × 7, and solid dots instead of the variety of elements used in Ouhnana et al. (2013)) and excluded as well local position adaptation by global jittering of the adaptor pattern. They assumed a filter-rectify-filter model, but of a slightly different implementation. They argued that Ouhnana et al. (2013) did not check the sign of the difference at the peak intensity of the distribution of spatial frequency channels response between the adaptor and the test pattern, which would explain the randomness aftereffect. They did not provide an explanation for the discrepancy between the two studies other than speculating that the smaller number of elements in the earlier study (49 instead of 256) may not have been sufficient to yield a strong difference in the peak response. We notice however, that the two studies used patterns with significantly different ‘element size to inter-element distance’ ratio. Also, none of the two studies examined the effect of very disordered adaptors.

Influenced by the work of Morgan et al. (2012), Ahmed et al. (2013), instead of positional regularity, studied perception of orientation regularity. They used ar-
rays of Gabor patches and varied the standard deviation of their orientation to control regularity. They found bidirectional adaptation aftereffects, i.e. adaptation to regular patterns caused less regular test patterns to appear even less regular, while adaptation to irregular patterns, caused more regular patterns to appear even more regular. Later (Ahmed et al., 2014), to exclude that contrast adaptation was causing the observed regularity adaptation, examined whether different amounts of positional randomisation of the Gabor patches, and so different amounts of overlapping between the adaptor and the test patterns, affected the adaptation aftereffect. They reported that adaptation did not depend on the spatial overlap, and thus concluded that orientation adaptation could not solely be accounted for by contrast adaptation.

The most relevant study to our work was conducted by Preiss and Vickers (2005). The researchers collected data on judgements of regularity of point patterns and examined whether a specific geometrical measure, the density-normalised mean nearest neighbour distance, correlated with them. The points were arranged on a square area, along a line or around the circumference of a circle, and were presented as blue hollow circles on white background. Patterns varied from regular to clustered, and totally random point patterns (Poisson) were also included. Participants rated the degree to which point patterns appeared clustered, random or regular, using an eleven point scale. It was found that ratings correlated well with the geometric measure but the relationship was not linear. There are well known biases associated with the method of direct rating that was employed (e.g. category effect, principle of consistency, stimulus-equalizing bias) (Gescheider, 1988). Also, rating scales do not allow testing of the structure of the collected data and are not appropriate for modelling perceptual noise. The small hollow circles very often overlapped and their intersections created junctions. Since the number of overlapping circles depends on the density and the type of pattern used, it is possible that observers used this as a cue. This measure was examined in our experiments (Section 3.3.2.1) and did not perform well (measure 1 in Table 3.2).
2.4 Psychophysics

Psychophysics originated with Fechner in 1860, when he introduced techniques for measuring sensations as an approach to relate the physical with the mental (Gescheider, 1997). In psychophysics “sensations themselves are attributes of the organism, comparable in many ways to the properties of physical objects. As such, these attributes are assumed to be highly stable and regular, and thus they can be subjected to careful analytical study and represented numerically in a manner analogous to physical measurement” (Luce and Krumhansl, 1988).

In order to quantify the perception of apparent order, we need to measure the sensation that is associated with a pre-defined set of stimuli (point patterns). Rating scales can be used to collect data on the perceptual attribute. The scale of allowed responses could be specified by a series of reference patterns, numbers, or linguistic specifications (e.g. ‘slightly ordered’). While rating has the advantage of directly estimating the value of the attribute in the stimulus, it has several disadvantages: there may be dependency on the reference patterns; there may be different interpretations of linguistic specifications; and the spacings of the items of the response scale are unknown. Apart from the well known biases of the direct-measurement techniques (Gescheider, 1988), it became clear in psychophysics that it is fundamentally impossible to measure sensation directly by any method, and this must be indirectly inferred from observers’ responses to sets of stimuli during specific tasks (Gescheider, 1988; McKenna, 1985).

A commonly-used indirect approach is to have observers rank sets of stimuli. The limiting cases of this are ranking the entire set, and pairwise ranking. Pairwise ranking is preferred as it does not make the assumption that judgements of the attribute will necessarily have a transitive structure, and instead allows this to be tested for in the collected data. Additionally, possible interactions in the perception of individual stimuli are minimised in number in pairwise comparison (David, 1988). The main disadvantage of the pairwise comparisons method is that they require a large number of total judgements if it is necessary to compare all pairs. Other options of comparing or estimating numerical relations of differences or ratios be-
between patterns presuppose that the scale of measurement has specific properties, something which is not known before the analysis of the data. After establishing that the perceptual dimension has sufficient structure, a comparison of differences or ratios would be a reasonable direction for further research.

In order to infer the magnitude of sensation from responses, through this procedure of indirect measurement, a psychophysical theory is necessary. This theory can be tested by examining its agreement with other theories of sensory processes and the experimental data.

2.4.1 Psychophysical Scaling Methods

“Psychophysical scaling methods are procedures for constructing scales for the measurement of psychological attributes.” (Torgerson, 1958). Two important linear scaling models for paired comparison data exist, Thurstone scaling and the Bradley-Terry models.

2.4.1.1 Thurstone Scaling Model

Thurstone was the first to develop methods for measuring sensory intensity when stimuli are not characterised by physical properties (Gescheider, 1997). Especially for our analysis, this is very convenient, since we do not know in advance how the interaction of different regularities determines the final degree of order. In Thurstone’s Law of Comparative Judgement (Thurstone, 1927) he proposed a scaling model for converting subjective comparative preferences with respect to a specific attribute into a uni-dimensional interval scale. This method has been widely used in diverse scientific areas, where an objective quantification method is required.

According to the Thurstone’s model, the attribute that is being measured is a normally-distributed random variable. When an observer estimates the amount of the specific attribute present in an object \( A \), this is equivalent to a random experiment, where the estimate \( \psi_A \) is a noisy realisation of the true scale value. Thurstone assumed a normal distribution for the noisy realisation, with the mean value \( M_A \) being the true scale value. Therefore, when a comparison between two objects \( A \) and
2.4. Psychophysics

When decision-making occurs, the decision about which of the two contains the specific attribute at a greater degree, will depend on the comparison of the two realisations, \( \psi_A \) and \( \psi_B \), or equivalently on the sign of their difference, \( \psi_A - \psi_B \). This difference is a normal random variable itself, with mean value \( M_{AB} = M_A - M_B \) and standard deviation \( \sigma_{AB} \) which depends on the standard deviations \( \sigma_A \) and \( \sigma_B \) of the two distributions and their correlation \( \rho_{AB} \) according to the following formula (Gescheider, 1997):

\[
\sigma_{AB}^2 = \sigma_A^2 + \sigma_B^2 - 2 \rho_{AB} \sigma_A \sigma_B \tag{2.1}
\]

According to this model, the probabilistic character of such judgements takes the form of a random experiment and the perceptual noise of the observer is explicitly described. Furthermore, the outcome can be predicted by the following integration:

\[
P(A > B) = \int_{-M_{AB}}^{\infty} \frac{1}{\sqrt{2\pi} \sigma_{AB}^2} e^{-\frac{x^2}{2\sigma_{AB}^2}} dx = \Phi\left( \frac{M_{AB}}{\sigma_{AB}} \right) \tag{2.2}
\]

Depending on the level of simplicity that can be assumed for this model, Thurstone distinguished five cases. In Case I a single judge is assumed, while in Case II it is assumed that, even with multiple judges, the distribution of the apparent attribute remains normal. In Case I and II, if the correlation coefficient depends on each pair of objects, the scale values cannot be determined, as each equation of comparisons adds one extra unknown. Therefore, the correlation coefficient is considered constant and then for a number of objects greater to five the problem is solvable when all possible pairs of objects are judged. In Case III, it is assumed that the degree of perceived quality on one object does not affect the perceived quality on the object of the same pair, so that the correlation can be considered zero. In Case IV, the standard deviation of the Gaussian distributions is assumed to vary in a linear way along the scale and in Case V, which is the simplest, a common standard deviation is assumed for all distributions. Since the scale is not changed
if multiplied by a constant, this common standard deviation can be used as the unit of the scale. Quality difference, $\hat{M}_{AB}$, in Case V can be directly estimated by the proportions of preference, $f_{AB}$, for objects A over B (Thurstone, 1927; Tsukida and Gupta, 2011) according to:

$$\hat{M}_{AB} = \Phi^{-1}(f_{AB})$$  \hspace{1cm} (2.3)

Figure 2.3 shows diagrammatically the concept of perceptual comparison with respect to the magnitude of a specific attribute for a pair of stimuli according to the Thurstone Case V model.

**Figure 2.3:** Diagrammatic presentation of the Thurstone Case V model. Bottom: Distributions of the noisy realisations of a perceptual attribute for two stimuli with real attribute values $M_A = 0$ and $M_B = 1$. The noise is Gaussian and on this scale the standard deviation is equal to 1. Top: The resulting preference function and the probability of preferring stimulus B over A when the difference between the real values is equal to 1.
2.4.1.2 Bradley-Terry Model

According to the Bradley-Terry model (Bradley and Terry, 1952) the attributes (qualities) \( \pi_A \) and \( \pi_B \), assigned to objects \( A \) and \( B \) are related to the probability of \( A \) being chosen over \( B \) with this simple formula:

\[
P(A > B) = \frac{\pi_A}{\pi_A + \pi_B}
\]  

(2.4)

With the transformation \( \pi_i = \exp(M_i/s) \), where \( s \) is a scale parameter \( (i = A, B) \), the probability that \( A \) will be chosen over \( B \) is expressed as the cumulative distribution function of the logistic random variable \( A - B \) with mean value \( M_A - M_B \):

\[
P(A - B > 0) = \frac{\exp(M_A/s)}{\exp(M_A/s) + \exp(M_B/s)} = \frac{1}{\exp(-\frac{M_A - M_B}{s}) + 1}
\]  

(2.5)

While in the case of the Thurstone model, the random variables \( A, B \) and their difference are all Gaussian, here, as Holman and Marley proved (in Luce and Suppes (1965)), \( A - B \) can be a logistic variable if \( A \) and \( B \) obey the Gumbel distribution. The logistic curve is very similar to the cumulative Gaussian but has slightly longer tails which means that proportions near 0 or 1 will result in larger scale differences in comparison to the Thurstone Model, while the opposite will happen for proportions near 0.5. (Tsukida and Gupta, 2011)

2.4.2 Model Fitting

The equations presented in the previous sections for both the Thurstone and the Bradley-Terry models provide exact solutions when only one pair is judged. However, this is seldom the case, as usually the scale values of a larger set of objects need to be estimated in a single experiment. The scale differences derived from each pair, may lead to inconsistencies because of the probabilistic nature of the
preferences. Mosteller (1951) provided a simple closed form solution based on a least squares optimisation, minimising the square error between the differences of the quality scores and the estimates from the Thurstone’s Law for the pairs. This solution suffers though from the fact that the inverse sigmoid function $\Phi^{-1}$ (probit or logit) returns $\pm \infty$ when the preferences are 0 or 1. The ways to circumvent this problem is either to exclude these values, a method which has the drawback of removing important information, or to fix these proportions by adding a small number, which alters the experimental matrix. Another option is to use Maximum Likelihood estimation for the score values. If we consider constant the chance of choosing one object over the other, then for each pair the probability for the frequency of a specific outcome is determined by a binomial distribution. For the whole experiment the log likelihood of the overall outcome can be defined as:

$$L(M|f) = \sum_{A,B} f_{AB} \log(\Phi(M_A - M_B)) + \text{constant}$$

(2.6)

with the constraint of zero mean for the score values. This is a convex optimisation problem with a unique solution (Tsukida and Gupta, 2011).

### 2.4.3 Linear Models

The Thurstone and Bradley-Terry models are special cases of a general group of linear models. In linear models the preference probabilities can be expressed as a symmetrical cumulative distribution (sigmoid) function of the differences of the assigned scores for each object. When this can be done, then the scores can be represented as points along a line, with their relative positions fixed (David, 1988). So, in theory, there is a wide range of possible symmetric CDFs to choose from, not all of which will represent an acceptable model. If, for example, longer tails seem to fit better the preferences for objects placed in longer distances, then a Laplacian CDF might give a better fit than a Gaussian CDF. However, some CDFs do not correspond to variables that can be expressed in a simple way as the difference of
two random variables of a reasonable distribution. The Laplacian, specifically, can be derived as the difference of two single side exponential distributions, a description which does not provide a meaningful interpretation of an underlying process. On the contrary, the realisations of the objects’ qualities during a judgement in the Thurstone model follow the normal distribution, and in the Bradley-Terry model the Gumbel distribution, both of which can have physical interpretation; the first as the mean of a large number of signals transmitted to the brain, and the second as the maximum value of a number of signals of a range of distributions, e.g. Laplacian. (Tsukida and Gupta, 2011; Luce, 1994; David, 1988).

### 2.4.4 Correction of a Model for Observers’ Lapses

During psychophysical experiments it is common that stimulus independent errors can appear. That means that even when the difference in the quality between two stimuli is large, there is a finite but small probability that the participant would register by mistake or lack of concentration the opposite choice. Maximum Likelihood estimation proves to be very sensitive in this type of error (Swanson and Birch, 1992). Avoiding the area of high performance (where the sigmoid function approaches 0 or 1) is a par solution to the problem as this area provides significant contribution to the estimation (Wichmann and Hill, 2001a). The phenomenon of lapsing can be incorporated in the model by rescaling the model function from \([0, 1]\) to \([\lambda, 1 - \lambda]\), where \(\lambda\) is the lapse parameter. With this adjustment these types of error do not strongly affect the overall fitting. \(\lambda\) does not always represent accurately the real lapsing rate, however, its use helps to avoid the afore-mentioned bias (Wichmann and Hill, 2001a).

### 2.5 Measures of Order

Analysis of a point pattern can be based directly on point locations, or on a geometric construction derived from the locations. In particular, triangulations or polygonal meshes induced by a point pattern are often used. The two most common, the
Voronoi diagram (Voronoi, 1908), and the related Delaunay triangulation (Delaunay, 1934), are described in the following paragraph.

**Voronoi Diagram and Delaunay Triangulation.** In Figure 2.4 we illustrate the Voronoi diagram and the related Delaunay triangulation for example data from the fly notum. The Voronoi diagram is the partitioning of the plane into convex polygons (cells) such that each polygon contains all locations closer to one point of the pattern, the *generator*, than to any other points. Voronoi cells are bounded by sides where two cells abut. Voronoi sides meet at junctions of typically three sides. The Delaunay triangulation of a point pattern is the unique triangulation such that no triangle contains any points of the pattern within its circumcircle. The Voronoi diagram and Delaunay triangulation are dual: Delaunay triangle edges are one-to-one with Voronoi cell sides; and Delaunay triangle circumcentres are coincident with Voronoi side junctions.

A plethora of systems are directly represented by polygonal tessellations (e.g. epithelial tissue, plant epidermis, fracture patterns on surfaces, etc.). Methods developed for their analysis can also be used for point patterns since these can be applied on the induced Voronoi and Delaunay tessellations.

In the following, we review methods for the quantification of order used in Biology (Section 2.5.1) and methods that appear in different scientific fields (Section 2.5.2). Most of these are based on point locations, triangulations and polygonal meshes. Some approaches cannot be applied directly to point patterns, but can inspire ideas for the design of novel measures.

### 2.5.1 Measures of Order Used in Biology

For the analysis of biological tissue, measures have been proposed that depend on distances, areas, shapes and topology. Most often the patterns are analysed statistically, and so the mean value, the variance, or the whole histogram of a chosen quantity is examined.

Distance measures may take into account the distances between all pairs of elements in a pattern or only the distances between nearest neighbours. Nearest
2.5. Measures of Order

![Figure 2.4: Demonstration of the Voronoi diagram and the Delaunay triangulation based on a point pattern. (a) Drosophila fly notum at 22h APF, (b) the point pattern derived from the image, (c) the Voronoi Diagram, and (d) the Delaunay Triangulation based on the point pattern.](image)

Neighbours can be those within a specified radius of an element, or the nearest elements up to a specific number (considering the single nearest neighbour is the most common). Nearest neighbours when the measure is applied on a polygonal mesh, can be defined as adjacent polygons. For a point pattern, these can be defined through a geometric construction (e.g. directly connected points in the Delaunay triangulation).

Mean nearest-neighbour distances are commonly used. In Raymond and Barthel (2004), and Allison et al. (2010) global mean nearest-neighbour distances are used in the analysis of the patterning of the planar mosaic array of cone photoreceptor spectral subtypes in the zebrafish retina. The global variance of nearest-neighbour distances is used as well, for example in Cohen et al. (2010), where...
2.5. Measures of Order

the spacing of the precursor bristle cells on the *Drosophila* dorsal thorax is analysed. Other measures use the *conformity ratio* which is the ratio of the mean nearest-neighbour distance to its standard deviation (Cook, 1996; Allison et al., 2010; Galli-Resta and Novelli, 2000). In some studies different quantities based on nearest-neighbour distances (mean, variance, etc.) are combined in a non-linear way to form a single value order estimate (Kayser et al., 2007).

Other distance based approaches that use all, rather than nearest-neighbour, distances are the Ripley’s L-function (Ripley, 1977; Shapiro et al., 1985) (see Section 2.5.2.13) and auto-correlation analysis (Galli-Resta and Novelli, 2000; Cook and Podugolnikova, 2001; Raven and Reese, 2003) (see Section 2.5.2.14).

Farhadifar et al. (2007) studying cell-cell interactions on proliferating *Drosophila* wing epithelium use the average area of the cells and its variance to quantify cell packing. Measures of area disorder are used as well in Raven and Reese (2003) to study the mosaic regularity of horizontal cells in the mouse retina. In Sudbø et al. (2000) a list of order/disorder measures is examined with the aim to detect statistically significant differences between normal mucosa and carcinomas. The variance of the Voronoi tessellation polygons area, the variance of the Delaunay triangles’ edges, and the variance of the number of neighbours contained within a specified radius are included in the list.

Shape-based approaches for the analysis of the retina include the distribution of Voronoi cell angles (Shapiro et al., 1985) and Delaunay triangle angles (Galli-Resta and Novelli, 2000). Farhadifar et al. (2007) and Patel et al. (2009) use the polygonal shape distribution (number of edges) to compare simulations with experimental tissue. In Marinari et al. (2012) average cell shape anisotropy is used to quantify the disorder of the midline of the dorsal thorax of the developing fly. Cell shape anisotropy is defined by the ratio of the long to short axis of the epithelial cell. Other shape approaches include the average roundness defined as $4\pi \frac{\text{area}}{\text{perimeter}^2}$ or its variance (Sudbø et al., 2000). Cell shape regularity in Sahlin et al. (2009) is quantified as the ratio between the squared length of a cell boundary and the area of
the cell for the examination of different cell division models and comparison with experimental data extracted from the shoot apex of Arabidopsis thaliana.

Analysis of topology using the distributions of number of neighbours has been applied to compare cells across epithelia of different species (Drosophila, Hydra, Xenopus, Cucumis, Allium, Euonymus, Dryopteris) and plant epidermis (Anacharis, Volvox) (Gibson et al., 2006; Gibson and Gibson, 2009) to study the emergence of geometric order in proliferating tissue. Distributions of numbers of neighbours for cells in plants' tissue have been compared with those of simulations for testing different division rules (Sahlin et al., 2009; Patel et al., 2009). Similarly, such distributions have been examined in studies of tumorigenesis and tissue homeostasis to test models of cellular interaction networks (Csikász-Nagy et al., 2013). When ‘non-cooperating’ cells interact with a population of cooperative cells, they have an advantage and proliferate faster than the latter. This affects the cell topology and the observed distributions of neighbours.

Guillaud et al. (2004) analysed samples of healthy and cancerous cervical epithelial tissue using quantitative methods on nuclear features. Apart from simple morphological features (area, mean radius, etc) a large set of ‘architectural’ features based on Voronoi polygons (mean area, standard deviation of area, area disorder, mean perimeter, etc) and Delaunay graphs (mean Delaunay nearest-neighbours distance, standard deviation of mean Delaunay nearest-neighbours distance) were used in linear discriminant analysis.

### 2.5.2 Measures of Order from Diverse Scientific Areas

#### 2.5.2.1 Statistical Analysis of Fracture Patterns

Shin et al. (2012) analysed fracture patterns observed in ancient wall paintings and compared them with simulated ones in order to validate fracture models. They analysed a list of geometrical quantities of the fracture pattern: i) fragment adjacency, defined as the number of fragments adjacent to each fragment, ii) fragment area, iii) fragment convexity, defined as the area of the fragment divided by the area of its convex hull, iv) fragment circularity, defined as $\sqrt{\frac{4\pi \text{area}}{\text{perimeter}}}$, v) edge
2.5. Measures of Order

length, vi) edge straightness, defined as the length of the straight line segment between adjacent junctions divided by the length of the path between the two edges, vii) edge orientation, viii) edge type, according to the types defined by McBride and Kimia (2003) depending on alignment of corners, ix) junction angle, and x) junction type, categorised according to the number of adjacent fragments. The histograms of all above quantities for real fracture patterns and simulated ones are extracted and compared.

2.5.2.2 Entropy Based Approaches

Many statistical approaches to quantify order rely on analysis of distributions. These are commonly summarised by the mean and/or the variance. The variance aims to capture the spread of the distribution. In general, low variance is associated with ordered states and high variance with disordered states. However, the variance is not a good measure of the uncertainty when the distribution is distinctly different from Gaussian. A more accurate quantity to express the uncertainty of values is the entropy. For the discrete case of \( n \) possible outcomes with probability mass function \( p(x_i) \), the Shannon (Shannon, 1948) definition of entropy for the random variable \( X \) is:

\[
H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \quad (2.7)
\]

In practice, even for a continuous distribution, a histogram can be used for the estimation of the entropy.

2.5.2.3 Dispersity

Some approaches to quantify spatial order have been developed in the field of materials science where apart from the well-established methodology for studying crystalline solids (Kittel, 2004), significant effort has been put into analysing the structure of materials that do not exhibit exact arrangement of their basic units, e.g.
2.5. Measures of Order

fluids and amorphous solids or powders. Most of the ideas are based on models of monodisperse (same radius) or polydisperse (varying radius) hard spheres or disks and are worth examining for whether they could be transferred to the perceptual domain and used in the analysis of different types of systems (e.g. biological).

The dispersity (or polydispersity index, PDI) is used as a measure of heterogeneity of the sizes of molecules or particles in a mixture in physical chemistry and polymer science (Gilbert et al., 2009). This notion can be extended to cases where similar elements are packed together as happens in powder grains or epithelial cells. It is defined as:

\[ PDI = \frac{\bar{d}_n}{\bar{d}_w} \] (2.8)

where \( \bar{d}_n \) is the average cell diameter and \( \bar{d}_w \) the average cell diameter weighted by cell area (Xu, 2011). For point patterns, the dispersity measure can be applied on the Voronoi cells, using the square root of the cell area as an effective diameter.

2.5.2.4 Bond Orientational Order Parameter \( Q \)

It is clear that the size distribution of elements alone, deprived of any information related to the position of the elements, is inadequate as a basis for a measure of order. Even monodisperse spheres can have configurations of different degrees of order as it is shown in studies of identical spheres packing (Cliffe and Goodwin, 2013; Torquato et al., 2000; Kansal et al., 2000; Steinhardt et al., 1983). In these studies the bond-orientational order has been used for capturing orientational symmetries in the relative positions of elements. Assigning spherical harmonics to each vector \( \vec{r} \) connecting neighbouring particles (bond), with \( \theta (\vec{r}) \) and \( \varphi (\vec{r}) \) the polar angles of the vector with respect to some reference system, the local-orientational-order parameter \( Q_{lm}(\vec{r}) \) is defined as:

\[ Q_{lm}(\vec{r}) = Y_{lm}(\theta (\vec{r}), \varphi (\vec{r})) \] (2.9)
2.5. Measures of Order

By averaging over all bonds in the sample and values of \(m\) where \(|m| \leq l\), a measure independent of the reference frame is defined as the \textit{global orientational order} parameter \(Q_l\):

\[
Q_l = \left( \frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{Q}_{lm}|^2 \right)^{1/2}
\]  

(2.10)

As the value of \(Q_6\) is known for the completely uncorrelated sample or ideal gas (perfectly disordered) and for the fcc packing (face-centred-cubic lattice, perfectly ordered packing), a normalised version of \(Q_6\) taking values in the range of \([0, 1]\) can be used to characterise the order of the system.

For the case of two dimensions, a single order parameter, the \textit{global orientational order}, takes the simpler form:

\[
Q_l = \left| \frac{1}{N_{\text{bond}}} \sum_j \sum_k e^{il\theta_{jk}} \right|
\]  

(2.11)

where \(j\) runs over all particles, \(k\) runs over all neighbours of \(j\), \(\theta_{jk}\) is the angle between some fixed reference axis in the system and the bond connecting the particles \(j\) and \(k\), and \(N_{\text{bond}}\) is the total number of bonds in the system (Kansal et al., 2000). Neighbours are commonly defined as particles connected in the Delaunay tessellation (Kansal et al., 2000). Index \(l\) is 4, for a fourfold symmetry (square lattice), and 6 for the sixfold symmetry (hexagonal lattice). For the perfect hexagonal lattice, \(Q_6 = 1\), while for the ideal gas (totally random arrangement of particles), \(Q_6 \sim N_{\text{bond}}^{-1/2}\).

2.5.2.5 Translational Order Parameter \(T\)

The translational order parameter \(T\) (Torquato et al., 2000) with respect to the perfect fcc lattice of the same density is defined as:
2.5. Measures of Order

\[ T = \left| \frac{\sum_{i=1}^{N_C} (n_i - n_i^{\text{ideal}})}{\sum_{i=1}^{N_C} (n_i^{\text{fcc}} - n_i^{\text{ideal}})} \right| \]  

(2.12)

where \( n_i \) expresses the occupation percentage of particles for spherical shells at specific radii \( r_i \) (which is equal to the distance of the \( i \)th nearest-neighbour for the fcc lattice of equal density) and with specific width for the system of interest, and \( n_i^{\text{ideal}} \), \( n_i^{\text{fcc}} \) are the same quantities for the case of ideal gas and fcc packing respectively. \( N_C \) is the total number of neighbour shells considered.

2.5.2.6 Translational Order Parameter \( T^* \)

The translational order parameter \( T^* \) is independent of a reference lattice (Truskett et al., 2000) and detects density-density correlations. It is based on the radial distribution function \( g(r) \) which expresses how density varies with respect to a distance from a reference point. For equal size sphere systems, it is defined as:

\[ T^* = \frac{\int_{r_o}^{r_C} |g(r) - 1| dr}{r_C - r_o} \]  

(2.13)

where \( r_o, r_C \), are the radius of the sphere and a radius cut off, limited by the size of the sample, respectively.

2.5.2.7 Structural Correlation Length

Sausset and Levine (Sausset and Levine, 2011) use entropy for the ensemble of states of patches to define a structural correlation length for amorphous materials. Using a superimposing and rotating rule together with a configuration distance threshold for patches of the same size \( V \), they construct equivalent classes whose entropy they calculate. By studying the dependence of this entropy on the size of the patch they define a correlation length which they use as a measure of order. It
is clear that this type of order is limited to structures where the extent of local order is important (e.g. polycrystals). Applying entropy on classes of patches is an interesting approach for quantification of order.

2.5.2.8 Mutual Information
Dunleavy et al. (2012) use mutual information as a structure-independent measure of order to study a 2D binary mixture of hard-disks (glass model). In this sense, order is equivalent to how the configuration in one part of the system can ‘restrict’ the configuration in another. Since both configurations are sampled from the same material, it is necessary to specify distinct regions or a rule (e.g. a distance between the parts) to estimate the mutual information for two variables. In the case of glass this approach could be appropriate as in the end a correlation length may be of interest. In the case of overall order of point patterns, however, it does not provide any useful information. Furthermore, as the number of possible configurations increases, this approach becomes extremely susceptible to low number bias. This approach is suitable for large numbers of elements where the spatial correlation of density is meaningful.

2.5.2.9 Electrostatic Potential
The electrostatic potential of the Coulomb interaction can provide a candidate measure of order. Based on the naïve observation that electrically charged particles of the same type of charge, when free to move in a confined space, tend to occupy positions which maximise their inter-particle distances, one could define a measure of order related to the total electric energy for a system of such particles (Griffiths, 2012). Considering point particles of the same electric charge, the formula for the total electrical energy can be simplified in the following form:

\[ U = - \sum_{i \neq j} \frac{1}{r_{ij} \alpha} \]  

(2.14)
where \( r_{ij} \) expresses the distance between two particles \( i, j \) while the electric charges \( q_i, q_j \) as well as the electrostatic constant \( k_e \), have been set equal to unity. The parameter \( \alpha \) can be adjusted, with the default value being \( \alpha = 1 \). A normalisation over the number of points is necessary.

### 2.5.2.10 Lennard-Jones Potential

A more realistic interaction between molecules is described by the Lennard-Jones potential. This interaction contains a term for attractive and a term for repulsive interaction. There is a specific distance \( r_m \) for a pair of molecules where the system is at the equilibrium. For a system of molecules, the total potential energy is of the form:

\[
U = -\sum_{i \neq j} \left[ \left( \frac{r_m}{r_{ij}} \right)^{12} - 2 \left( \frac{r_m}{r_{ij}} \right)^6 \right]
\]  

(2.15)

where again, \( r_{ij} \) expresses the distance between two particles \( i, j \).

### 2.5.2.11 Fano Factor

Neurons’ spike trains are often characterised by examination of a statistic called Fano factor (Fano, 1947; Eden and Kramer, 2010). It is defined as the ratio of the sample variance to the sample mean of the numbers of spikes in a set of time intervals. In the case that the spikes are produced according to a Poisson process of a constant mean rate, this statistic approaches unity. Higher values are related to clustered occurrences and lower values to disperse. Since this analysis is applied on one-dimensional data, in order to be applied in two-dimensional patterns, a procedure for transforming spatial information into a unidimensional quantity is required. One straightforward strategy would be to partition, using a grid, the available area in square or rectangular subregions and count the numbers of points which reside in each subregion. If the points are distributed according to a randomly uniform distribution, the equivalent Fano factor would have a theoretical value equal to one.
Similarly, values for the Fano factor greater than one would suggest spacing more regular than the uniformly random case, while values less than one would suggest clustering.

2.5.2.12 Hopkins Statistic

The *Hopkins* statistic is used in point pattern analysis to test for divergence from the completely spatially random case (Jain and Dubes, 1988). It is based on the comparison between nearest-neighbour distances from randomly picked probes (points of the point pattern) and uniformly randomly picked points (random points on the plane). For the definition of the Hopkins statistic, we denote by $d(p_1, p_2)$ the distance between two points $p_1(x_1,y_1)$ and $p_2(x_2,y_2)$, $S$ the set of $n$ total probes, $s_j = (x_j,y_j)$ a random sampling point and $\tilde{p}_k = p_{ik}$ a random chosen probe. We also denote by $d(p_j,S) = \min\{d(p_j,p_k), \forall p_k \in S\}$ the minimum distance of the point $p_j$ from all points belonging to set $S$. By choosing in random $m$ sampling points and probes we define the *Hopkins* statistics as:

$$H = \frac{U}{U+W}$$  \hspace{1cm} (2.16)

where $W = \sum_{j=1}^{m} d^2(\tilde{p}_k,S)$ and $U = \sum_{j=1}^{m} d^2(s_j,S)$. The values of $H$ are restricted in $[0, 1]$ with the completely random case giving $H = 0.5$ while existence of clustering would result in higher values. Lower values of this statistic would be the result of dispersion. The probability density function for $H$ is:

$$f(H) = \frac{H^{m-1}(1-H)^{m-1}}{\Gamma(m)^2 / \Gamma(2m)}$$  \hspace{1cm} (2.17)

where $\Gamma(n) = (n-1)!$ (Jain and Dubes, 1988). For the analysis of a single system, a bootstrap approach is used, in practice, to estimate the Hopkins statistic (Zhang et al., 2006). In a previous project (Protonotarios, 2011), it has been shown, using
specifically designed patterns, that this statistic can be easily tricked as an index of
deviation from the completely random case and its power is not considered suffi-
cient for relatively ordered patterns.

2.5.2.13 Ripley’s Statistics
Ripley’s functions are also used in point pattern analysis for testing deviation from
the completely spatially random case. For a set \( S \) of \( n \) points on the plane, we
estimate the intensity, as \( \lambda = \frac{n}{A(C)} \), where \( A(C) \) is the area of the set \( C \). With \( r > 0 \)
as a parameter, we define the set \( S_p(r) = \{ q \in S; 0 < d(p, q) \leq r \} \) and an indicator
function as \( I_p(r) = |S_p(r)| \) (i.e. the number of probes of \( S \) that are located in a
distance less than \( r \) from point \( p \)). Ripley’s \( K \) function is defined as:

\[
K(r) = \frac{Av(r)}{\lambda}
\]  

(2.18)

where \( Av(r) \) is the average indicator function for the parameter \( r \). Its expected value
for the completely spatially random case is: \( Av(r) = \frac{1}{n} \sum_{i=1}^{n} I_{p_i}(r) \approx \lambda \pi r^2 \).

Two other forms of Ripley’s \( K \) function are derived from \( K \) as:

\[
L(r) = \sqrt{\frac{K(r)}{\pi}}
\]

(2.19)

with expected value \( r \), and

\[
H(r) \text{ or } \tilde{L}(r) = L(r) - r
\]

(2.20)

with expected value 0.

To test for divergence from the completely spatially random case, the function
curves derived for a specific arrangement of points, are compared with the 95%
most common curves derived from simulated completely spatially random point-
sets. Divergence above the simulated curves represents clustering, and the value of \( r \) at maximum divergence gives an estimation of the cluster size. Divergence below the simulated curves represents “repulsion” between points (Kiskowski et al., 2009; Zhang et al., 2006; Ripley, 1977). While Ripley’s statistics can capture clustering and repulsion taking into account density variations, there is no simple way of extracting a single numerical value to represent the degree of order. Moreover, unless the number of points is very large Ripley’s function curves can be very noisy.

### 2.5.2.14 Autocorrelation

Autocorrelation is commonly used in signal or image processing for the detection of periodicity. For a continuous uni-dimensional signal \( s(t) \) between \( t_1 \) and \( t_2 \), autocorrelation is defined as:

\[
R(\tau) = \frac{1}{t_2-t_1} \int_{t_1}^{t_2} s(t) s(t+\tau) dt
\]

Since translational symmetry is a crucial aspect of spatial organisation, extension of autocorrelation to two dimensions offers a promising candidate measure of order. For the application of autocorrelation to point patterns, it is necessary to express each point with an appropriate localised function (e.g. with a Gaussian blob). The position of the first maximum represents the period of the pattern. The height of the maximum can vary depending on how much ‘periodic’ the pattern is. The autocorrelation is related to the Fourier transform of the signal (Wiener-Khinchin theorem).

### 2.5.2.15 Symmetry in Computer Vision

In computer vision, symmetry is very important as applications that take into account symmetry can be more efficient, and symmetry is considered a cue for recognition and grouping. Partial or approximate symmetry can be identified on the image and the elementary building blocks have to be matched. One approach to
quantify the degree of symmetry is to apply a transformation on the image based on one type of symmetry and then calculate an overlapping factor between the image and a reference pattern. Another measure of approximate symmetry is based on the amount of effort needed in order to transform a given object to a perfectly symmetric one. (Mitra et al., 2012; Zabrodsky et al., 1995). In point patterns, when there is point-to-point correspondence, similar measures could be applied. However, in some situations (e.g. when points are absent near expected positions) the application of such measures is difficult. Other applications of symmetry in computer vision include identification of structural defects in regular textures. Chetverikov and Hanbury (2002) used autocorrelation applied on a set of windows covering the image to identify these defects.

An interesting class of tessellations that find applications in computer vision (e.g. clustering, image compression) are the Centroidal Voronoi Tessellations (CVTs). These are Voronoi tessellations whose generators coincide with the centroids of the Voronoi cells. Given a density function these tessellations generate an evenly-spaced distribution of sites. Methods to compute them rely on the minimization of an energy function which depends on the average distance between the generator and the centroid of the cell (Liu et al., 2009). Since CVTs appear quite regular, using a measure depending on how close a Voronoi tessellation is to a CVT is a promising idea for quantification of order.

2.6 Conclusion

As it has been obvious from the introductory chapter of the thesis and from attempts to quantify order found in scientific literature, there is a compelling need, not only in Biology but also in other scientific fields, for an objective and validated measure of order. Current quantifications are not validated, and are not consistent. This hinders scientific research and communication. Occasionally, researchers are hesitant to trust measures they use, as they report an obvious disagreement between the measure and what they perceive visually when examining the organisation of a system (Cook, 2004). Nonetheless, there exist many candidate measures which could be
examined for agreement with perceptual order or approaches which could suggest novel measures. Measurement theory provides the framework for our approach of order quantification, and psychophysics provide the necessary experimental and analytical methods for examining perceptual order. Previously, there has been only one attempt to a geometric correlate of order (Preiss and Vickers, 2005); however no scaling model was employed and tested, observers agreement was not quantified, and no comparative assessment of different quantifications was conducted for a diverse set of patterns combining a variety of regularities.
Chapter 3

An Absolute Interval Scale of
Apparent Order for Point Patterns

3.1 Introduction
In this chapter we report experimental results and analyses (Section 3.2) that assess whether the perception of point patterns varies along a dimension of order with an interval scale structure. The analysis tests for this by attempting to place test patterns into a perceptual interval scale ($p$-scale, for perceptual) so that their relative positions predict judgements of the degree of order made by multiple observers. Our results will confirm the reality of the perceptual dimension and establish that it is well extended in terms of its number of distinguishable levels. We will then proceed to establish a pragmatic interval scale of objective order (Section 3.3), as assessed by a novel geometric algorithm ($g$-scale, for geometric), which accurately corresponds with the $p$-scale. We will demonstrate that previously proposed quantifications of order correspond less accurately. By scaling the $g$-scale, so that absolute disorder maps to 0 and perfect order to 10, we define a convenient absolute interval scale ($a$-scale) for the measurement of order (Section 3.4). In Chapter 4 we demonstrate the applicability of the derived $a$-scale by measuring the degree of order in the arrangement of cells on the fly dorsal thorax and plotting its timecourse during development.
3.2 Existence of a Perceptual Scale of Order

3.2.1 Methods

3.2.1.1 Psychophysical Experiments

For reasons explained in Section 2.4, we used pairwise ranking of point patterns as our method of collecting data on subjective perceptually-based order. Participants were not allowed to respond that two patterns were equal in order, so the task is of the two-alternative forced-choice type (2AFC). Since the space of possible point patterns is multiply infinite and contains many different perceptual types, difficult decisions about what patterns should be ranked had to be made. We lack the understanding and means to sample uniformly from the space, so the best that we can hope to do is study the degree of order of patterns from some restricted region of pattern space. However, if our experiments reveal lawful behaviour across the studied region we might be tempted rashly to hypothesize that this extends to the entire space. To prevent this we adopt the strategy of thoroughly studying a particular core region and less thoroughly a broader region. With this strategy we may be able to discover lawful behaviour over the core region and deviations within the broader region.

The aims of the experiment are to determine the reality of a $p$-scale for order, and to assess candidate $g$-scales. These aims place different demands on the experimental design. For testing for the $p$-scale, it is preferable to present a small number of patterns in all possible pairs. For testing $g$-scales, it is preferable to present as many patterns as possible. We reconcile these two considerations—core vs. border region, all pairs vs. many patterns—by an experimental design using a set $A$ of 20 patterns from a core region of pattern space presented in all possible $(190 = 19 \times 20/2)$ pairs, and a set $B$ of 240 patterns from a broader region presented in 120 fixed pairings.

Each point pattern for set $A$ and set $B$ was created with the same multi-step process: (1) a base lattice—triangular, rectangular, or hexagonal (honeycomb)—was chosen, (2) the lattice points were jittered using Gaussian positional noise, (3)
3.2. Existence of a Perceptual Scale of Order

The pattern was affinely distorted by stretching it a random amount along a random orientation, (4) a fraction of randomly chosen points were removed, or a number of randomly positioned points were added, (5) a non-linear positional warp, was applied, (6) a random centre for a circular window was chosen, and a radius was determined so that 180 points were visible within. The parameters of this pattern generation process were the (i) base lattice, (ii) jitter magnitude, (iii) orientation and magnitude of the stretch, (iv) fraction of points removed or added, and (v) five parameters for the non-linear warp which was implemented as a bicubic transformation of $x$-$y$ coordinates. For the positional jitter of the points a Gaussian distribution for the displacements was preferred as this resembles the effect of multiple random factors that could affect the position of an element in an evolving system.

We used different procedures to select the patterns for set $A$ and set $B$, reflecting different roles for these datasets. Set $A$ was used to test the reality of the $p$-scale and to determine its extent. Both set $A$ and set $B$ were used to test the accuracy of candidate $g$-scales. Set $B$ was used to determine the limits of applicability of the best $g$-scale.

To construct set $A$, a pool of 120 patterns were generated using a variety of parameter settings, avoiding those that were multiply extreme. From the pool a subset of twenty were chosen with the criteria that they should: be roughly uniformly spaced in order all the way from highly ordered to fully disordered; and be diverse in terms of the base lattice, the amounts of perturbation, deletion/addition of points and the degree of warp. The patterns of set $A$ are shown in Figures 3.1, 3.3 and 3.13 where they are numbered according to their estimated $p$-scale ranks (see Section 3.2.2.2). To construct set $B$, 120 pairs of patterns were generated using random parameters settings.

Patterns were displayed in pairs on a 40 cm diagonal LCD screen at a distance of 50 cm under comfortable room illumination. Each pattern was rendered using solid black dots of 1 mm diameter on a white circular disk of radius $r = 6.2$ cm. Pairs were on a grey background. Observers first viewed patterns from set $A$ (each of the $19 \times 20/2 = 190$ pairs was presented twice) followed by each of the set $B$
3.2. Existence of a Perceptual Scale of Order

Figure 3.1: Point patterns from set A, numbered by their estimated $p$-scale ranks (Section 3.2.2.2).

pattern pairs viewed once. Trials carried out within each block were performed in random order; in each trial the patterns were randomly oriented and allocated to left or right. Randomization aimed to reduce bias and effects of adaptation. Observers were given written instructions to use the keyboard to indicate the pattern which “appeared more ordered” to them, and to proceed at their own pace. All participants took 15-20 minutes to complete their 500 trials. Presentation of stimuli and recording of responses were controlled using the MATLAB Psychtoolbox (Brainard, 1997). Twenty observers (14 male), with normal or corrected-to-normal vision and at least undergraduate level education, took part. Participants’ age ranged from 18 to 38 years ($Mean = 25.5$ years, $SD = 5.5$ years). Our research adhered to the tenets of the Declaration of Helsinki for the protection of human subjects.

3.2.2 Results

3.2.2.1 Variability of the Data

We computed three measures of response variability: the intra, inter and maximum agreement rates. All three rates were computed for set A, which consisted of 20 patterns presented in all 190 pairs twice to each observer; whereas only inter and maximum rates were computed for set B as this consisted of 240 patterns in 120 fixed pairs presented only once to each observer. Rates are shown in Table 3.1.
3.2. Existence of a Perceptual Scale of Order

Table 3.1: Agreement rates.

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th>Set B</th>
</tr>
</thead>
<tbody>
<tr>
<td>intra (%)</td>
<td>85</td>
<td>-</td>
</tr>
<tr>
<td>inter (%)</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>maximum (%)</td>
<td>88</td>
<td>86</td>
</tr>
</tbody>
</table>

The *intra* agreement rate is the probability that a random observer would choose the same pattern both times when faced twice with the same random trial. The rate of 85% for set A demonstrates that observers do not always respond the same to a pair of patterns. This is not surprising since with 20 patterns in set A we would expect some to have similar levels of order.

The *inter* agreement rate is the probability that two observers will agree on which pattern of a pair is more ordered. These rates are within 1% for the two sets. On set A, the *inter* rate is only 2% less than the *intra* which shows that there is very little variation between observers over-and-above their personal variability. There is, therefore, a strong consensus in the perception of order.

The *maximum* agreement rate is the probability that on a random trial a random observer will agree with an optimal ranking of the patterns, where the optimal ranking is that which maximises this rate. Estimating the optimal ranking and computing the *maximum* rate on the same set of data could result in overestimation. For an unbiased estimation of this rate, we randomly split the data into two groups (each group corresponding to ten participants), determined the optimal ranking for the first group and then computed the agreement of the second group with this ranking (Kohavi, 1995). For an estimation of the complete cross-validation, we conducted 1,000 Monte-Carlo repetitions. The average over the 1,000 repetitions is the final *maximum* agreement value we report. For both sets of patterns, this rate is roughly 5% more than the *inter* rate. We include this rate as it will provide a context to the performance scores for geometric measures in Section 3.3.2.1.
3.2. Existence of a Perceptual Scale of Order

3.2.2.2 Validating the $p$-scale

Linear models to account for paired comparisons of stimuli relative to some indicated perceptual dimension originate with Thurstone (Thurstone, 1927). In linear models, the perceptual dimension is modelled as an interval scale within which each stimulus ($S_i$) is assumed to have a true attribute value ($M_i$). Each separate perception of a stimulus has its own attribute value ($\psi_i$) from the same scale, which is a noisy realisation of the true value. When an observer compares two stimuli ($S_i$ and $S_j$) they report which of the two noisy realisations ($\psi_i$, $\psi_j$) is larger, which can vary if the trial is repeated. Depending on assumptions about the realization noise and its correlations it can be shown that there exists a monotonic preference function $P : \mathbb{R} \to [0, 1]$ which maps the signed difference $\Delta M = M_i - M_j$ between the true values to the probability that $S_i$ will be preferred to $S_j$ (David, 1988). If the noise is assumed normally-distributed, stationary and uncorrelated (Thurstone Model Case V) then $P(\Delta M) = \frac{1}{2} \left( 1 + \text{erf} \left( \frac{\Delta M}{k_e} \right) \right)$ (Thurstone, 1927). If the noise is assumed Gumbel-distributed, stationary and uncorrelated (Bradley-Terry Model) then there will be a logistic preference function $P(\Delta M) = \left( 1 + e^{-\Delta M/k_l} \right)^{-1}$ (Bradley and Terry, 1952). The positive constants $k_e$, $k_l$ in the error and logistic function models control the units of the scale of attribute values. We choose them so that a unit distance on the scale corresponds to a just-noticeable difference (jnd). By convention the magnitude of a jnd is such that an observer will have a 75% chance of correctly ordering two stimuli whose attributes are different by this amount (Torgerson, 1958). In our notation, the requirement is that $P(1) = 0.75$, which is achieved if $k_e = 2.10$ and $k_l = 0.91$.

During psychophysical experiments, stimulus independent errors can occur (lapses), even when the difference in the quality between stimuli is large. Unless modelled, these can bias estimates of attribute values (Wichmann and Hill, 2001a,b). Lapsing at a rate $\lambda$ is incorporated in the model by using a rescaled preference function $\lambda + (1 - 2\lambda) P(\Delta M)$.

We estimated the parameters of the model ($M_i$ for each stimulus, and the lapse rate parameter $\lambda$) by Maximum Likelihood (ML) fitting to the psychophysical data.
using gradient descent with multiple random starts to check for stability. A common approach for the assessment of goodness-of-fit (GoF) is to calculate an error term, as the Pearson $X^2$, which asymptotically behaves as a $\chi^2$ distribution when the binomial distribution becomes Gaussian (Klein, 2001). However, in some trials the frequencies for one of the two patterns can be low (<5) and in this case this assumption does not hold, so the chi-square test cannot be trusted. Wichmann and Hill (2001a) provide an alternative option for the estimation of the GoF suggesting Monte Carlo simulations to generate the necessary distributions. Instead of the Pearson $X^2$, the log-likelihood ratio or deviance is proposed. Deviance of a dataset (original or simulated) is twice the difference between the log-likelihood of the dataset given the ML model, and the log likelihood of the dataset given a saturated model, which in this case specifies a separate ML probability for each trial. Unlike $X^2$, not a simple criterion exists for the test of the asymptotic behaviour for the deviance and so the goodness-of-fit is not defined on the basis of the $\chi^2$ distribution, but on a distribution created through Monte Carlo simulation. The deviance is calculated with the best fitting parameters and function, for each generated set of simulated data, for a large number of repetitions. Thus, we get a distribution for the deviance which is the one we would expect from responses that result from binomial distributions with the theoretical preference probabilities. The goodness-of-fit (GoF) of the ML model is assessed by comparing its empirical deviance to the distribution of deviances that result by generation of random datasets from the ML model (Wichmann and Hill, 2001a).

For set $A$, the empirical deviance for the ML fitted error function model is 223.5 and the 95% interval of acceptable deviances is [152.5, 220.5]. Since the empirical deviance is outside of the acceptable interval the model is rejected. The empirical deviance for the logistic model is 211.7 and the 95% interval of the random datasets is [144.6, 214.9], hence the model is accepted. Therefore, our hypothesis that the consensus of perceptually-based order has the structure of an interval scale is supported. We refer to this scale as the $p$-scale, and will treat the ML estimates of the true values of our stimuli as values on this $p$-scale. In Figure 3.2 we compare the
3.2. Existence of a Perceptual Scale of Order

experimental preference rates for trials using set $A$ to the predicted rates based on the $p$-scale values of the stimuli. Visually the fit is good which is consistent with the results of the GoF analysis. Having accepted the linear model with a logistic preference function, we now present the ML estimates of the $p$-scale values of set $A$ stimuli. Since only differences in $p$-scale values, not absolute values, are used to predict preference rates, ML fitting only estimates $p$-values up to an additive constant for the entire set. The $p$-scale values of set $A$ are shown in Figure 3.3. They cover a range of 9.86 jnds. The estimated lapse rate was 0.0008.

![Figure 3.2:](image)

**Figure 3.2:** Compares predicted and experimental preference rates. The horizontal axis is the signed difference in the estimated true attributes of two stimuli ($\Delta M$) in a paired comparison. The vertical axis is the probability that the first stimulus will be chosen as more ordered than the second. The solid curve shows the predictions of the ML logistic with lapsing model. There are 380 ($= 2 \times 190$) plotted points as the data for each trial using set $A$ is plotted twice.

We compare the resulting scales of the two models for set $A$ (error function and logistic) with a scatter plot of the estimated order values (Figure 3.4). Graph shows that these are highly correlated, meaning that the exact type of noise distribution does not affect significantly the relative spacing between the estimated attribute values.

To estimate the uncertainty of the fitted $p$-scale values we used a bootstrap method (Wichmann and Hill, 2001b; Efron, 1979). The method generates sets of synthetic experimental data from the fitted model, to each of which a new model is fitted. In the nonparametric bootstrap method the original set of data is resampled...
3.2. Existence of a Perceptual Scale of Order

Figure 3.3: At bottom is a visual representation of the $p$-scale of perceptually-based order. Order increases from left to right. The points and error bars indicate the estimated $p$-scale values of the twenty stimuli of set $A$. The point patterns for a selection of the stimuli are shown at top. The values span a range of 9.86 jnds, the size of which are indicated by the tick marks. Error bars indicate +/- one standard deviation of the uncertainty of the estimates. Numbers in circles next to each pattern correspond to their ranking according to their $p$-scale estimated values. (The same numbering system is used in Figures 3.1 and 3.13)

Figure 3.4: Comparison between the estimated attribute values derived with the error function and the logistic function models for set $A$. A straight line connecting the extreme order values shows that the two scales are highly correlated.
with replacement, assuming that the binomial probabilities are those exactly given by the preferences. In the parametric bootstrap a specific model is assumed and the probabilities with which the data are generated are those dictated by the model. The parametric model is considered a better choice as it is in accordance with the selection of the model without requiring additional parameters and it considers the data as noisy samples from a smooth and monotonic function. Therefore, if one performs two realisations of the same experiment the actual data may be different and thus the nonparametric bootstrap would result in different confidence intervals for each stimulus point; while if the model fits the data, even with two independent realisations the parametric bootstrap will result to identical estimates of the confidence intervals as these will depend only on the parameters and the model function (Wichmann and Hill, 2001b,a; Efron, 1979). To remove the effect of arbitrary additive constants, we align different bootstrap estimates of \( p \)-scale values by subtracting the mean of each set. We thus obtain a set of \( p \)-scale estimates for each stimulus, the standard deviation of which provides an estimate of uncertainty. Across the stimuli of set \( A \) the standard deviations ranged from 0.10 to 0.30 jnds, with 0.14 jnds being the (RMS) average. The 95% confidence interval, containing 3.84 standard deviations from the bootstrap sample, for the size of the range of \( p \)-scale values in set \( A \) was [9.08, 10.64] jnds. The 95% confidence interval for the lapse rate was [\( 10^{-8} \), 0.0013].

Figure 3.5 shows a diagrammatic summary of the steps followed for the estimation of the model parameters (fitted order values) and the GoF test (a), as well as the uncertainty of the estimated values (b).
Figure 3.5: Diagrammatic summary of the deviance GoF test (a), and the estimation of the uncertainty of the estimated order values (b).
3.3 Evaluating Geometrical Measures

3.3.1 Methods

3.3.1.1 Existing Geometrical Measures

As described in Section 2.5, methods for quantification of order can be based directly on the point locations, or on a geometric construction derived from them. Common constructions used to analyse biological patterns are the Voronoi diagram (Voronoi, 1908) and the related Delaunay triangulation (Delaunay, 1934) (Section 2.5, Figure 2.4). The Voronoi diagram is the partitioning of the plane into convex polygons (cells) such that each contains all locations closer to one point of the pattern than to any other points. The Delaunay triangulation of a point pattern is the unique triangulation such that no triangle contains any points of the pattern within its circumcircle. The Voronoi diagram and Delaunay triangulation are related: Delaunay triangle circumcentres are coincident with Voronoi side junctions.

Measures have been proposed that depend on distances, areas, shapes and topology. Distance-based measures use either nearest-neighbour or multiple neighbours (as defined by a triangulation or other structure). Examples include: global mean nearest-neighbour distance (Raymond and Barthel, 2004; Guillaud et al., 2004), global variance of nearest-neighbour distances (Allison et al., 2010; Cohen et al., 2010; Guillaud et al., 2004) and other measures of the dispersion of nearest-neighbour distances (Galli-Resta and Novelli, 2000; Cook, 1996). Other distance based approaches use all, rather than nearest-neighbour, distances such as Ripley’s L-function (Shapiro et al., 1985) and auto-correlation analysis (Cook and Podugolnikova, 2001; Raven and Reese, 2003). Area-based approaches include: variance of polygon area (Farhadifar et al., 2007; Guillaud et al., 2004), area disorder (Sudbø et al., 2000; Guillaud et al., 2004) and other measures of area distribution (Raven and Reese, 2003). Shape-based approaches use the distributions of polygons edges (Patel et al., 2009; Farhadifar et al., 2007; Guillaud et al., 2004), cell anisotropy (Marinari et al., 2012; Guillaud et al., 2004), polygon roundness (Marcelpoil and Usson, 1992; Guillaud et al., 2004), and Voronoi cell angles (Shapiro et al., 1985).
3.3. Evaluating Geometrical Measures

or Delaunay triangles angles (Galli-Resta and Novelli, 2000). Topological measures analyse the variation in the number of neighbours points have (Gibson et al., 2006; Sahlin et al., 2009; Sudbø et al., 2000).

Based on pilot studies (e.g. Protonotarios et al. (2012)) we have chosen to present the assessment of six representative methods from the above list: mean nearest-neighbour distance, variance of the number of neighbours, variance of Voronoi cell area, Voronoi cell area disorder, variance of nearest-neighbour distance and maximum autocorrelation. The autocorrelation method has a single tunable parameter (width of the point spread function); the other methods have none.

3.3.1.2 Novel Measures

As well as previously proposed measures of order, we also assess four novel measures of order that we have designed. The first two methods concern the local symmetry of the Voronoi diagram, and the second two concern the variability of Delaunay triangles. We report the tunable parameters for each method.

Pairwise local symmetry. This measure is the mean of a local symmetry score computed separately for each pair of adjacent Voronoi cells. The pair score measures their symmetry by transforming one cell so that it lies roughly on top of the other, and then measuring their area overlap (intersection divided by union). The transformation can be by reflection in the shared Voronoi side, or by 180° rotation about the Voronoi side midpoint; whichever transformation results in the greater overlap is used. Figure 3.6 provides a visual description of this process. We allow overlap scores to be transformed by a power law, and then compute their weighted mean with weights which are power law transformed areas of the involved cells. The exponents of both power laws are tuneable. The area overlap score is based on a standard formulation of sets similarity (Jaccard index, Jaccard (1901)). Originally Jaccard proposed this measure (coefficient de communauté) for the quantification of the similarity between alpine regions with respect to the number of species they shared. It was defined as the number of species common in the two regions divided by the total number of species in both regions (O’Reilly, 2013). Now this measure
is adapted for use in computer vision for the estimation of the accuracy in binary pixel classification (O’Reilly et al., 2011; O’Reilly, 2013).

Figure 3.6: Schematic of the pairwise local symmetry measure. This measure is based on the Voronoi tessellation of the point pattern. Left: A selected Voronoi cell (A) is shown (yellow) with its adjacent cells (pink). Right: For the pair of adjacent cells A and B the local score is computed as the overlap of the areas of cell A and the reflected in their common side cell B (B′) divided by the area of their union. This is formulated as \[
\frac{\text{area}(A \cap B')}{\text{area}(A \cup B')}
\].

Centroidal symmetry. While the previous measure assesses local symmetry based on pairs of adjacent Voronoi cells, this measure assesses order on the basis of each cell alone. The symmetry of a cell is measured by the distance between the generator of the cell (the point of the pattern which the cell surrounds) and the centroid of the cell, made dimensionless by dividing by the square root of the cell area. We allowed the same two tuneable power laws as for pairwise local symmetry. This measure is based on the observation that Centroidal Voronoi Tessellations (CVTs) appear quite ordered. Centroidal Voronoi Tessellations are the Voronoi tessellations whose generating points coincide with the centroids of the corresponding Voronoi cells (Du et al., 1999). The measure is demonstrated in Figure 3.7 where the positions of the generators and the centroids of the cells are shown for two patterns which differ in degree of order.

Delaunay sides entropy. Previous measures that use Delaunay side lengths pool all lengths together and assess their mean or variation. Our novel measure
Figure 3.7: Schematic of the centroidal symmetry measure. Top row shows two point patterns which differ only in the amount of positional jitter of their points. The one on the right, with the greater jitter, appears less ordered than the one on the left. Bottom row shows, at the same scale, the corresponding Voronoi tessellations along with the original points (generators), in blue, and the centroids of the Voronoi cells, in green. It is readily noticed that the average distance between the generators and the corresponding centroids is larger in the case of the more disordered pattern. Only the centroids of the cells that are fully contained in the circular disk are shown.
3.3. Evaluating Geometrical Measures

analyses three disjoint sets of Delaunay side lengths: one for the shorter sides in each Delaunay triangle, another for the longest, and the third for the intermediate length sides. We assess the variability of each set of lengths using an entropy measure based on a kernel density estimation of the distribution using a Gaussian kernel as illustrated in Figure 3.8. The method has one tuneable parameter, the width of the Gaussian kernel.

Delaunay triangles entropy. Methods based on nearest neighbour distances assess the size of the spaces between the points of a pattern using a single univariate histogram. In contrast, the Delaunay sides entropy measure of the previous section uses three univariate histograms, which gives it some sensitivity to the shape of the inter-point spaces as well as their size. However, information on shape and size is entangled and present across all three histograms. In this final measure, we separate these two aspects of the inter-point spaces: rather than three univariate length histograms we form one univariate size histogram (triangle areas) and one bivariate shape histogram (triangle shapes as represented by the lengths of the two shorter sides divided by the longest). As with Delaunay sides entropy, we use kernel-density estimation to produce smooth histograms from which stable entropy estimates are computed. The process is illustrated in Figure 3.9. The method has two tuneable parameters: the size of the two smoothing Gaussians.
3.3. Evaluating Geometrical Measures

Figure 3.8: Schematic of the Delaunay sides entropy measure. From (a) a point pattern, (b) the Delaunay triangulation is constructed. From the triangulation, side length histograms (c) are determined for shortest, intermediate and longest sides. Smoothed histograms are created by kernel-density estimation and used for entropy estimation.
3.3. Evaluating Geometrical Measures

(a) point pattern

(b) Delaunay triangulation

(c) Delaunay triangles size distribution

(d) Delaunay triangles shape distribution

Figure 3.9: Schematic of the Delaunay triangles entropy measure. (a) Two example patterns, (b) their Delaunay triangulations, (c) the triangles size smoothed histograms, (d) the triangles shape histograms (raw data and smoothed). The more disordered pattern (right) has wider histograms than the more ordered pattern (left).
3.3.2 Results

3.3.2.1 Assessing Candidate g-scales

A geometric measure can agree with the ordinal structure of the \( p \)-scale while differing in the interval structure. In such a case, if the raw outputs of the measure are transformed by an appropriate monotonic function the resulting outputs will have perfect interval structure. This monotonic transformation also allows a unified expression of all measures on the same scale and with the same units (those of the perceptual \( p \)-scale). For each of the measures presented in Sections 3.3.1.1 and 3.3.1.2, we will report the ordinal accuracy of their raw outputs and the interval accuracy of their optimally transformed outputs. While the linear models are evaluated with chi-square or deviance analysis, such analysis will not be practical for the evaluation of the \( g \)-scales since we are mostly interested in estimating their predictive power in perceptual order and not their ability to predict the experimental preferences.

To find the optimal transformation for each measure we use a parameterized representation of a smooth monotonic function implemented as a constrained polynomial. As well as the optimized parameters for this transformation, some of the measures have additional internal parameters that can also be tuned for optimal performance. For example, the methods that use the entropy of a distribution as a measure of its variability require a smoothing kernel bandwidth to be defined. We will refer to the combination of a measure and transformation as an algorithm. We can optimize the parameters of each algorithm by maximizing the likelihood of the experimental data, where the likelihood is computed using the probabilities that result by feeding the algorithm outputs into the same logistic function that was used when testing the \( p \)-scale. Each optimized algorithm defines a candidate \( g \)-scale.

We use leave-one-out cross-validation to make unbiased estimates of the accuracy with which each \( g \)-scale agrees with the \( p \)-scale. In the leave-one-out approach, we set aside one of the 20 patterns of set \( A \) and optimize the algorithm’s parameters with respect to the subset of experimental data that involves only the other 19 patterns. Next, we find the constant offset needed to align the 19 \( g \)-scale values
from the optimized algorithm with the 19 corresponding $p$-scale values. Then we compute the $g$-scale value of the left-out pattern using the optimized algorithm, apply the constant offset, and compare the result to the $p$-scale value of the left-out pattern. We repeat this, leaving out in turn each of the 20 patterns, and compute the RMS $g$-scale/$p$-scale difference as the interval accuracy for that $g$-scale.

The performance of the candidate $g$-scales are shown in Table 3.2. Rows 1-6 are previously presented measures as described in Section 3.3.1.1. Rows 7-10 are novel methods described in Section 3.3.1.2. The ‘number of parameters’ column gives $K_{\text{int}} + K_{\text{mon}}$, where $K_{\text{int}}$ is the number of internal parameters, and $K_{\text{mon}}$ the number of parameters of the monotonic transformation. In all cases we report $K_{\text{mon}} = 2$ since it gave consistently better performance than $K_{\text{mon}} = 1$, while $K_{\text{mon}} = 3$ gave marginal improvement for some methods and worse cross-validated performance for others. The ‘ranking performance’ columns show the fraction of psychophysical trials that agree with the ranking of the patterns according to each candidate $g$-scale. These rates cannot exceed the maximum agreement rates in Table 3.2. The ‘accuracy’ column gives the cross-validated RMS difference between the $g$-scale and $p$-scale values.

In comparing the performance of different geometric measures, our ultimate aim is to discover one with excellent interval accuracy for order. Such a measure will necessarily have excellent ranking accuracy. We therefore first assess measures by their ranking accuracy, and then amongst those that perform well, compare them by their interval accuracy. Ranking accuracy can be assessed on both sets but our primary aim is for good accuracy across set $A$, while the role of set $B$ is to allow the limits of measures to be assessed. Interval accuracy can be assessed on set $A$ but not on set $B$.

Table 3.2 shows that previous methods perform roughly similarly, with the exception of autocorrelation, which performs better. Autocorrelation achieves a ranking accuracy within 2% of the ceiling on set $A$ and 5% on set $B$, and an interval accuracy of 0.85 jnds. All our novel methods show equal or better ranking accuracy than previous methods. Amongst the novel methods, the best is Delaunay triangles.
3.3. Evaluating Geometrical Measures

Table 3.2: Performance of candidate g-scales for sets A and B. Methods 1-6 are previously described, 7-10 are novel.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Number of parameters</th>
<th>Ranking performance (%)</th>
<th>Accuracy (RMSE) /jnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 mean nearest-neighbour distance</td>
<td>0 + 2</td>
<td>81 65</td>
<td>1.84</td>
</tr>
<tr>
<td>2 variance of numbers of neighbours</td>
<td>0 + 2</td>
<td>81 75</td>
<td>1.52</td>
</tr>
<tr>
<td>3 variance of Voronoi cell area</td>
<td>0 + 2</td>
<td>79 73</td>
<td>1.99</td>
</tr>
<tr>
<td>4 Voronoi cell area disorder</td>
<td>0 + 2</td>
<td>78 73</td>
<td>1.92</td>
</tr>
<tr>
<td>5 variance of nearest-neighbour distances</td>
<td>0 + 2</td>
<td>81 63</td>
<td>1.64</td>
</tr>
<tr>
<td>6 autocorrelation</td>
<td>1 + 2</td>
<td>86 81</td>
<td>0.85</td>
</tr>
<tr>
<td>7 pairwise local symmetry</td>
<td>2 + 2</td>
<td>86 83</td>
<td>1.17</td>
</tr>
<tr>
<td>8 centroidal symmetry</td>
<td>2 + 2</td>
<td>87 83</td>
<td>0.97</td>
</tr>
<tr>
<td>9 Delaunay sides entropy</td>
<td>1 + 2</td>
<td>88 81</td>
<td>0.58</td>
</tr>
<tr>
<td>10 Delaunay triangles entropy (preferred)</td>
<td>2 + 2</td>
<td>88 83</td>
<td>0.49</td>
</tr>
</tbody>
</table>

entropy, which achieves perfect ranking performance on set A and within 3% of set B, with an interval accuracy of approximately half a jnd. In the remainder of the chapter we will focus on the Delaunay triangles entropy g-scale, and will refer to it as the g-scale.

We have examined the cross-validated residuals between the g-scale and the p-scale: their mean is small (0.03 jnds); they pass Kolmogorov-Smirnov (Kolmogorov, 1933; Stephens, 1992) and Jarque-Bera (Jarque and Bera, 1980) normality tests ($p > 0.05$); and their correlation with the p-scale values is not significantly different from zero ($p > 0.05$). We conclude that, assuming that the patterns to be studied belong in the same class as set A, g-scale values estimate p-scale values with an unbiased, unstructured random error with a standard deviation of half a jnd. To illustrate what this precision represents we compute that the precision of a-scale values is such that patterns can be ranked according to their approximate order, in agreement with the population consensus, as accurately as a committee of seven independent human rankers. Expressing the performance of the measures in jnd units of interval accuracy allows their direct comparison on a linear scale, which is
3.3. Evaluating Geometrical Measures

convenient for measurement purposes. It also relates the measures in a straightforward way to the human performance, and makes it easy to express the accuracy as a fraction of the whole range of order/disorder. It is worth mentioning that for the amount of collected experimental data, deviance analysis suggests rejection for all candidate \( g \)-scales as valid models for the perception of order at the 0.05 level.

As examples of the performance of the \( g \)-scale, in Table 3.3 the observed preference rates from the psychophysical experiment and the predicted preference rates based on the estimated \( p \)-scale and \( g \)-scale values are shown for the pairs of patterns from set \( A \) which are presented in Figure 3.1.

**Table 3.3:** Preference rates for selected pairs of patterns of set \( A \) (Figure 3.1). Pattern numbers correspond to their estimated \( p \)-scale ranks (Section 3.3.2.1).

<table>
<thead>
<tr>
<th>Pattern pair</th>
<th>Observed (%)</th>
<th>( p )-scale (%)</th>
<th>( g )-scale (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(11, 8)</td>
<td>60 - 40</td>
<td>59 - 41</td>
<td>49 - 51</td>
</tr>
<tr>
<td>(14, 12)</td>
<td>68 - 32</td>
<td>63 - 37</td>
<td>63 - 37</td>
</tr>
<tr>
<td>(7, 4)</td>
<td>80 - 20</td>
<td>71 - 29</td>
<td>76 - 24</td>
</tr>
<tr>
<td>(16, 9)</td>
<td>88 - 12</td>
<td>88 - 12</td>
<td>77 - 23</td>
</tr>
<tr>
<td>(19, 5)</td>
<td>98 - 2</td>
<td>100 - 0</td>
<td>99 - 1</td>
</tr>
</tbody>
</table>

There is a wide range of physically inspired measures of order in literature (e.g. Zhang et al. (2006); Cliffe and Goodwin (2013); Truskett et al. (2000); Sausset and Levine (2011)) that could directly suggest or inspire candidate measures for our study. Some are not straightforward to adapt to our case due to the relative small number of elements in the patterns in comparison to physical systems. Apart from initial ideas and their variations that were examined in pilot experiments and which led to the novel measures presented in this chapter, some additional measures based on electrostatic and Lennard-Jones potentials, Hopkins statistics, radial distribution function and bond order parameters \( Q4 \) and \( Q6 \) were tested; their ranking performance was lower than that of the novel measures we present.
3.3.2.2 Failure cases

The patterns of set $B$ were more numerous ($N_B = 240$) and more diverse than those of set $A$ ($N_A = 20$). We examined the performance of the Delaunay triangles entropy $g$-scale on set $B$ in order to characterize its limitations. For each of the 120 pairs of set $B$, we used an exact binomial test to assess whether the empirical preference rate was consistent with the preference rate as predicted by the $g$-scale. Figure 3.10 shows the histogram of log probabilities of observing the empirical data, or more extreme, given the prediction of the $g$-scale. At a 5% level of significance, 96 pairs agree with the $g$-scale and 24 deviate; but at a Bonferroni corrected significance level of 0.04% (= 5% / 120) 109 agree and only 11 deviate. Figures 3.11 and 3.12 show two representative examples from these 11 cases. Based on informal examination of the failure cases we tentatively offer two possible characterizations: the $g$-scale under-estimates the order of (a) patterns with multiple levels of perceptual organization, for example widely-spaced curves of closely spaced points, or (b) highly-ordered (non-jittered) patterns disrupted by substantial deletion or addition.
3.3. Evaluating Geometrical Measures

Figure 3.10: Histogram of the log-probabilities for the exact binomial test (two-tailed) of the consistency of empirical preferences with $g$-scale predictions for set $B$ pattern pairs. The vertical line represents the Bonferroni corrected significance level. Preference frequencies for the pairs whose probability logarithm lies on the right side of the line cannot be predicted by the Delaunay triangles entropy $g$-scale.

Figure 3.11: First example of poor $g$-scale performance for pair of patterns in set $B$. The $g$-scale under-estimates the order of the left pattern. Experimentally this pattern was chosen as more ordered 70% of the times (20 trials) while the $g$-scale predicts that it would be chosen 10%. The left pattern contains widely-spaced curves of closely spaced points.
3.4. An Absolute Interval Scale of Apparent Order

Since the interval structure of the $g$-scale is preserved by linear rescaling, we can scale it so that certain significant patterns are anchored to memorable values. We will call the result an **absolute interval scale of apparent order** ($a$-scale).

For the lower anchor, we considered patterns arising from two-dimensional spatial Poisson processes, which are random systems of points such that the distribution of counts within any region is Poisson-distributed, and the distributions for disjoint regions are independent. Finite point patterns that arise from such a Poisson process were found to vary modestly in their $g$-scale value. In particular, for patterns of 180 points (as in our stimuli) the standard deviation of the $g$-scale value was 0.29 jnds. We established the $a$-scale so that on average Poisson patterns of 180 points have a value of zero.

For the upper anchor, we consider Bravais lattice patterns which appear as regular arrays of parallelograms (with special cases of squares, diamonds, etc.). For any Bravais lattice, all the Delaunay triangles are congruent so the $g$-scale value attains its maximum possible. We established the $a$-scale so that Bravais lattice patterns have a value of 10.

---

Figure 3.12: Second example of poor $g$-scale performance for pair of patterns in set $B$. The $g$-scale under-estimates the order of the left pattern. Experimentally this pattern was chosen as more ordered 75% of the times (20 trials) while the $g$-scale predicts that it would be chosen 15%. The left pattern contains highly-ordered (non-jittered) structure disrupted by substantial deletion.
Since the difference in \( g \)-scale value between Poisson patterns and Bravais lattices is 10.4 jnds, we estimate that 1 unit on the \( a \)-scale (\( a \)-unit) is equal to 1.04 ± 0.07 jnds. Thus, for practical purposes: 1 \( a \)-unit ≈ 1 jnd. Since the \( g \)-scale agreed with the \( p \)-scale to an RMS accuracy of 0.49 jnds, taking into account the re-scaling of the \( a \)-scale to the anchor points, the precision of the \( a \)-scale is 0.47 \( a \)-units (≈ 0.49 / 1.04). In Figure 3.13, the \( a \)-scale values of a diverse set of patterns are shown.

**Figure 3.13:** Middle row: the proposed absolute scale (\( a \)-scale) of order. Upper row: point patterns of diverse origin (from left to right): (a, b) precursor bristle cells of *Drosophila* (a: early stage, b: later stage), (c) S-cones on the macaque retina (Shapiro et al. (1985), reprinted with permission from Taylor & Francis Ltd, [http://www.tandf.co.uk/journals]), (d) eye-shaped spots on the peacock train (image kindly provided by Meena Uba), and (e) planar-packed metallic spheres of the same radius. Lower row: patterns from set \( A \) (numbers show their estimated \( p \)-scale ranks).

The constructed absolute interval scale of order can be provided as a measurement tool to natural scientists for the analysis of static and evolving patterns. The expression of order in \( a \)-units can improve communication of results and facilitate comparisons. As efficient function does not always coincide with perfect order, a common scale for the quantification of intermediate levels of order can help in the identification of optimal system-specific ranges of order. Also, comparisons of order between different systems can highlight common principles in self-organising systems. The MATLAB code for the computation of the \( a \)-scale
order value of a given point pattern can be downloaded using the following URL of The Royal Society: http://rsif.royalsocietypublishing.org/highwire/filestream/23423/field_highwire_adjunct_files/3/rsif20140342supp4.rar.

3.5 Discussion

The schematic descriptions of the entropy-based measures (Figures 3.8 and 3.9) indicate why their performance is higher than the variance-based measures. When the distribution of a quantity is not unimodal, entropy provides a better measure of variability. Variance, for example, will overestimate variability if the distribution contains two very sharp and well separated peaks.

In the entropy-based measures, the smoothing of the histograms has been necessary for two reasons. The first is that the performance of the measures with respect to the bin size can vary significantly for small changes in the bin size. Figure 3.14 shows how the ranking performance of the Delaunay sides entropy measure changes with respect to the bin size. For comparison, in the same figure, the corresponding performance is shown for the smoothed version of the measure for varying kernel size. It is obvious that the smoothed version presents a more stable performance for the equivalent range of sizes. The second reason for which smoothing is preferred is that the estimation of entropy suffers from bias when the numbers of items in the histogram is low (Bonachela et al., 2008). Hence, for point patterns with fewer points, the computed entropy can be severely underestimated. Various approaches for correction of this bias have been proposed (Bonachela et al., 2008; Schürmann, 2004). To estimate how strong the effect is in the case of our entropy-based measures, we conducted simulations where the generated point patterns differed in the amount of positional jitter and in the number of points. For the non-smoothed histograms, the low number bias became noticeable for patterns of around 150 points. This bias could not be adjusted using the Grassberger and Bonachela corrections (Bonachela et al., 2008; Schürmann, 2004). In general, for all measures that rely on entropy estimation based on histograms, the best approach has been to smooth
the histogram with a kernel of an optimized size. With the smoothed histograms approach the bias in the entropy estimation became very small even for numbers of points as low as 25.

**Figure 3.14:** Graphs show the ranking performance of the Delaunay sides entropy measure with the entropy computation based on the simple (left) and the smoothed (right) histogram for varying bin and kernel sizes respectively. The performance of the smoothed version is less sensitive to changes of the size parameter. In order to make a valid comparison, for the same range of the size parameter, the width of a rectangular pulse (for the simple histogram) and that of a Gaussian kernel (for the smoothed histogram) of the same area have to be matched. This is done by adjusting their widths so that the mean squared error between the two shapes is minimised. The computed pulse size is 2.9 times the standard deviation of the Gaussian curve. This gives a good match visually as well.

Similarly to texture analysis approaches (Ojala and Pietikäinen, 2001), where pixels are the focus, we can classify measures of order according to the complexity in the relationships between points they take into account. The simplest approach, takes into account first order properties of the points, i.e., considers each point separately ignoring spatial interactions. Density of points is a first order property. Often, local first order properties are compared at a larger scale and spatial correlations can be identified. A suggested measure (Dunleavy et al., 2012) based on this principle uses mutual information between densities of types of particles in different regions of the pattern for the quantification of order in amorphous materials. Measures taking into account second order properties analyse relationships between pairs of points (e.g. distances or orientations of the linear segments joining them). Most commonly, nearest neighbour relationships are considered. Our preferred g-scale makes use of third order relationships; Delaunay triangles are defined
3.5. Discussion

by the relative spatial positions of triads of points. Still, the analysis of the Delau-
nay triangles’ properties by splitting them into unidimensional histograms does not
perform as well as the Delaunay triangles entropy. A particular example can high-
light a reason for the lower performance: A mild warping applied on an initially
regular pattern will cause spread in the triangles’ one dimensional histograms. The
pattern, however, will look regular since the local variation will be smooth and the
shapes of the Delanay triangles will remain similar across the pattern. Considering
the sharp bivariate shape histogram, our preferred measure manages to overcome
this difficulty and nevertheless identifies this pattern as relatively ordered. Extend-
ing this method by including more complex relationships between points without
disentangling them into separate dimensions seems a reasonable approach. Such a
technique would be able to capture more complex levels of order. However, for the
given number of points (180) a higher dimensional space for such features will be
sparsely populated.

In the proposed absolute scale of order, when we estimated the position of the
zero anchor point that corresponds to the average Poisson pattern, we only needed
to compute the mean value and the standard deviation of the optimized preferred
g-scale for a set of randomly generated patterns (10,000 point patterns). Although
we are only interested in these two values, it is tempting to visually examine the his-
togram of the a-scale order values for the whole set of generated patterns. There is
no reason we should expect this to follow a specific distribution; the characteristics
of the distribution (e.g. skewness) are determined by the combination of the raw
values of the geometrical algorithm and the monotonic transformation that aims to
match the perceptual scale. Interestingly though, as Figure 3.15 shows, the result-
ing histogram approximates very well the Gaussian distribution. Application of the
Kolmogorov-Smirnov test (Kolmogorov, 1933; Stephens, 1992) for normality gives
a p-value of 0.74, which verifies the good visual match. Raw values of measures
do not, in general, exhibit this behaviour. For example, the values of the variance of
the nearest-neighbours distances, for the same set of patterns, gives a skewed his-
togram. The Kolmogorov-Smirnov normality test in this case results in a p-value of
4.12 \cdot 10^{-5} and thus the hypothesis of Gaussian distribution is rejected. The symmetry of the histogram around the zero point of the $a$-scale means that the degree to which a pattern is distinguishable from the central ones is associated with the probability of its $a$-scale value (at either side). The method for the construction of the perceptual scale has not made use of comparisons between Poisson patterns and therefore the symmetry, at least, around zero is surprising. Further research with the use of dense sampling of patterns around the zero point of the $a$-scale could verify whether this symmetry is linked to real perceptual symmetry and also examine whether the Poisson patterns with the most common $a$-scale value of zero have special perceptual properties, e.g., if they can be identified as the most random.

![Figure 3.15: Histogram of $a$-scale values for 10,000 randomly generated Poisson patterns. The theoretical fitted Gaussian curve fits very well the distribution.](image)
3.6 Conclusion

Using pairwise comparisons of the order of point patterns we established that: there is a strong consensus in the perception of order, these perceptions have the structure of an interval scale, and that scale extends over roughly 10 just-noticeable differences (jnds) between disorder and order.

We compared existing and novel geometric methods for quantifying order. The mismatch between existing methods and perception was 0.85 jnds. The novel Delaunay triangles entropy method achieved a mismatch of less than 0.5 jnds. This method works by quantifying the variability of the spaces between points as captured by the size and shape of the Delaunay triangulation of the points.

We proposed an absolute interval scale (a-scale) for order based on the Delaunay triangles entropy method, but scaled so that Poisson disorder has on average a value of zero, and perfect Bravais lattice order has a value of 10. Each a-unit of this scale is approximately 1 jnd for human perception. The RMS precision of a-scale values, as determined by our method, is 0.46 a-units. Our scale quantifies order with an agreement to the consensus as accurate as a committee of seven raters, while being fully objective and repeatable. Its scope is limited to point patterns that do not present multiple levels of organization.
Chapter 4

Application of the Absolute Scale of Apparent Order

4.1 Introduction
A major goal of the approach described in Chapter 3 was to develop a scale of order that would be of widespread utility in the study of biological patterning during development and disease. Although a variety of quantifications exist in literature, as presented in Chapters 2 and 3, none of these have been tested for agreement with visual perception for intermediate levels of order and for diverse point patterns. Whether the human visual system has been evolved in such a way that apparent order presents distinguishing advantages as a basis for scientific analysis is difficult to prove. However, such quantification can be proposed as an additional tool to natural scientists with the hope that for particular classes of evolving systems and ranges of intermediate order it will exhibit higher sensitivity. This can only be verified if our measure is widely accepted and compared for its performance with alternative approaches. In this chapter we illustrate the application of the absolute interval scale of apparent order (a-scale) we constructed to a selected biological system. The aim is not to examine its performance in contrast with the other approaches, but to show that, in principle, the proposed method can provide valuable results. The particular system we study is the dorsal thorax (notum) of the developing fruit fly (Drosophila melanogaster). A brief presentation of this system and the mechanisms that are related to its specific patterning follows.
The notum of Drosophila melanogaster as a model system. The notum of *D. melanogaster* is considered a model system for the study of development (Cohen et al., 2010). It is ideal for this type of analysis since it is simple and experimentally tractable, and has been used for nearly 60 years in studies of pattern formation, its genetics and its evolution (Gómez-Skarmeta et al., 2003). The pace of its development is such that observation of cellular dynamics and gene expression changes can be achieved in real time by employing novel live imaging techniques which do not interfere with the developmental process (Marinari et al., 2012; Georgiou and Baum, 2010).

Bristle cell patterning. In the fully developed fly, the notum is covered by a population of short mechanosensory bristle cells (*microchaetes*) arranged in a fairly regular array; something that is likely to contribute to organismal fitness (Wootton, 1992; Usui-Ishihara and Simpson, 2005). At an earlier stage of development (12 h after pupa formation (APF)) the notum tissue is clearly identifiable but is in a state of relative disorder without an established pattern (Figure 1.2). A process of cell competition through lateral inhibition, based on Delta-Notch signalling, drives the system towards a patterned final state in which each Delta-expressing bristle precursor cell is surrounded by cells with active Notch signalling. The activation of intracellular Notch signalling by the membrane-tethered Delta protein requires cell-cell contact which is accomplished by the action of dynamic, basal actin-based protrusions (*filopodia*). This intermittent signalling by transient filopodia that can extent several cell diameters is required for the observed spacing of the bristle cells (de Joussineau et al., 2003; Cohen et al., 2010). Moreover, it has been proposed that the final well-ordered state is reached through a process of gradual pattern refinement which relies on the stochastic character of the filopodia contacts (*structured noise*) and fluctuations in the expression of the inhibitory signals (*temporal noise*) (Cohen et al., 2011). This phenomenon is similar to solid-state phenomena of grain growth with annealing, where high temperature enables movement of crystal defects (e.g. vacancies and dislocations) and leads to a more ordered configuration of lower energy (Miodownik, 2001; Cohen et al., 2011). During the final stages of
4.1. Introduction

In this refinement process, the fate of each cell, i.e., whether it will remain a normal epithelial cell or it will give rise to a mechanosensory organ (bristle), is decided.

**Epithelial tissue organisation.** In parallel to the progressive arrangement of the precursor bristle cells in the regular final array, another process of spatial organisation is observed. At 12 h APF epithelial cells, as defined by their apical cell-cell junctions, appear disordered with the midline of the notum exhibiting a more anisotropic geometry than the rest of the tissue. During development, cycles of division, shape change, and neighbour exchange drive finally (26 h APF) the system to a well-ordered packing which approximates a hexagonal lattice (Fristrom, 1988; Marinari et al., 2012). These processes induce a significant topological rearrangement whereas overall form and size of tissue are not affected (Cohen et al., 2011). Two mechanistically distinct processes with different time scales have been identified recently which contribute to the tissue refinement (Marinari et al., 2012). The final stage of both is a selective removal of cells from the epithelial tissue (*delamination*). In the one process, a proportion of cells first undergoes junctional loss and then after a progressive loss of apical area is pushed out by their neighbours. Whole process takes place over an extended and variable period of time. Simulations based on mechanical forces showed that crowding drives cells to compete for the limited space by rearranging junctions and finally cells with smaller apical area are squeezed out (Marinari et al., 2012). In the second process, cells only undergo a gradual apical area loss without changes of their neighbours. Delamination appears more frequently at areas with cellular anisotropy (midline) and thus facilitates transition to a more ordered arrangement of cells. Again noise, in the form of stochastic fluctuations in junctions, is necessary to trigger neighbour exchanges. These delamination processes are of great importance as they balance variations in cell growth (Marinari et al., 2012).

A quantitative analysis of the evolution of the spatial organization of the notum’s tissue during development is important for biologists. Results can be informative about (i) self-organising developmental mechanisms that lead to healthy
organisms, and (ii) the robustness of these processes with respect to genetic and environmental manipulations.

In the following we apply our method of quantification of order, first, to the bristle cells pattern formation and, second, to the whole set of the background epithelial cells. We use these two systems to demonstrate the applicability of our method. Next, we apply two different measures of order to simulated point patterns to demonstrate inconsistencies and so emphasise the necessity of a validated measure for scientific use.

4.2 Methods
Imaging is by confocal laser scanning microscopy. Neuralised-Gal4, UAS-Moesin-GFP (Neu:GFP) was used as a marker for the Delta expression of bristle precursor cells, and ubiquitously expressed E-Cadherin-GFP was used to visualise apical cell-cell junctions. For live imaging, a window was cut in the pupal case; consequently the earliest time that could be imaged was 12 h APF. Confocal laser scanning microscopy provides depth selectivity which can be an advantage in case depth sectioning is required; on the other hand, this translates in high focus sensitivity along the \( z \)-axis. Therefore, it is possible that during live imaging the quality of the acquired images becomes impaired as the surface of the scanned tissue is subject to movement, and does not remain precisely vertical to the \( z \)-axis. After the scanning procedure, for each time point, the best image among those of a stack along the \( z \)-axis is selected. In some cases, two or more images need to be combined for better quality. From 12 h APF, images were acquired roughly every 30 minutes for ten hours and assembled into a movie. Each movie frame appeared as a mosaic of epithelial cells with bright borders (Figure 1.2). Roughly 95% of these cells express no fluorescent protein, causing them to appear dark, while the remaining 5% exhibit varying degrees of lightness due to differing levels of proneural gene expression. During the movie, the cellular mosaic continuously changes, and new cells appear through division; as cells start to express markers of bristle fate they become visible (becoming potential bristle precursor cells), and cells expressing the
fluorescent marker can lose this expression. Although the distinction is not sharp, the fluorescently-labelled cells can be segregated into dim and bright classes. Over the course of the movie, a number of cells transit between the dark, dim and bright states. Bright cells at the end of the movie are those that will become bristle cells; dark and dim cells will become ordinary epithelial cells. Therefore, the process of bristle patterning involves the refinement of the pattern of dim and bright cells. Four movies (A, B, C and D) were kindly provided to us from the Baum Lab at UCL. Movie A was provided to us first and was of higher quality than the ones that followed. Analysis of the bristle cell patterning was conducted on all four movies, while analysis of the background epithelial tissue was conducted only on movie A, as this was the only one where cell-cell junctions were fully discernible.

4.2.1 Bristle Cell Patterning

To generate point patterns for the bristle patterning analysis we first localised cells expressing the fluorescent marker, and then classified them as dim or bright. For the localisation, the open-source software package imageJ with the Particle Picker plugin was used (Abràmoff et al., 2004) for manual marking of the cells. For the classification, we developed an automatic algorithm based on the initial dataset (movie A). The first step of the algorithm includes a brightness histogram matching for all frames, since the overall brightness may vary from frame to frame. The second step includes the comparison of the local brightness in the area of the centroid of the cell with the local background brightness. This is done to compensate for any spatial variation in the background brightness which appears because the surface of the tissue is not absolutely flat and perfectly aligned with the plane of the focus of the laser microscope. Two parameters are used in this algorithm: (i) the number of pixels selected from a square of a fixed size centred around the centroid (inner area) whose mean value represents the cell’s brightness, \( b_{\text{cell}} \), and (ii) the size of a larger centred square (outer area) whose median value of brightness represents the local background brightness value, \( b_{\text{background}} \). The relative brightness \( b_{\text{rel}} \), which is the output of the algorithm, is the difference of these two values, \( b_{\text{cell}} - b_{\text{background}} \).
We can train the algorithm using more than two classes of brightness. The two parameters are tuned by maximising the Kendall tau rank correlation coefficient, which is a measure of concordance between the groups of the ground truth and the output of the algorithm. This way of evaluating the performance of the classifier has been extended by Obuchowski for the case of non-binary classifiers (Hanley and McNeil, 1982, 1983; Agresti, 2010; Obuchowski, 2005). The threshold values for each class are selected so as to minimise misclassification cost. An advantage of the automated method is that groups of cells of different brightness ranges can be effortlessly selected in a continuous way by adjusting the threshold parameters. Figure 4.1 shows the results of the algorithm’s classification for two frames when three levels of brightness are considered.

![Figure 4.1: Automatic classification of potential precursor bristle cells into three levels of brightness: dim (blue), intermediate (green), bright (red) for an earlier (left) and later (right) stage of development. The number of bright cells is increased at the later stage.](image)

After examining the whole set of movies, we favoured manual classification, instead of automatic, as variation across movies in brightness, contrast and blur as well as in cells’ size and shape was large. A graphical user interface was designed for higher efficiency. While localisation was done with full frame viewing, classification was done by viewing individual cells clipped out of the image along with their immediate contexts in a randomised order for all frames of a movie. This was decided so that the viewer was not influenced by the positioning of a cell within some wider pattern and/or the stage of development. A reference cell was displayed next to the clipped image so as to allow direct comparison of brightness levels and
thus avoid drifting of threshold over time. Figure 4.2 shows the results of the manual classification into *dim* and *bright* cells.

![Figure 4.2: Manual classification of potential precursor bristle cells. Left: original image, right: depiction of the result of the classification into *dim* (blue) and *bright* (red) classes.](image)

### 4.2.2 Epithelial Tissue Organisation

A complete analysis of the order of the epithelial tissue would require development and validation of measures of order appropriate for polygonal arrays. This is an interesting line of research but is outside the scope of this work. Here, we evaluate the order of the background epithelial tissue taking into account only the location of the cells. We use the centroid of a cell to define its location. Given that all three main processes of tissue rearrangement (neighbour exchange, cell division, and delamination) affect the location of cells, we expect that analysis based on location, although partial, captures an essential aspect of the degree of organisation of the tissue.

As explained in the introduction of Section 4.2, for the analysis of the background epithelial tissue only movie A was used. Again, the open-source software package imageJ with the Particle Picker plugin was used for the manual marking of the centroids. In contrast to the bristle cell localisation, the distances between centroids are not significantly larger than the size of cells. This could introduce
considerable variation in the marking, and the subsequent computations. To get an estimate of the method’s robustness and of its possible effect on the derived curves of order, we applied the manual localisation process to a randomly chosen image-frame twice, and computed the order values for both. The difference was $3.1 \cdot 10^{-2}$ $a$-units, which is a very small number. For comparison, the observed changes in the curves were at least $0.5$ $a$-units. Thus, we conclude that the manual process of centroid marking is adequate for the analysis.

4.2.3 Region of Interest, Density, and Edge Effects
Before applying the $a$-scale of order on specific patterns derived from real data, two important issues must be addressed. The one is how to properly define the region of interest (ROI), and the second how to treat the points which are near the edge of this region.

4.2.3.1 Region of Interest and Density Estimation
In order to remove scale dependency in algorithms estimating order, the density of the pattern is required. The naïve approach of estimating density by dividing the number of points by the area of the pattern is not adequate, as the question of what exactly the boundaries of a biological point pattern are, is not trivial. The boundaries of the region of interest have to be determined in a way so that the estimation of density is not biased, especially when the number of points is small.

The choice of the convex hull of the points as the boundary of the pattern seems reasonable, but does not take into account the omitted area around the boundary points. A fix would be to keep convex hull as boundary but when estimating density instead of the actual number of points use an effective one. Simulations showed that low number bias is reduced when we count interior points with a weight of 1 and border points with a weight of 0.5, and then subtract one unit to compensate for the equivalent $2\pi$ radians for points in the “corners”. As border points we consider those on the convex hull and as interior points the rest. Considering as border points not only those on the convex hull, but also those whose Voronoi cell’s cen-
troid is outside the convex hull gives better results. Additionally, since the number of convex hull points may depend on the specific arrangement and on the degree of order of the pattern, the latter approach avoids a possible systematic error in the estimation of order. Figure 4.3 shows an example Poisson pattern and the two options for the characterization of border points. The one adopted matches better to what we would visually consider “border” points of a pattern.

![Figure 4.3](image_url)

**Figure 4.3:** Figure demonstrates, for a randomly generated pattern of 50 points, border and interior points according to the two approaches described in Section 4.2.3.1. Left: only points on the convex hull are considered border points. Right: points on the convex hull but also those whose Voronoi cell’s centroid is outside the convex hull are considered border points. Second approach captures as border points also those close to the convex hull and is less biased in the estimation of density. The effective number of points is 44.5 for the approach shown on the left, and 41 for the approach shown on the right.

### 4.2.3.2 Edge Effects

In point pattern analysis literature (e.g. Perry et al. (2006)) the problem of edge effects arises because the theoretical distributions assume boundless area while the analysed patterns are obviously bounded. For example, statistics which estimate numbers within a circle of specific radius will contain fewer points when the circles exceed the bounded area. Similarly, for order estimation, geometrical algorithms have to take into account edge effects. Three methods are used in point pattern analysis for treating edge effects: (i) the weighted edge correction (Ripley, 1977), according to which less weight is given for pairs of points \((i, j)\) when the circumfer-
ence of the circle centred at point \(i\) and passing through point \(j\) is not fully contained in the bounded area, (ii) the toroidal wrap (Yamada and Rogerson, 2003), according to which the up and bottom parts are joined together and the same is done for the left and right parts, and (iii) the guard area correction (Yamada and Rogerson, 2003; Perry et al., 2006), where the points in this area, which can be internal or external to the region of interest, are considered only as ‘destinations’ in the measurements of distances. The first approach is not preferred because taking into account points with varying weight is not applicable for all measures. The second approach cannot be applied on arbitrary shapes of boundaries and the resulting pattern will not be perceptually comparable to the initial pattern. Therefore, we adopt a modified version of the guard area correction. The guard area is external to the region of interest and defined indirectly as the area which contains the points that are considered border points according to the definition of the previous section. Although in this chapter we mainly present applications of the \(a\)-scale, this approach is also adjusted to application of different geometrical algorithms: (i) for measures that use inter-point distances, the points in the guard area are included only as ‘destinations’ of distances, (ii) for those that require Voronoi tessellation, the cells whose generators are in the guard area are not taken into account (so, even for measures that consider the Voronoi cells of neighbours, these neighbours are neglected), and (iii) for measures based on the Delaunay triangulation, a triangle with all three points in the guard area is excluded.

### 4.3 Bristle Cell Patterning

#### 4.3.1 Results

In this section we present the application of the \(a\)-scale to the measurement of order of the Neu:GFP expressing cells (\(dim\) and \(bright\) cells) during development. We first examine their numbers and then the timecourses of order.

As shown in Figure 4.4 (left panel), the number of \(dim\) plus \(bright\) cells averaged over the four movies increases over time until around 18 h APF where it
4.3. Bristle Cell Patterning

reaches a plateau. The mean number of bright cells increases during the whole process with the slope approaching zero near the end (22 h APF). The proportion of bright cells (Figure 4.4, right panel) shows a smooth increase through the whole duration. In the initial stages of development (15 h APF), the number of bright cells is roughly equal to the dim cells, and in the final stages (22 h APF) more than 90% of the cells expressing Neu:GFP are bright.

Figure 4.4: Timecourses of cell numbers (left) and proportion of bright cells (right), averaged over four movies.

Figure 4.5 shows the timecourses of the $a$-scale values of order for the bright, bright plus dim, and random subsets for all four movies, along with their regression lines. The random subsets are chosen from the dim plus bright cells, equinumerous with the bright cells. The $a$-scale values plotted are the average for ten random subsets at each timepoint. The quality of the linear fits was assessed separately for each fit using a $\chi^2$—test, taking into account the 0.46 units precision of $a$-scale values, and the number of parameters of the fit. The sum of the squared residuals between the predicted values based on the linear fit and the real values was computed with variance estimated from the 0.46 units of precision. The number of degrees of freedom was corrected by the two parameters of the linear fit. Eleven of the twelve linear fits pass the $\chi^2$—test at a 95% significance level; the twelfth, the fit to the bright cells of movie D (Figure 4.5, top-right), just fails, but passes after
the removal of one of its 17 points. Figure 4.5 shows that all four movies displayed similar slopes of order for different classes of cells: The slopes corresponding to the bright population are considerably greater than those of the bright + dim or random population.

![Figure 4.5: Data points and linear fits for order versus time for the four movies for (a) bright cells, (b) bright + dim cells, and (c) the mean of random selections of cells equinumerous with the bright cells in each image.](image)

### 4.3.2 Discussion

To get a clearer view of the typical behaviour we combined the data from all four movies and computed new linear fits which are shown in Figure 4.6. This should be viewed in conjunction with Figure 4.7 which shows point patterns for one of the movies at indicated times (arrows in Figure 4.6).

Figure 4.6 shows that the population of bright + dim cells starts at an $a$-scale level of 3.0 at 15 h APF, which is a modest level of order but distinctly different from perfect disorder, and rises to 3.5 over the following 8 h. While the increase in
4.3. Bristle Cell Patterning

Figure 4.6: Linear trends of order when all movies are taken into account for bright, bright + dim, and random selection of all cells.

order of only 0.5 $a$-units is small, Figure 4.4 (left) shows that this is achieved while increasing the number of these cells (by 43%). Adding points randomly to a pattern will decrease its order, so it seems that there must be processes that cause the new cells to appear in non-random positions and/or processes that cause the bright + dim population to re-arrange to maintain their order. Observation of the movies reveals little re-arrangement so the non-random emergence of new bright + dim cells is the prevalent mechanism for maintain order in this population even while it increases in number. This process can be observed in the middle row of Figure 4.7.

Turning to the bright cells, Figure 4.6 shows that they start out at a lower level of order than the bright + dim cell population of which they are a subpart, but their order increases so that by 18 h APF it has reached the same level as the bright + dim cells, and by 23 h it has exceeded it by a full two $a$-units. This behaviour can be understood by comparing the timecourse of the order to that of the random cells. This comparison shows firstly that the initial low level of order of the bright cells
Figure 4.7: Locations of cells in movie A at times 15, 18.5, 22 h APF, which correspond to the arrows in Figure 4.6. The number above each box is the order in a-units.

is compatible with them being a random subset of the bright + dim (Figure 4.7 top-left), and secondly that the increase in the fraction of bright + dim that are bright would account for only a modest increase in the order of the bright cells, much less than that observed (compare top-right and bottom right of Figure 4.7). Therefore, the observed large increase in order of the bright cells is a result of a mechanism selecting only the ‘right’ cells from the population of bright + dim.

In summary, the timecourses of order are consistent with two mechanisms operating in parallel. One mechanism adds cells to the bright + dim population while maintaining its order; the other mechanism changes cells from dim to bright and
vice versa so that the *bright* sub-population increases in proportion and level of order.

Thus, using the \(a\)-scale of order, we have been able to decipher a two-step process by which the pattern of bristles is refined during the generation of biological order.

### 4.4 Epithelial Tissue Organisation

#### 4.4.1 Results

In this section we present the application of the \(a\)-scale to the development of the fly’s notum epithelium. As in Section 4.3, the analysis is made with respect to cell numbers and centroids.

Figure 4.8 shows the number of the background epithelial cells between 12.5 h and 21.5 h APF. Cells remain relatively constant in number (\(\approx 700\)) from the beginning till 15 h APF, when the effect of division appears. Between 15 h and 18 h APF the number of cells doubles plateauing at slightly below 1400. The number of *internal* points/cells, and not their actual number, is plotted since the actual number of cells presents larger fluctuations due to uncertainties in the marking of cells close to the border.

Figure 4.9 shows the change of order over time. It is evident that, overall, order increases. The initial value is 5.5 \(a\)-units at 12.5 h APF, and near the end (21.5 h APF) it reaches a value of approximately 6.9 \(a\)-units. A transient drop of almost half an \(a\)-unit is noticed which coincides with the beginning of the division. Order starts to recover while still division takes place and increases monotonically with decreasing rate till the end (21.5 h APF).
4.4. Epithelial Tissue Organisation

Figure 4.8: Epithelial tissue. Number of cells during development increases from $\approx$700 to slightly below 1400. Increase takes place during cell division, from 15 h to 18 h APF.

Figure 4.9: Epithelial tissue. Overall order increases during development (1.4 $a$-units), with a transient drop (0.5 $a$-units) coinciding with cell division.
4.4.2 Discussion

Application of the $a$-scale to the epithelial background cells showed an overall increase in order during development and revealed a transient drop which coincided with cell division. This non-monotonic change in order during development shows that although division causes a significant disruption, it allows fast recovery of the system. This suggests that division does not impact crucial aspects of the pattern’s organisation, or furthermore that it even facilitates the transition to the well-ordered final state, as after the initial drop, the increase rate becomes very high ($\sim 0.7$ $a$-units per hour). This hypothesis is consistent with mechanical simulations where the system’s capacity for junctional rearrangements enhances its final order (Mari­nari et al., 2012).

Additionally to average order, we are also interested in estimating its spatial distribution and thus obtain information about the scale of the processes. Specifically in the fly’s notum, the difference in the degree of organisation between the midline and the rest of the tissue highlights different mechanisms of transition to order through the two distinct mechanical delamination processes (Mari­nari et al., 2012). To acquire a spatial visualization we apply a modification of the pairwise local symmetry measure that was presented in Section 3.3.1.2. Although this measure was not the best performing, its performance was relatively high. By assigning to each cell a number between 0 and 1 expressing its local symmetry, we can acquire a representation of order across space and over time. The numerical value is computed as the average overlapping factor (intersection divided by union) for each Voronoi cell over all its adjacent cells (without the power law transformation employed in the pairwise local symmetry measure). A value of 1 indicates perfect overlap and thus perfect local symmetry around the cell. Low numbers imply local asymmetry. In Figure 4.10 the spatial distribution of local symmetry, is depicted for two images corresponding to early and later stages of development. We notice that view is similar to that described in Marinari et al. (2012), where the more informative polygonal array is analysed using a cells’ anisotropy measure (long/short axis ratio). Figure shows that at the early stage of development the tissue is dis-
ordered with the midline area being less ordered than the rest of the tissue. Later in development, the midline is less distinguishable and average order is considerably increased. It is additionally noticed that for disordered areas of the tissue the discrepancy between the epithelial polygonal array and its centroid-based Voronoi tessellation is apparent, while for more ordered areas the two meshes almost coincide.

Figure 4.10: Spatial representation of local symmetry for the fly’s notum epithelial tissue. Top row: microscopic images of the epithelium at early (left) and later (right) stage of development. Midline appears horizontally, approximately in the middle of the microscopic images. In left pattern, midline is more distinguishable as it is more anisotropic than the rest of the tissue. Overall, pattern on the right presents a more ordered, well-packed, arrangement. Bottom row: corresponding Voronoi tessellations based on the centroids of the epithelial cells, coloured according to average overlap score. Left pattern appears more yellowish than right pattern (see colour map legend) and less uniform with top and bottom horizontal areas more reddish than middle.

Application of the $a$-scale, as well, can offer spatial information with respect to order, but at a coarser resolution, provided that the number of cells is large enough so that parts of a pattern can be analysed separately. Given the symmetry of the
tissue along the midline, two straightforward ways of partitioning the images into simple rectangular subparts can be employed. The first is to consider vertical stripes, orthogonal to the axis of symmetry; these will be similar in terms of portions of the midline and the outside of the midline tissue. The second is to consider horizontal stripes, parallel to the axis of symmetry. In this case, given the dimensions of the midline, only one stripe will contain it, while the others will contain the rest of the tissue. We present both of these analyses as each highlights different aspects of the spatio-temporal evolution of order.

In the vertical partitioning each subregion includes areas of different degree of organisation and thus each is practically representative of the whole. Since the partitioning is not biased with respect to order, this analysis allows to examine whether the observed transient drop described in section 4.4.1 is a real effect or the result of random fluctuations. Partitioning of four stripes was chosen as a good balance between spatial resolution and sufficient numbers of cells in each subregion. Stripes were numbered 1-4, right-to-left. Widths were adjusted so that each subregion contained roughly equal number of points and were kept fixed for all frames of the movie. Figure 4.11 shows the number of cells of each sub-image (cell division is evident) over time (12.5 h to 21.5 h APF) and Figure 4.12 shows the changes in order for the same period. Order appears to increase in all curves, and the transient drop during division appears consistently in all of them. This provides support that the effect is real. Moreover, it is noticeable that the drop (local minimum) for each subregion is temporally shifted. The sequence of delays is in agreement with the similar sequence of delays of the onset of division shown in Figure 4.11. Considerable increment in numbers is observed first in subregion ‘1’ (right side of the tissue), ‘2’ and ‘3’ follow, and subregion ‘4’ is the last one (left side of the tissue).

This regular pattern of change in order appearing sequentially in adjacent sub-regions is compatible with the division appearing as a wave on the tissue (Marinari et al., 2012). Similar waves of division that facilitate the formation of highly ordered tissue have been observed during development of the *Drosophila* ommatidia
4.4. Epithelial Tissue Organisation

**Figure 4.11:** Epithelial tissue spatial analysis (vertical partitioning). Number of cells during development for the vertical subregions 1-4 (numbered right-to-left).

**Figure 4.12:** Epithelial tissue spatial analysis (vertical partitioning). Order during development for the vertical subregions 1-4 (numbered right-to-left). Order increases with a transient drop, time-shifted for each subregion.
(Ready et al., 1976) and during cone differentiation in the zebrafish retina (Hu and Easter, 1999).

Therefore, with spatial analysis based on the $a$-scale we have been able to not only detect a significant transient drop of order during development, but also to identify a pattern of a moving wave of disorder, which, after a temporary disruption, leaves behind tissue of higher organization.

For the horizontal partitioning we examined the simplest option, i.e., three sub-images. The middle sub-image was centred around the midline and similarly to the vertical case, widths were adjusted so that each contained roughly the same number of points and kept constant for all frames. Figure 4.13 shows the number of cells for each subregion over time (12.5 h to 21.5 h APF), cell division is evident, and Figure 4.14 shows the changes in order for the same period. We notice that the curves of order for the upper and lower horizontal subregions are very similar. Order for both is consistently higher than that of the midline, presenting a slow increase which gets interrupted by a small drop during division. The order of the midline is almost steadily increasing from approximately 4.5 $a$-units (12.5 h APF) to around 7 $a$-units near the end (21.5 h APF) approaching the level of order of the other two subregions.

Thus, with the use of the $a$-scale we have been able to detect differences in order between the midline and the rest of the tissue as well as their dissimilar rates of increase. Results are consistent with the different delamination rates reported in Marinari et al. (2012).
4.4. Epithelial Tissue Organisation

Figure 4.13: Epithelial tissue spatial analysis (horizontal partitioning). Number of cells during development for the upper (blue) and lower (red) horizontal subregions, and the midline (green).

Figure 4.14: Epithelial tissue spatial analysis (horizontal partitioning). Order during development for the upper (blue) and lower (red) horizontal subregions, and the midline (green).
4.5 Does the Choice of Order Measure Matter?

Often in the analysis of physical systems, different measures of order (regularity, anisotropy, etc.) correlate very well (e.g. Guillaud et al. (2004); Kapfer et al. (2010)). It is wrong to conclude that the correlated measures are equally valid. In this section we demonstrate that such correlation does not provide support for a legitimate generalisation of their applicability. To do this we present a series of simulations (four cases) where two algorithms quantifying order may agree or radically disagree depending on the exact transformations that an evolving point pattern undergoes. The two measures we compare in this analysis are the raw values of the nearest-neighbour distances variance and the Delaunay triangles entropy measures (i.e., without the monotonic mapping to the perceptual scale described in Section 3.3.2.1).

To keep a common perspective across the four simulations, it is convenient to imagine the sequence of instances of a point pattern during transformations as a path in a multidimensional point pattern space starting from an initial point A (initial state of the system) and ending at a final point B (final state). To make it easier for the reader to follow, each simulation case is presented along with the corresponding figure on the same page.
4.5. Does the Choice of Order Measure Matter?

Figure 4.15: Quantification of order for the transformation described in simulation case I. The two measures, nearest-neighbour distances variance (left) and Delaunay triangles entropy (right), are highly correlated. (The sign of the algorithms’ output has been inverted so that high values correspond to high order.)

Simulation case I. In this simulation, the initial point pattern is a perfect square lattice of 400 points (array of 20 × 20 points). At each time-step the position of the points is randomly jittered. This is done with the addition of independent random numbers drawn from a uniform distribution to the x-y coordinates of the points. Figure 4.15 shows the order curves of the two algorithms for thirty steps. Independently of the exact numerical values and ranges, when appropriately scaled, the curves show a similar change in order throughout the whole process and the outputs are highly correlated. Therefore, if a real system was evolving along this restricted path in pattern space, the two measures would quantify the progression to disorder similarly.
4.5. Does the Choice of Order Measure Matter?

Figure 4.16: Quantification of order for the transformation described in simulation case II. The two measures, nearest-neighbour distances variance (left) and Delaunay triangles entropy (right), for the same sequence of transformations show opposite change. The arrows depict the change between initial and final states. (The sign of the algorithms’ output has been inverted so that high values correspond to high order.)

Simulation case II. This simulation is based on two point patterns of the psychophysical experiment described in Chapter 3. These are the patterns numbered ‘12’ and ‘15’ from set A, shown in Figures 3.1 and 3.13 respectively. We construct a path in the point pattern space starting from pattern ‘12’ (pattern A), and ending at pattern ‘15’ (pattern B). We then plot the evolution of order along this path for the two measures. These patterns have been generated from the same algorithm with the use of different parameter values. For each parameter $q$, the two different parameter values $q_A$ and $q_B$ define an interval. By partitioning this interval into a large number of equal subintervals for each parameter we define intermediate states that form a path connecting patterns A and B. Figure 4.16 shows the order curves derived by the two measures when pattern A is transformed into pattern B following a sequence of twenty step-transformations along the above-described path. In both graphs, order initially increases and after four time-steps decreases. However, when comparing initial and final states, the nearest-neighbour distances variance shows an overall decrease in order while the Delaunay triangles entropy shows an increase. Hence, in this transformation the two measures disagree.
4.5. Does the Choice of Order Measure Matter?

![Figure 4.17: Quantification of order for the transformation described in simulation case III.](image)

For the three types of transformation, the relative slopes given by the two measures, nearest-neighbour distances variance (left), and Delaunay triangles entropy (right), are different. (The sign of the algorithms’ output has been inverted so that high values correspond to high order. Pattern instances are plotted every two steps for better visibility.)

**Simulation case III.** In this simulation, we apply three different types of transformation in succession. The initial pattern A is a perfect square lattice of 400 points (array of 20 × 20 points), same as in case I. Each type of transformation is applied for ten steps. We first apply a small positional jitter on the positions of the points (same as in case I), secondly a non-uniform stretching along the x-axis, and thirdly a random removal of points. The non-uniform stretching is implemented by displacing the points towards the x-axis direction at each time-step by an amount which is proportional to the square of their x-coordinate. Figure 4.17 shows the corresponding order curves for the two measures. The different types of transformation are depicted in distinct colours. We notice that for the three types of transformation the relative slopes are different. This cannot be corrected with a simple rescaling. Having not matched the two algorithms at the interval level the curves of order for specific transformation paths are not meaningful. However, the inconsistency between the two measures is more fundamental and even an ideal monotonic transformation would not be able to resolve it. This is explained in the next simulation.
4.5. Does the Choice of Order Measure Matter?

Figure 4.18: Quantification of order for the two transformations described in simulation case IV. Starting from the same pattern we apply either a non-uniform stretching (red) or a process of random points removal (green). We observe that the rates of decrease predicted by the two measures, nearest-neighbour distances variance (left) and Delaunay triangles entropy (right), for the two transformations are inconsistent. (The sign of the algorithms’ output has been inverted so that high values correspond to high order. Pattern instances are plotted every two steps as in case III.)

Simulation case IV. In this simulation, we apply two different types of transformation to the same initial pattern. We use as initial pattern the final pattern of the previous case, IV. Each type of transformation is applied for ten steps. The one transformation, $T_1$, is a non-uniform stretching along the $x$-axis. The other transformation, $T_2$, is a random removal of points. Figure 4.18 shows the changes in order for each of the two paths for the two measures. We notice that for the nearest-neighbour distances variance the rate of decrease of order is higher for the $T_2$ than the $T_1$ process. The opposite holds for the Delaunay triangles entropy; the rate of decrease is higher for the $T_1$ than the $T_2$. This disagreement cannot be resolved, as any monotonic transformation will preserve this inconsistency.
With the use of simulated point patterns we showed that while two measures can agree for a restricted range of transformations (case I), significant differences can occur when the pattern evolves through more general processes (case II). Thus agreement of two measures does not provide reassurance about their validity. Since, in natural systems we do not have access to the actual mechanisms, but on the contrary, most often these are the aim of research, an acceptable measure is one which would be validated in the widest possible range of relevant patterns. Further, we showed that a mismatch between two measures on the interval level, invalidates comparison of rates (case III), but still this may be impossible to rectify even for transformations where measures exhibit perfect rank correlation (case IV).

The described simulations include stochastic processes. Repetition of the simulations with different random seeds results in curves which are very close to the ones showed. The analogue for the analysis of natural patterns is the property of repeatability. In this aspect, the simulations emphasize that the high repeatability of an algorithm in estimating order is unrelated to its validity.

4.6 Conclusion

We demonstrated the use of the $a$-scale on pattern formation data from the fly. Analysis of the bristle cell patterning allowed an objective quantification of the timecourses of order for different sub-populations of the pattern. The qualitative structure of these timecourses is consistent with existing hypotheses about distinct pattern-refinement processes operating in parallel. Analysis of the epithelium allowed similar quantification for different sub-parts of the tissue in agreement to recent findings of mechanistic processes driving the system to a well-packed arrangement. With the use of simulated point patterns we highlighted fundamental controversies that can occur when applying different measures of order to evolving systems.
Chapter 5

Difference Magnitude of Order in Point Patterns is not Measured by Discrimination Steps

5.1 Introduction

When analysing a perceptual dimension (e.g. lightness, glossiness, or regularity), a fundamental question is whether the intensity of the percept can be represented quantitatively, i.e., as a numerical value on a scale. In Chapter 3 (Protonotarios et al., 2014) we have shown, using pairwise comparisons of point patterns from a diverse set, that observers are highly consistent in their judgements of order, and that their judgements are compatible with an interval scale structure. The derived scale is expressed in units of internal noise (just-noticeable differences or jnds), and its construction relies on the confusion between adjacent stimuli. Although comparisons of all pairs of patterns were used in the construction of this scale, the process relies particularly on judgements of sub-threshold differences while supra-threshold differences do not affect the scaling apart from validating its ordinal structure. Figure 5.1 shows diagrammatically the concept of perceptual comparison according to the Thurstone Case V model with respect to the magnitude of a specific attribute for two stimuli that are placed far apart on the perceptual scale.
A natural question, therefore, is whether large intervals on this scale correspond to direct perceptual differences (Luce and Krumhansl, 1988). There is no fundamental reason why these two types of judgement should depend on the same sensory and/or cognitive mechanism and therefore be predicted by a common perceptual scale. To examine this it is necessary to collect psychophysical data for both types of judgement.

For the collection of data, direct scaling methods pre-suppose a structure in the scale and so indirect scaling methods are preferred. With these the structure of the scale can be validated and the internal noise can be modelled explicitly. Two methods of indirect scaling can be distinguished. The first one—magnitude
comparison—is based on pairwise comparisons between stimuli and originates with Thurstone (Thurstone, 1927). Judgements are collected from a series of presented pairs of stimuli where the observer selects the one that contains the attribute in question at the greatest degree. Analysis of such judgement data yields a discrimination based scale similar to the one we derived in Chapter 3. The second method—magnitude difference comparison—has been considered by Maloney and Yang (Maloney and Yang, 2003; Knoblauch and Maloney, 2008) and requires observers to consider quadruples of stimuli arranged as two pairs and to select the pair that shows the greater perceptual difference. For this task, it is necessary that the perceptual difference between stimuli of the same pair is evident (supra-threshold). This method has been applied to a range of perceptual dimensions: colour (Maloney and Yang, 2003; Brown et al., 2011; Lindsey et al., 2010; Yang et al., 2008), quality of compressed images or video (Charrier et al., 2007, 2011, 2012; Menkovski and Liotta, 2012), similarity between surface texture pairs for gradual phase randomization (Emrith et al., 2010), glossiness (Obein et al., 2004), transparency (Fleming et al., 2011), strength of the watercolour effect (Devinck and Knoblauch, 2012; Devinck et al., 2014), similarity between pairs of faces (Rhodes et al., 2007), correlation in scatterplots (Knoblauch and Maloney, 2008), auditory stimulus duration (Yang et al., 2008) and emotional intensity (Junge and Reisenzein, 2013).

For the construction of perceptual interval scales from indirect scaling data, one supposes the existence of a perceptual continuum where the true values of the attribute in question reside. Each time a judgement concerning a configuration of stimuli is performed, noisy realizations of the true values of the attributes are perceived and then compared, either directly or after a differencing step. Assuming the realization noise is stationary along the scale, and each perceptual realization is independent, a link function with a sigmoid form converts scale value differences, or differences of differences, into response probabilities. The form of noise distribution determines the exact shape of the link function. With a dataset of responses to a set of scaling tasks, which has been well-chosen with respect to stimuli spacing and is sufficiently numerous, the noisy realization model can be critically tested using
likelihood analysis. If the model is found to account adequately for the data, scale values for the experimental stimuli can be estimated by maximum likelihood model fitting.

In this chapter we investigate whether there is a single internal interval scale of order for point patterns that underlies the perception of order for both discrimination and perceptual difference tasks. In the following sections we present two experiments we conducted for the collection of magnitude comparison and difference comparison data for the perception of order in point patterns and the construction of the corresponding scales. We then investigate whether a common perceptual scale can account for the data of both experiments.

5.2 Methods

5.2.1 Stimuli

The space of approximately ordered point patterns is vast. In Chapter 3 (Protonotorios et al., 2014) we used a multistep process to synthesize diverse examples from this space. Starting from perfect lattices of points, which could be either triangular, rectangular or hexagonal, we independently perturbed the position of the points, randomly deleted and added varying proportions of them, and finally applied a smooth nonlinear positional warp. In this study, however, we are not concerned with a thorough investigation of the interaction of these factors in the final percept of order but rather aim to compare the two methods of scaling. Demonstrating a disagreement between the two scaling methods in a subset of the space of point patterns would be enough to show that the two tasks utilize distinct mechanisms. Additionally, we noticed in pilot experiments that in difference comparison, the effect of secondary parameters of the stimuli (e.g. type of lattice) can be significant. In particular, a quadruple of patterns may appear where one pair exhibits a large difference in the degree of order but both patterns are based on the same type of lattice (e.g. triangular), while the other pair exhibits a smaller difference in order but its patterns are based on different types of lattice (e.g. triangular and square).
In this case, it is possible that the participant may be influenced by the similarity of the first pair with respect to the secondary parameter and judge the second pair as the one showing the larger difference.

Given these considerations we chose a simple method for the generation of stimuli with varying degree of order and minimum presence of secondary perceptual dimensions. We started with a point pattern based on a perfect square lattice and introduced independent Gaussian positional jitter on the points. The physical parameter that defined the strength of the jitter was the standard deviation of the Gaussian distribution (on each coordinate) expressed as a fraction of the spacing of the perfect square lattice. The final step was a random selection of a circular window containing exactly 180 dots.

![Figure 5.2: Boxplot of $a$-scale order values for generated point patterns for varying jitter level. The latter is expressed as the standard deviation of the Gaussian perturbation of the dots as a fraction of the dot spacing of the perfect square grid. (Horizontal axis has been inverted so that high order values are on the right hand side.)](image)

For our experiment to be effective it is ideal if we use a set of stimuli that are uniformly spaced in perceptual order. Such a sampling scheme on the discrimina-
5.2. Methods

tion scale maximizes average overlapping of stimuli and provides uniform sampling at least in one of the scales, in particular the discrimination scale. Even with the simple jittered square lattice we employ this is not trivial to achieve for two reasons. The first is that patterns can vary in how ordered they appear even at a fixed level of jitter. The second is that it is unjustified to expect that average perceptual order depends linearly on jitter magnitude. We have dealt with those problems by using the absolute interval scale for the measurement of order in point patterns we have constructed in Chapter 3 (a-scale). On this scale the anchored values 0 and 10 correspond to total randomness (Poisson point patterns) and perfect Bravais lattice respectively, and each unit corresponds roughly to a jnd. The basis of the scale is a geometric algorithm that assesses the variability of the spaces between the points taking into account the distributions of the sizes and shapes of the triangles defined by their Delaunay triangulation (Delaunay, 1934). After examining a list of pre-existing and designed geometrical measures, this algorithm was identified as the one which correlates very well with human perception. The a-scale predicts the perceptual order of patterns on a discrimination-based interval scale.

By generating many patterns at different jitter levels and computing their a-scale values, we have determined 21 jitter levels that on average are uniformly spaced on the a-scale, and have selected patterns that have close to the mean a-scale value for their level of jitter. Figure 5.2 shows the boxplot of the a-scale values of the generated point patterns for varying jitter levels. Figure 5.3 shows the a-scale order values for the selected patterns for the magnitude and difference comparison experiments against the corresponding jitter levels. The patterns at the highest jitter level (disordered end) were generated from a Poisson process which is equivalent to applying an infinite amount of jitter to a square lattice. We additionally excluded patterns that contained points that were so close that they would overlap when displayed as dots; this happened with increasing frequency for larger amounts of jitter.

For the magnitude comparison task the full set of 21 patterns was used (set A). For the difference comparison task, the 11 odd-numbered patterns were used (set B).
5.2. Methods

Figure 5.3: Shows $a$-scale order values for selected point patterns for the magnitude and difference comparison tasks against the corresponding jitter level. The latter is expressed as the standard deviation of the Gaussian perturbation of the dots as a fraction of the dot spacing of the perfect square grid. For the Poisson pattern (most disordered), the equivalent amount of jitter is infinite. (Horizontal axis has been inverted so that high order values are on the right hand side.)

The number of patterns used was chosen so that the resulting number of pairwise comparisons for a balanced task is not experimentally prohibitive while giving a relatively dense set. This is necessary since the scaling method based on the magnitude comparison task relies on overlap of the estimated values of order.

5.2.2 Observers

Twelve observers (5 females) with normal or corrected-to-normal vision participated in both of the experiments. Participants’ age ranged from 18 to 40 years ($Mean = 25.4$ years, $SD = 6.0$ years). Our research adhered to the tenets of the Declaration of Helsinki for the protection of human subjects.
5.2.3 Procedure
The patterns were presented in pairs or quadruples depending on the task, at a comfortable viewing distance of approximately 50 cm on a 40 cm diagonal laptop screen of resolution $1920 \times 1080$. They were presented as solid black dots of 0.5 mm diameter, on white circular disks of radius $r = 4.0$ cm, on a grey background. Participants viewed the patterns under comfortable room illumination. In figures, the size of the dots in the patterns has been increased for visibility in reproduction. Presentation of stimuli and recording of responses were controlled using the MATLAB Psych toolbox (Brainard, 1997). In both tasks participants were given unlimited time to respond for each trial.

5.2.4 Magnitude Comparison Experiment
Observers first completed the magnitude comparison task (discrimination). Each of the $(21 \times 20/2 =) 210$ pairs was presented in random order in two blocks (Figure 5.4), resulting in 420 comparisons in total per participant. In each trial the patterns were randomly allocated to left or right in the first block and then in the opposite way in the second block. Each pattern was randomly oriented at integral multiples of 90 degrees. Randomization aimed to minimise learning of the patterns and to reduce bias and effects of adaptation.

Participants were given written instructions to use the keyboard (left or right arrow keys) to indicate which of the patterns “appeared more ordered” to them. They were able to return to the previous trial if they wished to correct a keystroke error. They were free to control the pace of the experiment, but all took 12-20 minutes.

5.2.5 Magnitude Difference Comparison Experiment
The magnitude difference comparison task followed the magnitude comparison task after a short break. Observers were presented with quadruples of patterns arranged as two horizontal pairs separated by a thick black horizontal line (Figure 5.5). Each
of the quadruples was presented in random order in two blocks. The two pairs were randomly allocated to the upper or lower part of the screen in the first block and the opposite way in the second block (Knoblauch and Maloney, 2012). Each pattern was again randomly oriented at integral multiples of 90 degrees. Randomization, as in the first experiment, aimed to minimise learning of the patterns and to reduce bias and effects of adaptation. With each pair defining an interval on the perceptual continuum, all quadruples with non-overlapping pairs of intervals were presented. This gives 330 distinct quadruples and since all were presented twice the total number of comparisons for the magnitude difference comparison task per participant was 660.

Participants were given written instructions to use the keyboard (up or down arrow keys) to indicate which of the pairs of patterns “showed the largest difference in the degree of order”. They were able to return to the previous trial if they wished to correct a keystroke error. They were free to control the pace of the experiment, but all took 30-50 minutes.

The method of quadruples for the magnitude difference scaling is the original suggested by Maloney and Yang (Maloney and Yang, 2003). A variant using triads has also been used for the construction of difference scales (Devinck and...
5.2. Methods

Figure 5.5: Participants were asked to select the pair of patterns (upper or lower) that shows the greatest perceptual difference in order. For this example, the upper pair gained 3 responses and the lower pair gained 21.

Knoblauch, 2012). In this variant the observer still compares intervals but one of the three stimuli serves at the same time as both the high endpoint of one interval and the low endpoint of the other. In the triads method, the number of possible comparisons is lower but this does not provide any benefit since the standard deviation of the estimates depends on the number of actual trials (Maloney and Yang, 2003). Furthermore, in the triads method direct comparison is conducted only for adjacent intervals and not for intervals which are further apart.
5.3 Results

5.3.1 Agreement Rates

We computed two measures of response variability, the *intra*- and *inter*- observer agreement rates, shown in Table 5.1. The *intra*-agreement rate expresses the probability that a random participant would repeat the same judgement when faced twice with the same random trial. Observers do not always respond the same to a pair or quadruple of patterns. The rate of 91% for magnitude comparison with set A demonstrates that there is high consistency. If we restrict the estimation of the *intra*-agreement rate by subsampling on the set B, we observe a slight increase to 93%. This increase is expected since the fewer patterns of set B are more widely spaced in order. The *inter*-agreement rate is the probability that two observers will agree on the same random trial. For the magnitude comparison task, the 91% for set A, and the 93% for set B, show that there is no variation between participants over-and-above their personal variability. For the difference comparison task, the agreement rates are lower. The *intra* rate is 76% indicating lower consistency of observers in comparison to the discrimination task and the *inter* rate is 70%, indicating an additional 6% of inter-personal difference over and above personal variability.

<table>
<thead>
<tr>
<th></th>
<th>Magnitude comparison</th>
<th>Difference comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Set A (21 patterns)</strong></td>
<td>intra (%)</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>inter (%)</td>
<td>91</td>
</tr>
<tr>
<td><strong>Set B (11 patterns)</strong></td>
<td>intra (%)</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>inter (%)</td>
<td>93</td>
</tr>
</tbody>
</table>
5.3. Results

5.3.2 Magnitude Comparison Scaling

5.3.2.1 Model fitting

For the magnitude comparison task, each stimulus $S_i$ is assumed to have a true value $M_i$ on a scale and each separate perception $\psi_i$ of it is a noisy realization of the true value. When a magnitude comparison between two stimuli $S_i, S_j$ takes place, the two noisy realizations are compared and the observer reports which one is larger (or smaller). Assuming the noise for each perception is identically distributed and independent, then for magnitude comparison, there exists a monotonic preference function which maps the signed difference between the true values, $\Delta M = M_i - M_j$, to the probability that the one will be preferred to the other. This preference function is the cumulative distribution function of the realization noise distribution convolved with itself. When the noise distributions of the set have sufficient overlap then the preference rates will not all be 0 or 1 and fitting a model to the preference rate data allows the interval, not just ordinal, structure of the true values of the set to be estimated. If the noise is assumed normally-distributed (Thurstone Model Case V) the preference function has the form of cumulative Gaussian distribution function (Thurstone, 1927) whereas if for example the noise is assumed Gumbel-distributed (Bradley-Terry Model) then the preference function has a logistic form (Bradley and Terry, 1952; David, 1988). In this chapter, we will only consider Gaussian noise.

As when establishing the perceptual scale of order based on discrimination for the diverse set of patterns in Chapter 3, we fit models by maximization of the likelihood of the data. The likelihood is computed using binomial probabilities for each stimulus pair. It has been shown that Maximum Likelihood (ML) fitting is extremely sensitive to probabilities near 0 or 1 (Harvey, 1986; Wichmann and Hill, 2001a). To incorporate in the model the independent of the attribute differences errors (lapses), we rescale the preference function as $\lambda + (1 - 2\lambda) P(\Delta M)$, where $\lambda$, positive but small, is the lapse rate parameter (Wichmann and Hill, 2001a). The model is parameterized by (i) the unknown true values of each stimulus, (ii) the lapse rate parameter, and (iii) any additional parameters used to vary the form of $P$. We used gradient descent for model fitting, with multiple random starts to check for
5.3. Results

stability. The equal variance Gaussian noise model we used required no additional parameters.

Since the interval scale is invariant under linear transformation we can choose the unit distance on the scale to correspond to a just-noticeable difference (jnd). By convention, the magnitude of a jnd is such that an observer will have a 75% chance of correctly ordering two stimuli whose attributes are different by this amount (Torgerson, 1958).

The fitted values for the set $A$ with respect to the physical parameter (amount of positional jitter of the points) are shown in Figure 5.6. This figure shows the estimated degree of perceived order on the discrimination scale in jnd units; the degree of perceived order increases as the amount of jitter decreases. Since only differences in fitted values, not absolute values, are used to predict preference rates, ML fitting only estimates values up to an additive constant for the entire set. In particular, the value zero has been selected for the Poisson pattern which is judged as the most disordered. The estimated value for the most ordered pattern is 16.5 jnds, meaning that the whole range of order for set $A$ is 16.5 jnd units. The ML estimated lapse rate parameter was 0.3%. We note that 16.5 jnds is larger than the 10 jnds we found between order and disorder in Chapter 3 (Protonotarios et al., 2014). We attribute this difference to the greater variation in pattern types in the previous experiment.

Our ML estimates for the scale values of our stimuli are uncertain because we have collected only a finite number of responses to each stimulus pair. To estimate this uncertainty we used a parametric bootstrap method (Efron, 1979; Wichmann and Hill, 2001b). The method uses the ML-fitted equal variance Gaussian model (with lapsing) to generate sets of synthetic experimental data. A new model is fitted to each synthetic data set. To remove the effect of arbitrary additive constants, we align different bootstrap estimates by subtracting the mean of each set. We thus obtain a set of estimates for each stimulus, the standard deviation of which provides an estimate of uncertainty. Across the stimuli of set $A$ the standard deviations ranged from 0.46 to 1.04 jnds, with 0.67 jnds being the (RMS) average. The size of the range of the fitted values in set $A$ is $16.5 \pm 1.2$ jnds. The fitted values for the pat-
5.3. Results

Figure 5.6: Fitted values of order based on the discrimination task plotted against the amount of jitter. The latter is expressed as the standard deviation of the Gaussian perturbation of the dots as a fraction of the dot spacing of the perfect square grid. For the Poisson pattern (most disordered), the equivalent amount of jitter is infinite. (Horizontal axis has been inverted so that high order values are on the right hand side.)

Patterns and the resulting 95% confidence intervals are shown in Figure 5.7. Patterns are numbered according to their predicted $a$-scale values with “1” corresponding to the least ordered and “21” to the most ordered. We observe that the graph is approximately linear, confirming the effectiveness of our $a$-scale computation of approximate order for jittered square lattice patterns which is a different population than used to establish the $a$-scale.

5.3.2.2 Goodness of fit

Having determined a maximum likelihood model we assess its goodness-of-fit (GoF) by comparing its empirical deviance to the distribution of deviances that result by Monte Carlo generation of random datasets from the ML model. If the empirical deviance lies within the range of simulated deviances that encompass the
Figure 5.7: Fitted values of order based on the discrimination task. Bars show 95% confidence intervals as estimated from bootstrap analysis (200 repetitions). Patterns are numbered according to their predicted $a$-scale values with “1” corresponding to the least ordered and “21” to the most ordered.

95% most common then we accept the model. Deviance of a dataset (original or simulated) is twice the difference between the log-likelihood of the dataset given the ML model, and the log-likelihood of the dataset given a saturated model. The saturated model in this case specifies a separate ML probability for each trial.

For the equal variance Gaussian model, the empirical deviance is 145 and the 95% interval of acceptable deviances for 10,000 repetitions is [107, 170], hence the model is accepted. Therefore, the consensus of discrimination-based order for set $A$ has the structure of an interval scale. We refer to the resulting scale as the discrimination scale.
5.3.3 Magnitude Difference Comparison Scaling

5.3.3.1 Model fitting

The difference comparison scaling method is based on a stochastic model of difference measurement (Krantz et al., 1971; Maloney and Yang, 2003). Similarly to the Thurstone scaling assumptions, each stimulus is associated with a number on an interval scale expressing the real attribute contained in the stimulus. The aim is to recover these numbers for a set of stimuli by collecting direct interval comparison judgements.

We again denote the set of stimuli as $S_1, S_2, \ldots, S_N$, numbered in such a way so that the physical parameter, $\phi_i$, related to each stimulus is ranked as $\phi_1 < \phi_2 < \ldots < \phi_N$ (here $N = 11$). We assume that each stimulus, $S_i$, in the set evokes a perceptual response for the degree of order which can be numerically represented as $\psi_i = M_i + N(0, \sigma)$, with $M_i$ being the true value of the attribute. We assume that when an observer compares the perceptual difference between pairs $(S_i, S_j)$ and $(S_k, S_l)$, they respond on the basis of the sign of $|\psi_j - \psi_i| - |\psi_l - \psi_k|$. The probability of choosing the first pair over the second is given by $Pr(|\psi_j - \psi_i| > |\psi_l - \psi_k|) = Pr(|\frac{|\psi_j - \psi_i|}{|\psi_l - \psi_k|}|^2 > 1)$, where $|\frac{|\psi_j - \psi_i|}{|\psi_l - \psi_k|}|^2$ is a random variable following a doubly non-central F distribution, with one degree of freedom in both numerator and denominator (Morgan et al., 2015). If the pair differences $|M_j - M_i|$ and $|M_l - M_k|$ are always sufficiently larger than the noise level then observers will never make an error about which pattern of a pair is the more ordered, and the response probability arises from a link function depending on $(M_j - M_i) - (M_l - M_k)$. The link function is the cumulative distribution function of a Gaussian of variance four times the variance of the realization noise. We are justified in using this link function for our data because of the spacing of patterns in set $B$ used for the difference task. Within the discrimination data, pairs of patterns from set $B$ are correctly ordered in 96% of trials. In this chapter, we only consider Gaussian noise when examining the magnitude difference comparison scaling; however Maloney and Yang showed, with the use of simulations, that the resulting scale
is robust with respect to the distributional assumptions for the noise (Maloney and Yang, 2003).

Similarly to the discrimination scale fitting we rescaled the link function to incorporate the lapse rate. We used Maximum Likelihood method with gradient decent and multiple random starts for stability checking to estimate the parameters $M_1, M_2, \ldots, M_N$ and $\sigma$.

5.3.3.2 Goodness of fit

For the GoF assessment we compared the empirical deviance with the histogram of deviances based on Monte Carlo simulations using 10,000 repetitions. Pooling data for all observers did not result in an acceptable model for the equal-variance Gaussian model (empirical deviance = 494 with 95% interval of acceptable deviances: [303, 402]). A model of a common scale for all observers was accepted when we allowed different sensitivities ($1/\sigma$) for each of the two sessions completed by each observer (a noticeable variation in the sensitivity per observer has been found also in other studies where the same difference scaling method was applied, e.g. Devinck and Knoblauch (2012)). For this model the empirical deviance was 7051, inside the 95% interval of acceptable deviances [6898, 7213]. Therefore, also for the case of the difference comparison task, a common scale for the stimuli can fit the collected data for all participants, when independent sensitivities are allowed per session.

The fitted attribute values for the common scale with respect to the physical parameter (amount of jitter) are shown in Figure 5.8. In the figure, the fitted values of the difference scale have been normalised so that patterns of extreme order/disorder get the values of 1/0 respectively (standard scale (Knoblauch and Maloney, 2012)).

To exclude the possibility of a change in sensitivity because of learning, we examined the distribution of sensitivities for the first and second session per observer. For comparison, we normalised the sensitivities so that the highest sensitivity of all sessions is set to 1. Learning is expected to cause an increase in sensitivity. When, for each observer, the sensitivity of the second session is plotted against the sensitivity of the first session, it is clear that there is no overall tendency for increase
Figure 5.8: Fitted values of order based on the difference comparison task plotted against the amount of jitter. The latter is expressed as the standard deviation of the Gaussian perturbation of the dots as a fraction of the dot spacing of the perfect square grid. The fitted values have been scaled so that patterns of extreme order/disorder get the values of 1/0 respectively. For the Poisson pattern (most disordered), the equivalent amount of jitter is infinite. (Horizontal axis has been inverted so that high order values are on the right hand side.)

(Figure 5.9). The number of data points positioned above the identity line is 5, while the other 7 are positioned below the line. In the same graph it is also noted that the average sensitivities for the two sessions are highly correlated meaning that the observers exhibit consistently low, medium or high sensitivity.

5.3.4 Common Interval Scale Model for Both Experiments
Having established that both the discrimination and the difference comparison data can be described by interval scale models with Gaussian noise, we attempted models of a common interval scale that would account for both sets of data simultaneously. Examining whether a common signal detection model for both tasks is possible is important, as this would imply a common underlying mechanism for
both supra- and sub-threshold judgements. We did not expect that such a model would be accepted though for the combination of tasks for all observers, since even for the difference comparison task alone such a model was not possible unless the per session sensitivities were adjusted. Indeed, by constraining the noise level to be the same for discrimination and difference data, we achieved a deviance of 800 which is outside the 95% interval of acceptable deviances [444, 566] computed using 10,000 repetitions. For a fixed noise level for the discrimination scale, if we allowed different noise levels for each of the difference comparison task sessions, the empirical deviance was 7361, just inside the 95% acceptable interval of [7067, 7389], for 10,000 repetitions. Thus we cannot reject this model based solely on the total deviance analysis. However, by focusing on the deviance residuals distributions for the two experiments, we notice that for the difference experiment the empirical deviance of 7059 was inside the 95% interval of acceptable deviances.
5.3. Results

[6901, 7216], while for the discrimination experiment the empirical deviance value was 302, far outside the 95% interval of acceptable deviances [141, 203]. This is a strong indication that this model does not describe the data for both tasks accurately.

The different noise level for the two tasks is equivalent to a linear transformation between the two scales. Since the empirical deviance value for the total model does not reject it, and since a non-linear relationship linking the two scales could be preferred, we examined a model with a simple non-linearity; we implemented a quadratic transformation linking the two scales’ values, which required the addition of one parameter. This model achieved a deviance of 6974, which fell inside the 95% interval of acceptable deviances [6780, 7115]. Examining the deviances per experiment, for the difference comparison task the empirical deviance was 6797, inside the 95% interval of acceptable deviances [6640, 6970], and for the discrimination task, we achieved a better result than in the linear case, as the empirical deviance value of 177 fell just inside the 95% interval of acceptable deviances [114, 178].

Although, it is not expected that the simple quadratic transformation captures the exact relationship between the two scales, the non-linear model can be used for the rejection of a common interval scale for both the discrimination and the difference comparison tasks. The linear and the simple non-linear models form a nested pair, the linear one being a special case of the non-linear, and therefore the linear relationship (null hypothesis) between the two scales can be tested with the use of the likelihood ratio test (Wilks, 1938; Kingdom and Prins, 2009). For the application of the likelihood ratio test, we need to assume that the statistic $W$, which is twice the difference of the log-likelihoods of the two models in comparison, follows a $\chi^2_d$ distribution, where $d$ is the difference in the number of parameters between the two models (in our case the two models defer by one degree of freedom). Although this is asymptotically true, it is not easy to estimate the number of observations that are necessary to provide a good such approximation (Wichmann and Hill, 2001a). We can, however, simulate the distribution of the statistic $W$. By accepting the linear model which has been estimated with the ML method, we generated a large number
of simulated data and refitted both the linear and the non-linear models. For 500 repetitions, we computed the $W$ statistic and examined how well the simulated distribution is approximated by the $\chi^2_1$ theoretical density. In Figure 5.10 we can see that this approximation is satisfactory in shape and extent. The log-likelihood for the estimated linear model was -3789.31, and the log-likelihood for the estimated non-linear model was -3595.97. This resulted in a value for $W = 386.68$. Getting a value of $W$, which is much greater than the 95% cut off of the $\chi^2_1$ distribution, 3.84, means that we have to reject the linear model in favour of the non-linear one. Since the non-linear transformation distorts the spacing between the scale values in one of the two scales, we conclude that no common interval scale model with Gaussian noise is able to describe both the discrimination and the difference comparison data.

Figure 5.10: Simulation of the $W$ statistic distribution (500 repetitions). The curve corresponds to the probability density function of the $\chi^2_1$. 
5.4 Discussion

Plotting the difference comparison scale against the discrimination scale provides a visual explanation of the failure of a common interval scale to describe both tasks (Figure 5.11). It is evident that the plotted curve deviates smoothly and systematically from a linear relationship. In general, the slope is higher near the disordered end and lower near the ordered end. This indicates that, when matched for discrimination jnd steps, supra-threshold differences near disorder appear larger than near order.

![Graph showing the difference comparison scale against the discrimination scale.](image)

**Figure 5.11:** Scatter plot of the estimated values of order for the point patterns (set B) derived from the discrimination task and from the difference comparison task.

A simple way to illustrate that the two scales are indeed inconsistent with each other is to focus on the mid-pattern of each one (Figure 5.12). By comparing the mid-pattern of the discrimination scale with the mid-pattern of the difference comparison scale we see that the first one is noticeably more ordered than the second. We perceive the mid-pattern of the discrimination scale as being more towards the
ordered end, and conversely, the pattern which we directly perceive as equidistant from the ends is not placed an equal number of jnds from the ends.

Figure 5.12: Illustrates the mismatch between discrimination- and difference-based scales by focussing on end- and midpoints of the scales. We expect that the reader will consider the difference AC as substantially greater than CB, even though they are almost equal in jnd steps. Also, the difference AD will be considered as similar to the difference DB even though they are quite different in jnd steps.

Allowing varying noise across a common perceptual scale for the two tasks is expected to affect the spacing of the stimuli on the two scales to a different degree. The discrimination scale being sensitive in the local extent of the noise will be affected more; the difference scales are relatively robust to whether noise is constant or varying (Kingdom and Prins, 2009). On such a common scale, equal differences at different positions on the scale will not correspond to equal discrimination performance. However, if the particular function that defines the extent of the noise distribution at each point is known, then the jnd width at a particular location can be computed. The simplest implementation of such a model assumes a constant rate of change in noise intensity. Such a model, together with the linear model, which was examined in the previous section, form a nested pair, and thus we can repeat
the same type of likelihood ratio analysis we followed for the quadratic transformation. The varying noise model has one additional parameter in comparison to the linear model—the rate of change. Fitting with Maximum Likelihood we get a log-likelihood value of -3713.11, which results in a value $W_2 = 152.40$, a number much greater than the 95% cut off of the distribution, which is 3.84. This means that the linear model should be rejected in favour of the varying noise model. Although we can reject a common equal variance model in both types of analysis: (a) quadratic transformation of the difference scale, and (b) common perceptual scale with varying noise, it is not possible to decide between the two options by comparing the values of $W_1$ and $W_2$. A major reason is that the proposed functional forms are not necessarily the best among all possible. The two models offer two different interpretations: According to (a), we can assume that observers use different visual mechanisms or employ different criteria when judging sub- and supra-threshold differences. Both scales are consistent with an interval scale structure but are different from each other. Alternatively, assuming a common mechanism and a common transducer function, the internal noise has to vary across the perceptual dimension. For the constantly increasing noise model we examined, (b), there is approximately a two-fold increase (1.85) in the standard deviation of the noise distribution in the direction towards disorder. Both interpretations agree qualitatively on their consequences about how jnd width varies with respect to the difference scale.

Devinck and Knoblauch (Devinck and Knoblauch, 2012) previously made a comparison of scales based on magnitude and difference comparisons and contrary to our results found them compatible. They collected comparison judgements for the strength of the fill-in percept of the watercolour effect (Pinna et al., 2001) as they varied the luminance ratio between the two components of the generating contour. For difference comparison data, they found it necessary to construct separate scales for each observer. They compared the discrimination and difference comparison scales and claimed that they were compatible in contrast to our findings. However, they tested the discrimination performance of each observer only at a narrow range of the physical parameter. It is our view that to exclude the possibility of only a
local agreement between the two scales a verification over a wide range of the scale is necessary, as the nonlinear relation between the two scales is only apparent when a larger extent is considered.

The method of Maximum Likelihood Difference Scaling (MLDS) as suggested by Maloney and Yang (2003) provides a convenient tool for the construction of perceptual scales based on supra-threshold judgements. This method in combination with the traditional Thurstone type scaling can address the important psychophysical question regarding the existence of a unique perceptual scale for both discrimination and appearance. These two methods do not require the existence of a physical parameter, and therefore are particularly useful when such a parameter does not exist or the stimuli are generated by the use of multiple parameters. The disagreement between the two scales can be relevant in cases where the experimental method involves both sub-threshold and supra-threshold comparisons in the same task. For example, recently Jogan and Stocker (2014) suggested a new two-alternative forced-choice (2AFC) method for the characterization of the perceptual bias caused by stimuli secondary parameters, where a common equal-variance signal detection model is assumed for both short and long stimulus distances. Not verifying the validity of such a model can confound the analysis as depending on the perceptual dimension under study the effect may be significant.
5.5 Conclusion

Analysing judgements of perception of order in point patterns we have shown that it is possible to construct separate interval scales for the description of magnitude comparison and magnitude difference comparison data. A common interval scale with Gaussian noise model describing both is not possible. The assumption of Gaussian noise does not affect the generality of our results since the Thurstone and the Maximum Likelihood Difference Scaling methods generate scales which are robust with respect to the distributional assumptions (Stern, 1992; Maloney and Yang, 2003), while it simplifies the analysis because the random variables corresponding to perceptual comparisons remain Gaussian.

Since the mismatch between the scales is most apparent in comparisons involving the endpoints and midpoints, it seems unlikely that a common scale can be achieved by any more complex model that properly treats the interval structure of the scales, or the use of non-Gaussian noise. Therefore, with regard to the degree of order in point patterns, the magnitude of large differences is not determined by the number of jnd steps.
Chapter 6

Discrimination Performance for Order in Point Patterns

6.1 Introduction

Point patterns are abstract mathematical entities, but concrete choices have to be made about how they are displayed to be accessed by vision. To visualise a point pattern, we have to depict a finite number of points contained in a limited area using visual elements of non-zero size. Elements need to be small to define position accurately and need to carry the least information over and above their location. Suitable choices are dots (small circular disks) or other simple elements with circular symmetry (e.g. Gaussian blobs) since circular symmetry minimises the bias in the estimation of the position (Bingham and Muchisky, 1993a,b; Baud-Bovy and Soechting, 2001). Dot patterns are commonly used in various scientific fields, and we have adopted the same approach. By presentation parameters we refer to (i) the size of the dots (diameter), (ii) the number of the dots, and (iii) the radius of the circular area that the point pattern occupies when displayed. There are, of course, additional parameters that determine the appearance of a dot pattern (e.g. contrast), but in this work we will only consider the afore-mentioned set.

In the psychophysical experiments described in Chapters 3 and 5, we had selected the number of dots (180) to be large enough so that the patterns can be statistically examined, but not so large that quadruples of patterns cannot be simultaneously displayed on an LCD monitor. This allowed, for the experiments of Chapter 5,
180 dots of diameter = 0.5 mm to be displayed within a circle of radius = 40 mm without dot overlapping even in arrangements of high disorder. For these presentation parameters, the analysis of the magnitude comparison data showed that humans can distinguish up to 16.5 jnd levels for point patterns that ranged between perfect order (perfect square lattice of points) and total disorder (Poisson point pattern). The set of stimuli consisted of patterns that differed only in the amount of the positional jitter of their points. The number of jnds between the two extremes of order is a measure of the average discrimination performance of the visual system for pattern pairs of similar order values. Since the choice of the presentation parameters, although reasonable, has been arbitrary, in order to generalise our result, we need to examine the robustness of our estimation for a range of presentation parameters.

In this chapter we describe experiments conducted to determine the effect of presentation parameters on discrimination. We used the same class of patterns (jittered square lattice of points) as in Chapter 5. Our hypothesis is that for a reasonable parameter range, discrimination of order is relatively stable. If this is verified, our aim is to determine a smooth function of the presentation parameters that predicts this performance.

6.2 Methods

6.2.1 Experiments

Ten experiments, in total, were conducted. All were based on the same two-alternative forced choice (2AFC) task. Observers were presented with pairs of point patterns and had to select the one that appeared more ordered to them. Each experimental condition was characterized by a specific choice of the presentation parameters for the displayed dot patterns. As mentioned in the introduction of the chapter, these parameters were: (i) the dot size, (ii) the number of dots, and (iii) the radius of the circular area that contained the pattern. Only patterns with the same presentation parameters were compared with each other.
6.2. Methods

Experiments are divided into two groups (Group 1 and Group 2) based on the choices of the presentation parameter values. In Group 1, which consists of experiments 1 to 7, we examine discrimination performance for varying dot size and number of dots, keeping dot density constant. After analysing the derived perceptual scales of these cases, we ran the experiments of Group 2. In Group 2, which consists of experiments 8 to 10, we varied dot size and pattern area, keeping dot number constant.

6.2.2 Stimuli Generation

The stimuli were generated with a method similar to the one described in Section 5.2.1. Patterns were created using a square lattice as a basis and increasing amounts of Gaussian positional jitter to introduce increasing disorder. We express jitter as the standard deviation of the Gaussian distribution (on each coordinate) as a fraction of the lattice spacing. The final step of this process was a random selection of a circular window containing the required number of points. Although we are interested in comparing the scales across experiments, use of a single set of stimuli for all experiments was excluded. Extensive exposure to them would induce learning and so judgements would not be independent. Conveniently, the chosen method of pattern generation permits selection of multiple similar patterns at each level of order for the different experiments. Additionally, since these patterns exhibit spatial uniformity, control of the number of points can be conducted in an unbiased way by selecting a random subregion.

As in Chapter 5, we used the $a$-scale (Section 3.4) to select patterns that are roughly uniformly spaced on the discrimination scale. We first generated a large number of patterns at different jitter levels. After computing their $a$-scale values, we determined 31 jitter levels that on average were uniformly spaced on the $a$-scale. We then selected patterns that had close to the mean $a$-scale value for their level of jitter. We excluded patterns that contained points that would overlap when displayed as dots. To avoid a per-experiment bias in this process, we applied the same criteria for all experiments. We checked for overlap assuming the maximum
parameter values for both dot size and dot number. Finally, ten selected patterns for each of the 31 jitter levels were randomly allocated to the ten experiments. The 31 patterns of each experiment were numbered from ‘1’ to ‘31’ in increasing a-scale value. The pattern of the highest jitter level, ‘1’, was practically created as a Poisson point pattern.

6.2.3 Experimental Design Matrix

Since, we cannot predict beforehand the discrimination performance for each presentation condition, we chose a large number of patterns to ensure sufficient perceptual overlap. Given the large number of patterns and the multiple experiments we aimed to conduct, in order to minimise the per-observer number of trials, we excluded uninformative judgements, i.e. those between pairs of large difference in order. Figure 6.1 shows the experimental design matrix. Filled disks represent pairs that were included in the experiment and non-filled disks represent excluded pairs.

6.2.4 Stimuli Presentation and Task

Stimuli were presented on a 40 cm diagonal LCD screen under comfortable room illumination. The screen resolution was 1920 × 1080 pixels and the viewing distance was approximately 50 cm. Each pattern was rendered using solid black dots on a white circular disk. Patterns were displayed in pairs on a grey background with their centres at the same height separated by a horizontal distance of 17.3 cm. Written instructions were presented on the display in the beginning of each experiment and participants had the opportunity to perform some test trials before they start the actual experiment. In each trial, observers selected “the most ordered pattern from the pair” with the use of the keyboard (left or right arrow keys). Selection initiated next trial. Observers were able to return to previous trials for correction of keystroke errors and were free to control the pace of the experiment. Each of the 189 pairs was presented in random order in two blocks, resulting in 378 comparisons per participant in each experiment. At each trial the patterns were randomly allocated to left or right in the first block, and then the opposite way in the second block. Each pattern
6.3 Methods for First Group of Experiments

6.3.1 Presentation Conditions

In the first group of experiments (experiments: 1-7), we varied separately dot size and dot number. The density of dots in all experiments of the group was kept constant, so the radius of the pattern area was adjusted in accordance to the number of
6.3. Methods for First Group of Experiments

dots. Table 6.1 shows the presentation parameters for the set of the seven experiments. The actual pattern radius on the display was larger by 1.2 mm in all ten experiments to avoid dots appearing on the border.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>dot size (mm)</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.2</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>dot numbers</td>
<td>180</td>
<td>180</td>
<td>180</td>
<td>180</td>
<td>80</td>
<td>125</td>
<td>245</td>
</tr>
<tr>
<td>pattern radius (mm)</td>
<td>72.0</td>
<td>72.0</td>
<td>72.0</td>
<td>72.0</td>
<td>48.0</td>
<td>60.0</td>
<td>84.0</td>
</tr>
<tr>
<td>subgroup</td>
<td>A</td>
<td>A</td>
<td>A, B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

Based on the parameter set, it is convenient to further organise Group 1 into two subgroups (A and B) of four experiments each; experiment 3 is included in both subgroups. Subgroup A includes experiments 1 to 4. In these, the dot size varies, taking values 0.6, 0.8, 1.0, and 1.2 mm, while the dot number remains constant, with the default value of 180. Subgroup B includes experiments 3, and 5 to 7. In these, the dot size remains constant with the default value for dot diameter, 1.0 mm, while the dot number varies, taking values 80, 125, 180, and 245. This consideration allows separate examination of each parameter (dot size, dot number). As already mentioned, for all seven experiments dot density remained constant.

For each participant, experiments were performed in random order. The four experiments containing patterns of the default number of dots (180) were interleaved with the three experiments containing patterns with 80, 125, and 245 dots. Random sub-sequences of the experiments which contained monotonic change in either dot size or dot number were excluded so as to avoid a systematic effect on discrimination performance due to learning or fatigue.

6.3.2 Observers

Eight observers (three females) with normal or corrected-to-normal vision participated in the experiments. Participants’ age ranged from 19 to 41 years.
(Mean = 25.6 years, SD = 7.1 years). Our research adhered to the tenets of the Declaration of Helsinki for the protection of human subjects.

6.4 Results for First Group of Experiments

6.4.1 Agreement Rates
As in previous experiments (Chapter 3 and 5), we computed two measures of response variability, the intra- and inter-observer agreement rates. The intra-agreement rate expresses the probability that a random participant would repeat the same judgement when faced twice with the same random trial. The inter-agreement rate expresses the probability that two observers will agree on the same random trial. The rates are shown in Table 6.2 for all experiments of Group 1. The experiments are arranged in their Subgroups (A and B) and sorted so that the corresponding presentation parameter increases from left to right. We notice that the agreement rates do not differ by more than 8% across experiments; they all range between 71% and 79%. We also notice that for each experiment, the intra- and inter- rates do not differ by more than 2%. This means that there is little or no variation between participants over-and-above their personal variability. The estimated agreement rates are much lower than the agreement rates found in the discrimination experiment of Chapter 5 where the same class of patterns was used. This can be explained by the fact that in these experiments many ‘easy’-to-judge pattern pairs have been excluded and only those which differ in order by a small number of jnds are included. Therefore, this drop in agreement rates is expected. In general, agreement rates show no clear trend with respect to the varying parameter within each subgroup.

6.4.2 Discrimination Scales
We derive the discrimination scales from collected judgements applying exactly the same method we used in the discrimination experiment of Chapter 5 (Section 5.3.2.1). In short, we assume a Gaussian noise model and we fit with Maximum Likelihood. We again modify the preference function to incorporate lapsing
### Table 6.2: Agreement rates for Group 1.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Experiment</th>
<th>A</th>
<th></th>
<th></th>
<th></th>
<th>B</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>dot size (mm)</td>
<td></td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.2</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>dot numbers</td>
<td></td>
<td>180</td>
<td>180</td>
<td>180</td>
<td>180</td>
<td>80</td>
<td>125</td>
<td>180</td>
</tr>
<tr>
<td>intra (%)</td>
<td></td>
<td>71</td>
<td>78</td>
<td>73</td>
<td>76</td>
<td>74</td>
<td>76</td>
<td>73</td>
</tr>
<tr>
<td>inter (%)</td>
<td></td>
<td>71</td>
<td>76</td>
<td>73</td>
<td>76</td>
<td>74</td>
<td>75</td>
<td>73</td>
</tr>
</tbody>
</table>

in the model. In the experiments presented in this chapter, we are not particularly interested in the specific fitted values of patterns at intermediate levels of order, but we mainly focus on the difference of the fitted values of the two extremes. This difference in jnd units is our estimate of the discrimination performance. To simplify the calculations and the visual comparison of scales across experiments, we set the free constant parameter in each fitting so that pattern ‘1’ in all experiments gets the value zero. Then, the discrimination performance value is simply the highest fitted order value. Figures 6.2 and 6.3 show the scales for Subgroups A and B. Figures are presented on the same page to facilitate comparison. Focussing on Subgroup A, we report discrimination performance values ranging between 12.9 and 17.7 jnds. The lowest value corresponds to the smallest dot size (0.6 mm), while the highest value corresponds to the largest (1.2 mm). Focussing on Subgroup B, we report similar discrimination performance values for all experiments; all range between 16.7 and 17.8 jnds.

### 6.5 Discussion on the First Group of Experiments

Neglecting range variation, derived scales look similar for all experimental conditions, and they exhibit almost linear increase with respect to pattern number (Figures 6.2 and 6.3). This validates the efficiency of our \(a\)-scale, whose construction was not based on this class of patterns. For the examined parameter range, the discrimination performance is relatively stable. This is interesting given the two-fold change in dot size and the three-fold change in dot number. Furthermore, it does
6.5. Discussion on the First Group of Experiments

Figure 6.2: Discrimination scales for experiments in Subgroup A. Dot number for all experiments: 180.

Figure 6.3: Discrimination scales for experiments in Subgroup B. Dot size for all experiments: 1.0 mm.
not seem to be affected by dot number, while seems to depend systematically on dot size. Using as reference the experiment with the largest dots (1.2 mm), the performance of experiment 1 (dot size of 0.6 mm) is by almost 5 jnd units lower (27% decrease). It is thus tempting to assume that discrimination performance depends solely on dot size. This assumption, however, contradicts the result of Chapter 5, where, with an even smaller dot size (0.5 mm), discrimination performance was much higher $16.5 \pm 1.2$ jnds. In that experiment the radius of the pattern area was 40 mm, while in experiment 1 the radius is much larger, 72 mm. Given that both experiments displayed the same number of dots (180), their comparison suggests that both average dot spacing ($= \sqrt{\text{pattern area} / \text{number of points}}$) and dot size influence discrimination performance. Since their effect seems opposite, we examine their ratio, ‘dot diameter / average dot spacing’, which we refer to as the aspect ratio of a displayed dot pattern. This is obviously the same for all displayed patterns in an experiment. Examining its values for the eight experiments (seven of Group 1 + the one from Chapter 5), we confirm that the lowest value of $6.3 \times 10^{-2}$ corresponds to the lowest discrimination performance of experiment 1. For all other experiments, values are higher ranging between $8.4 \times 10^{-2}$ and $12.6 \times 10^{-2}$. In Figure 6.4 we plot the discrimination performance against the aspect ratio. Linear fit shows positive correlation; however, this mainly relies on the single point on the left hand side of the graph. In order to establish with higher confidence the positive correlation between these two quantities, more datapoints at the lower range of aspect ratio are required. For this reason, we conducted three additional experiments (Group 2). Our aim is to collect sufficient number of datapoints ($11 = 7 + 1 + 3$) which will cover more evenly the parameter range. The second group of experiments and its analysis are presented in the following section.
6.6 Methods for Second Group of Experiments

6.6.1 Presentation Conditions

For the second group of experiments, we adjusted the presentation parameters to achieve pre-defined aspect-ratio values: $6.0 \times 10^{-2}$, $8.0 \times 10^{-2}$, and $10.0 \times 10^{-2}$.

As in the subgroups of previous experiments, we keep one of the three presentation parameters constant. Since we are interested in examining the discrimination performance with respect to dot size and aspect ratio, we decided to manipulate dot size and pattern radius, and so keep dot number unchanged. For greater diversity in parameter values, we chose different dot number than the previous default of 180. For constant dot number and dot size, the aspect ratio can be lowered by increasing the area of the pattern. Adjustment of the radius to achieve the desired aspect ratio for the extreme numbers of 80 and 245 dots, would result in too large and too small

Figure 6.4: Discrimination performance for experiments of Group 1 vs. aspect ratio (blue), and linear fit (green). The discrimination performance of the previous experiment of Chapter 5 is also shown.
patterns. We, therefore, chose 125 as the dot number value. Dot size of 1.2 mm was similarly excluded due to display area limitations. Thus, the dot sizes for the three new experiments were: 0.6, 0.8, and 1.0 mm. To confidently exclude the possibility that dot size controls discrimination performance, we combined the two low aspect ratio values ($6.0 \times 10^{-2}$, $8.0 \times 10^{-2}$) with large dot size, which was previously associated with high discrimination performance, and the highest aspect ratio value ($10.0 \times 10^{-2}$) with the smallest dot size, which was previously associated with the lowest discrimination performance. Table 6.3 shows the presentation parameters for Group 2 along with the corresponding aspect ratio values in increasing order. Experiment 6, from Group 1, containing the same number of dots (125), is included in the Table and in the scales’ plot for comparison. Again, for each participant, experiments were run in random order. Random experimental sequences of monotonic change in aspect ratio were excluded to avoid a systematic effect of learning or fatigue on discrimination performance.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>dot size (mm)</td>
<td>0.8</td>
<td>1.0</td>
<td>0.6</td>
<td>1.0</td>
</tr>
<tr>
<td>dot numbers</td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
</tr>
<tr>
<td>pattern radius (mm)</td>
<td>84.0</td>
<td>78.8</td>
<td>37.6</td>
<td>60.0</td>
</tr>
<tr>
<td>aspect ratio ($10^{-2}$)</td>
<td>6.0</td>
<td>8.0</td>
<td>10.0</td>
<td>10.5</td>
</tr>
</tbody>
</table>

### 6.6.2 Observers

Eight observers (3 females) participated in the new set of experiments. Six out of the eight had taken part in the previous set of experiments (Group 1). It was desirable to include as many as possible from the previous set so as to minimise observer variation. Participants had normal or corrected-to-normal vision and their age ranged from 20 to 41 years ($Mean = 26.7$ years, $SD = 6.5$ years). Our research adhered to the tenets of the Declaration of Helsinki for the protection of human subjects.
6.7 Results for Second Group of Experiments

6.7.1 Agreement Rates

The agreement rates for the four experiments are shown in Table 6.4. They are on average lower than those of Group 1, whose experiments were of higher average aspect ratio. In general the agreement rates correlate positively with aspect ratio. Comparing the intra- and inter-agreement rates for each experiment, we verify that there is no variation between participants over-and-above their personal variability.

Table 6.4: Agreement rates for Group 2.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>dot size (mm)</td>
<td>0.8</td>
<td>1.0</td>
<td>0.6</td>
<td>1.0</td>
</tr>
<tr>
<td>pattern radius (mm)</td>
<td>84.0</td>
<td>78.8</td>
<td>37.6</td>
<td>60.0</td>
</tr>
<tr>
<td>aspect ratio ($10^{-2}$)</td>
<td>6.0</td>
<td>8.0</td>
<td>10.0</td>
<td>10.5</td>
</tr>
<tr>
<td>intra (%)</td>
<td>71</td>
<td>72</td>
<td>74</td>
<td>76</td>
</tr>
<tr>
<td>inter (%)</td>
<td>71</td>
<td>72</td>
<td>74</td>
<td>76</td>
</tr>
</tbody>
</table>

6.7.2 Discrimination Scales

We repeated the same process as in Section 6.4.2 to derive the discrimination scales. Figure 6.5 shows the corresponding scales for all experiments with 125 dots (experiments 8, 9, 10, and 6). In comparison to previous experiments, the discrimination performance associated with dot sizes 0.8 and 1.0 mm: 16.7 and 16.7 jnds, is now lower: 13.8 and 14.1 jnds, respectively. Conversely, the discrimination performance which corresponded to the dot size of 0.6 mm (12.9 jnds) is now increased (16.4 jnds). The signs of these changes are consistent with the aspect ratio changes. Analysis of the discrimination performance for the whole set of experiments is presented in the next section.
Figure 6.5: Discrimination scales for experiments in Group 2. Top-to-bottom legend displays the dot size parameter for experiments 8, 9, 10, and 6 respectively.

6.8 Discussion

Having completed sufficient scaling experiments to cover a range of aspect ratios, we can now examine its relationship with discrimination performance. To estimate the uncertainty of the discrimination performance values, we used a parametric bootstrap method (Efron, 1979; Wichmann and Hill, 2001b). The method uses the ML-fitted Gaussian model (with lapsing) to generate sets of synthetic experimental data. A new model is fitted to each synthetic data set and the fitted order values are aligned at their minimum value (zero). We thus obtain a set of estimates for each stimulus, the standard deviation of which provides an estimate of uncertainty of the order values. The uncertainty of the fitted value of the most ordered pattern is the uncertainty of the discrimination performance.

Figure 6.6 shows the discrimination performance values against aspect ratio for all experiments (Group 1, Group 2, and discrimination experiment of Chap-
A weighted linear fit is shown along with the datapoints. The $p$-value of the $\chi^2$-test for the linear fitting is 0.72, and the R-squared value is 0.77, meaning that the linear relationship describes a substantial fraction of the data variation. The slope is approximately the same as in Figure 6.4, where datapoints of Group 2 were not considered. The fitted slope expresses an increase of 0.75 jnds in discrimination performance per 0.01 increase in aspect ratio.

Figure 6.6: Discrimination performance for all experiments vs. aspect ratio (blue) and linear fit (green). Bars indicate standard deviations based on parametric bootstrap.

Aspect ratio is controlled by both dot size and average dot spacing. Increasing dot size while keeping dot spacing constant, according to our findings, results in higher discrimination performance. This suggests that while each dot of small size defines its absolute position with higher precision (Bingham and Muchisky, 1993b), larger dots facilitate order estimates which rely on relative placement. It is not clear which mechanisms control the dependency of discrimination performance on aspect ratio. At higher aspect ratio values, both dot edge length and dot area increase in relation to dot spacing. This affects both hyperacuity mechanisms and
luminance. Testing this further would require additional experiments and more precise control of geometry, luminance and contrast. We can still, however, attempt, to examine whether aspect ratio affects discrimination sensitivity more at the ordered end of the scale, or at the disordered end. Order judgements between highly ordered patterns, similar to those appearing in our experiments which are based on a reference pattern (perfect square lattice) and involve point-to-point comparisons, could employ different visual mechanisms than judgements between disordered patterns. Therefore, a non-uniform effect of aspect ratio on discrimination performance may be present. Though it is worth noticing that our algorithm for quantifying order in agreement with perception (a-scale) does have a single mode of operation across the scale.

To test for a non-uniform effect of aspect ratio on discrimination performance, we compared the high performance scales with those of low performance. Since scale values appear noisy, we combined in one dataset the three scales of lowest performance (experiments: 1, 8, 9), and in a second dataset the three scales of highest performance (experiments: 4, 5, 6). For the comparison, we applied a polynomial fit of 6\textsuperscript{th} degree so as to obtain smooth curves. Figure 6.7 shows datapoints and polynomial fits. Appropriate rescaling of the low performance curve results in almost perfect coincidence with the high performance one. This implies a constant increase in discrimination sensitivity across the whole order range.
We remind the reader that observers were free to control the pace of their experiments. To confirm that differences of discrimination performances across experiments are not the result of variation in duration, we present in Figure 6.8 the performance against the average duration of experiment. Observing the datapoints, it is clear that higher discrimination performances are not associated with longer experiment durations. On the contrary, their correlation is negative, with Pearson correlation coefficient $\rho = -0.26$ (with 95% CI estimated by bootstrap $[-0.73, 0.46]$).
6.9 Conclusion

We examined the effect of presentation parameters (dot size, dot numbers, and pattern radius) on average discrimination performance between pairs of patterns of similar order values. We concluded that this is relatively robust for a reasonable parameter range, and we identified the aspect ratio, which we defined as the ratio ‘dot diameter / average dot spacing’, as the primary parameter that controls it. Our finding has a practical application suggesting high aspect ratio for easier order discrimination between displayed dot patterns.

Figure 6.8: Discrimination performance vs. average experiment duration. Observing the datapoints (blue), it is clear that higher discrimination performances are not associated with longer experiment durations. On the contrary, the slope of a linear fit (red) is negative.
Chapter 7

Conclusions

7.1 Contributions

As presented in the introductory chapter of the thesis, in this work we have made five main contributions:

1. We have validated a high inter-observer agreement in judgements of order for a diverse set of point patterns.

2. We have validated that the consensus of judgements of order is consistent with an interval scale structure.

3. We have constructed an absolute interval scale for the quantification of order, based on a geometric algorithm, that predicts order with an accuracy of 0.5 jnds (just-noticeable differences). We have demonstrated its applicability by analysing the progress of organisation of the developing fly notum epithelium.

4. We have rejected a common perceptual scale for both sub-threshold and supra-threshold judgements of order.

5. We have identified, for a common class of point patterns (square lattice with positional jitter), the ratio of dot size to average inter-dot distance (aspect ratio) as the main parameter that affects discrimination performance.
7.2 Critical Appraisal of Contribution

7.2.1 Perception of Order in Point Patterns

For the first two contributions: (i) high inter-observer agreement rate, and (ii) interval scale structure of the consensus of order judgements, to our knowledge, this has been the first time that such consistency is established for a diverse set of point patterns. It is exactly because of this consistency that perceptually-based quantification of order is possible. This finding implies common visual and cognitive mechanisms across humans for order estimates, even when multiple transformations are applied in combination on an initially perfect lattice of points.

We do not expect this consistency to hold for arbitrary sets of point patterns. An indication for this is provided by a comparison between the experiments of Chapter 3 and 5 which were based on different stimuli sets. The discrimination performance of 10 jnds estimated in the first experiment is significantly different than the 16.5 jnds estimated in the magnitude comparison experiment of Chapter 5. In Section 5.3.2.1 we attributed this difference to the greater diversity of patterns in the first experiment, and this assumption is confirmed by the results of Chapter 6, according to which presentation choices for the displayed dot patterns alone cannot explain such a large discrepancy.

It is very probable that when patterns of higher complexity are compared, judgements may not be consistent with an interval scale structure, and may further depend on individual differences and/or the cultural background of the observers. It is not hard to imagine arrangements of points forming structures of increasing complexity that cannot be described simply by the distribution of spaces around points. Such arrangements can form, for example, linear structures like spirals or arrays of clusters of points. At the extreme, patterns of high numbers of points roughly equally spaced can perceptually approach what we characterise as textures or images (Figure 7.1). Order then becomes impossible to qualitatively describe, not to say properly quantify. Since point patterns can express such a vast variety of
arrangements, a strict characterization of the space where our results can be generalised to, is very hard.

Figure 7.1: Point patterns forming textures or images. Left: two textures, the inner (rectangular area) and the outer (background) are easy to segregate based solely on the degree of order in the positional arrangement of the dots; density in both is the same (image inspired from Ouhnana et al. (2013)). Right: Dots of equal size form an easily discerned object, a ball, on a white background (image taken from Secord (2002)). While the textures on the left are generated by points that have same density but vary in degree of order, the image on the right is generated by points that differ in density but maintain constant local order across the image (stippling).

An approach to cope with this difficulty would be to examine pattern similarity for a wide range of point pattern generation algorithms as well as for patterns directly derived from natural systems. Such approach could attempt to partition the point pattern space into equivalence classes independently of the measured attribute, and thus allow an objective and more formal way of defining the limits of applicability of our results. Ideas for similarity measures exist but these are not validated (Mateu et al., 2015; Hagedoorn and Veltkamp, 1999). Perceptually-validated pattern similarity would be a project of similar nature to ours. To avoid the same problem of precise characterization of applicability limits for this project, pattern sets should be defined only on the basis of their generation method. Analysing perception of order, it would be interesting to examine where the limits of inter-observer agreement lie and whether judgements between more complex patterns would still be consistent with an interval scale structure. An interesting question would be whether judgements between stimuli would still be able to be described
by fitted order values on an interval scale but the noise model would have to be adjusted so that its intensity would depend on the similarity of compared patterns.

7.2.2 Absolute Scale of Apparent Order
The constructed absolute interval scale for the quantification of order (\(a\)-scale) predicts order with an accuracy of 0.5 jnds. The algorithm is 70\% more accurate than the best available measures and agrees to the consensus as accurately as a committee of seven independent raters, while being fully objective and repeatable. Since it is intended as a measurement tool for natural scientists for the analysis of static and evolving systems, we demonstrated its applicability; we used it to analyse the progress of organisation of the developing fly notum epithelium.

Similarly to the first two contributions, it is hard to define in a strict way limitations for the applicability of the \(a\)-scale. Using set \(B\)—the second set with the more diverse point patterns in Chapter 3—we identified cases where our preferred geometric algorithm failed to rank pattern pairs in agreement with humans. These involved widely-spaced linear structures with small percentages of randomly added points or severe point deletions that did not however disrupt the percept of linearity. The positional organisation of the points in these patterns was of higher order, involving relationships between large numbers of points and non-local interactions. The scope of our scale, therefore, is limited to point patterns that do not present such multiple levels of organisation. Despite mentioned failures, our algorithm is the only one thoroughly validated for order with a diverse set of point patterns.

Although the \(a\)-scale was constructed for measurement purposes, it can itself help in the study of perception. By design, it reserves a privileged position for the totally random case: the zero of the scale. Interesting paths for further research on perception of order and randomness are suggested by this decision. First, is this point a real zero on the scale? So, is it possible to represent order/disorder on a ratio scale? Second, do total disorder and total randomness coincide perceptually? What, if anything, resides on the negative “mysterious” side of the \(a\)-scale? According to our geometric algorithm, patterns on this side are characterized by even more
spread distributions of Delaunay triangles than the totally random case. That would mean, for example, more small triangles and more large triangles than random; arrangements which are consistent with clustered patterns. What classes of clustered patterns can be considered at the same time more random and more disordered than Poisson? Section 3.5 suggests an unbiased way for generating such patterns. The distribution of $a$-scale values for simulated Poisson patterns, which was found symmetric around zero, made use of patterns of 180 points. Fewer points would result in wider spread of the distribution, and equivalently in higher probability of creating by chance patterns further on the negative side. Experiments containing patterns at both sides of zero generated this way could investigate whether these are perceived as more disordered when compared to those corresponding to the most common $a$-scale values, and whether the latter are recognised as the most random.

Expressing the order of systems in $a$-units improves communication of results and facilitates comparisons, especially for intermediate levels of order. As described in the introductory chapter, while many developmental processes seem to aspire to exact order, this is not universal. In the mammalian eye, for example, parafoveal receptors’ spacings are less than perfectly regular, while foveal cones are regularly packed so as to achieve high density and therefore high resolution. The optics of the eye act as a low pass filter attenuating spatial frequencies greater than 60 cycles per degree and so Moiré-like aliasing artefacts that could occur because of the regular spacing of the foveal cones are prevented. This protection is not sufficient for the less dense parafoveal receptors which would present interference effects at lower spatial frequencies, due to their lower density. However, the arrangement of the parafoveal receptors at a less than perfect level or order, prevents this interference and at the same time keeps density sufficiently high (Wässle and Boycott, 1991). This is an example where mechanisms balancing opposing requirements in a natural system, result in intermediate levels of order. Expressing order in $a$-units helps in identifying optimal system-specific ranges, instead of characterizing these as deviation, to a varying degree, from perfect reference arrangements (e.g. hexagonal packing).
Further, use of the $a$-scale can highlight self-organising principles and common mechanisms in apparently different systems. As we saw in the analysis of the bristle patterning in *Drosophila*, Delta-expressing cells use active protrusions (*filopodia*) to discourage nearby cells from becoming bristle cells. If similar mechanisms control inter-element spacing in different systems, we would expect similar levels of order in their final states. Moreover, analysis of the parameters that control the efficiency of these mechanisms on one system, would allow predictions for the evolution of order on another system. For instance, analysis of the effect of filopodia contacts’ stochasticity (*structural noise*) on the evolution of order on the fly dorsal thorax (Cohen et al., 2011) would allow quantitative generalisation of the results to other systems that rely on analogous mechanisms.

### 7.2.3 Discrimination and Difference Scales of Order

Our fourth main contribution is the rejection of a common interval scale for both sub- and supra-threshold judgements of order. The assumption of a common equal-variance signal detection model for both discrimination and difference comparison data is violated, while separately both types of judgements are consistent with an interval scale structure. Perception of order in point patterns offers a neat example for this type of rejection as by focussing on the mid-patterns of the two scales their disagreement becomes evident (Figure 5.12). The two scales do not coincide; as Figure 5.11 shows, there is a smooth monotonic relationship between the two.

Our method which combines the Maximum Likelihood Difference Scaling as suggested by Maloney and Yang with the traditional Thurstone type scaling, provides a relatively easy way for rejecting the hypothesis of a common equal-variance Gaussian signal detection model. Provided that stimuli extend to a sufficient range of the attribute, this method can detect systematic deviations from linearity between the two scales. Moreover, it does not require the existence of a physical parameter, and so can be used when such a parameter does not exist or the stimuli are generated by the use of multiple parameters.
7.2. Critical Appraisal of Contribution

Unfortunately, depending on the stimuli and the attribute, the task of difference comparisons, can be hard for the observers. These judgements are also prone to secondary attributes affecting the perceptual difference. In our study these limitations did not allow use of diverse point patterns.

7.2.4 Discrimination Performance of Order

By examining the average discrimination performance between pairs of patterns of similar order values for a common class of point patterns (square lattice with positional jitter), we concluded that this is relatively robust for a range of presentation parameters (dot size, dot numbers, and pattern radius). We identified the ratio of dot size to average inter-dot distance (aspect ratio) as the parameter that most affects it. A practical application of this finding is to prefer displaying dot patterns using high aspect ratio (i.e. big dots) so that comparisons between point patterns of similar order are facilitated.

Although there exist visual phenomena that are not scale invariant (e.g. Oppel-Kundt illusion (Wackermann and Kastner, 2010)), in our analysis we have not examined how discrimination performance changes when stimuli are uniformly shrunk or expanded.

When Morgan et al. (2012) analysed discrimination thresholds for near-regular patterns, they used Gaussian blobs for the depiction of the point patterns. They found that sensory noise was equivalent to a Gaussian random variable with a standard deviation of approximately 5% of the inter-element spacing. Size increment of the Gaussian blob would affect the accuracy in the estimation of its centre but also the edge of the blob, which would be less localised. In the case of circular disks, the edge is always precisely defined. It would be interesting to examine the effect of element size on their results and compare it with ours.
Bibliography


