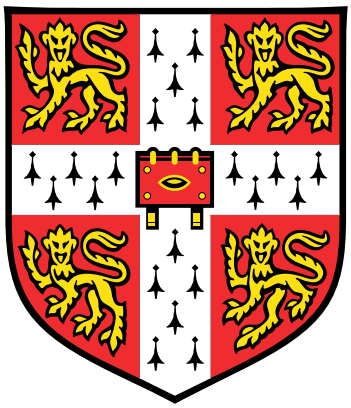
Symmetry and Other Structures

Topics in Nonparametric Regression



Louis Goldwater Christie

Supervisor: Prof. Sir John A. D. Aston

Statistical Laboratory
Department of Pure Mathematics and Mathematical Statistics

University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

St Catharine's College

September 2024

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Louis Goldwater Christie
September 2024

Abstract

In regression we measure covariate samples X_i and response variables Y_i and seek to estimate the conditional mean, or regression function, $f(x) = \mathbb{E}(Y_i \mid X_i = x)$ within some hypothesis class $\mathcal{F}$ of possible regression functions. Estimation of a nonparametric regression function $f : [0,1]^d \to \mathbb{R}$ becomes increasingly difficult as the dimension of the covariate space increases. Specifically, the number of samples n required to ensure that $\sup_{f \in \mathcal{F}} \mathbb{E}(\|\hat{f} - f\|_2^2) \leq \eta$ for a fixed $\eta > 0$ grows exponentially in d for all estimators $\hat{f}$ in many usual problem contexts. To overcome this, we often seek to find structure within the function f we can exploit. The two main structures that have been previously studied are *Sparsity*, where f depends only on $s < d$ of the covariates' coordinates, and *Additivity*, where f is a sum of functions on each of the covariates' coordinates independently. In this thesis we present a new framework for lower dimensional structures that can capture much of the same benefits but now in non-linear situations.

This framework exploits the structure of the level sets of the regression function. If we know the level set $[x]_f = \{u \in [0,1]^d : f(x) = f(u)\}$ then we can average any sufficiently local estimator f_n over this level set to obtain an estimator with a pointwise risk that is bounded in terms of $d - \dim[x]_f$ rather than d. In general it is hard to learn $[x]_f$; however, even a simple plug in estimator (i.e., that uses the level sets of some estimate p_n of f) can be shown to achieve these faster rates (up to log factors) under some strong conditions on f and demonstrate remarkable performance improvements in some finite sample experiments. Instead of placing conditions on f, this thesis gives sufficient conditions on structured hypothesis classes for the level sets $[x]_f$ under which we can achieve the same fast rates of convergence.

This key idea is to generalise sparsity, multi-index models, and other non-linear lower dimensional patterns in f in terms of transformations of the covariate space:

$$\Psi_f \subseteq \{\text{Bijections } \psi : \mathcal{X} \to \mathcal{X}\} \tag{1}$$

for which $f \circ \psi = f$ for all $\psi \in \Psi_f$. Functions that satisfy this property have *Symmetry*: in the same way that we can translate a sparse function along the unimportant axes we can transform f using Ψ_f without effect. Any such f must have level sets $[x]_f$ that contain the

sets $\{\psi(x) : \psi \in \Psi_f\}$, which we can seek to exploit.

We show that an oracle that knows any symmetry group Ψ_f of f performs with worst case L^2 risk decay that depends on a smaller dimension $s \leq d$, rather than the dimension d of the original covariate space. The regression function f can have many possible symmetry groups, but when f is continuous there is always a unique maximal symmetry group Ψ_f within the classes considered. We then show that we can use an M-estimator of this maximal Ψ_f in order to achieve the dimension reduction adaptively under some technical conditions on Ψ_f and the hypothesis class of symmetries considered.

We then show that we can give an inferential framework for this maximal Ψ_f, giving both hypothesis tests for:

$$H_0(\Psi) : f \circ \psi = f \text{ for all } \psi \in \Psi \tag{2}$$

for any transformation group Ψ, as well as confidence regions for the maximal Ψ_f in the hypothesis class of possible Ψ we consider. This also leads to another estimator of Ψ_f that has the computational advantage of not estimating f directly. We demonstrate the performance of these tools both on synthetic datasets and on real data from both the Earth's magnetic field and temporospatial locations of sunspots.

Acknowledgements

This work would not have been possible without the exceptional support from my family, friends, and colleagues. Thank you to my parents for encouraging me to pursue my mathematical inclinations from early childhood and building up my ambitions, and to my siblings for trying to keep me humble through it. Thank you to Madeleine Strait, for being a loving partner through this journey and reminding me of the joys of life beyond mathematics. Thanks to Sam Stephenson, Charlotte Guffick, Luke Cullen, and Veronika Golovina, and many others around Cambridge for creating a community that I'm glad to be a part of. Thank you to the David and Claudia Harding Foundation for awarding me the Harding Distinguished Postgraduate Scholarship allowing me to complete this degree and for connecting me to a collection of incredible students from across the university. Lastly, thank you to my supervisor John Aston, who has had incredible patience to work through all of my bad ideas in order to find the good ones.

Table of contents

List of figures **xiii**

List of tables **xvii**

List of Common Symbols **xix**

1 Introduction **1**

2 Background **7**

2.1 Nonparametric Regression 7

2.1.1 Local Polynomial Estimators 10

2.1.2 Projection Estimators 12

2.1.3 Cubic Smoothing Splines 14

2.1.4 Sparsity 14

2.1.5 Additive Models 15

2.1.6 Other Dimension Reduction Techniques 16

2.2 Manifolds and Differential Geometry 16

2.3 Symmetries and Group Theory 18

2.3.1 Important Examples of Groups 22

2.4 Invariant Functions 24

2.4.1 Compact Symmetries 26

2.4.2 Non-Compact Symmetries 26

2.5 Examples of Symmetry in Statistics 27

2.5.1 Symmetry in Distributions 27

2.5.2 Symmetry in Hypothesis Tests 29

2.5.3 Symmetry in Spaces 30

2.5.4 Symmetries in Density Estimation 33

2.5.5 Symmetries of Regression Functions 34

2.6 How to Use Symmetry in Regression 36

2.7 Symmetry Discovery 38

2.7.1 Auto-Augment 38

2.7.2 Augerino . . . 39
2.7.3 SCALE . . . 39
2.8 Estimation of Level Sets . . . 40

3 Regression Using Level Sets 43
3.1 Introduction . . . 43
3.1.1 Main Contributions . . . 45
3.2 Background . . . 45
3.2.1 Geometry of Level Sets . . . 46
3.3 Manifold AVeraged Estimator (MAVE) . . . 46
3.3.1 Constructing Spaced Points . . . 47
3.3.2 Theoretical Properties . . . 48
3.3.3 Robustness to the Choice of Manifold . . . 51
3.3.4 Increasing the Number of Averaging Points . . . 52
3.4 Estimation of the Level Sets . . . 54
3.4.1 Local Polynomial Level Set Estimation . . . 54
3.5 Numerical Experiments . . . 56
3.6 Discussion . . . 57

4 M-Estimation of Maximal Symmetries 61
4.1 Introduction . . . 61
4.1.1 Main Contributions . . . 64
4.2 Estimation using a Particular Symmetry . . . 66
4.2.1 Defining Partially Symmetrised Estimators . . . 67
4.2.2 Pointwise Convergence Rates of Partially Symmetrised Estimators . . . 68
4.2.3 Integrated Risk of Partially Symmetrised Estimators . . . 70
4.2.4 Computing Partial Symmetrisation Operators . . . 71
4.3 Adaption to Maximal Symmetries . . . 72
4.3.1 Metrising Subgroup Space . . . 73
4.3.2 Defining the Error Minimising Symmetries . . . 75
4.3.3 Global Rates for the Best Symmetric Estimator . . . 77
4.3.4 Local Rates for the Best Locally Symmetric Estimator . . . 78
4.4 Finite Sample Performance . . . 79
4.4.1 3D Rotations . . . 79
4.4.2 Translational Symmetry . . . 80
4.5 Discussion . . . 81
4.5.1 Low Smoothness Case . . . 82
4.5.2 Local Symmetries . . . 82
4.5.3 Improvements on the Constants . . . 83

4.5.4 Lower Bounds on the Estimation of Maximal Symmetries 83
4.5.5 A remark on equivariance 84
4.6 Proofs 84
4.6.1 Proofs in Section 4.2 84
4.6.2 Proof in Section 4.3 90

5 Estimation of Maximal Symmetries via Subgroup Lattices 93
5.1 Introduction 93
5.1.1 Main Contributions 95
5.2 Inference for Subgroups 96
5.2.1 Lattices 96
5.2.2 Subgroup Lattices 96
5.2.3 Confidence Regions in Lattices 98
5.2.4 Lattice Bases 99
5.3 Estimation of Maximal Invariant Subgroups 99
5.3.1 Estimating Over a Finite Lattice 100
5.3.2 Computation of $\hat{K}_\ell$ 101
5.3.3 Computation of $\hat{G}$ 101
5.3.4 Choosing the Lattice Base 102
5.3.5 Choice of α_i 103
5.3.6 Theoretical Properties of $\hat{G}_\ell$ and $\hat{K}_\ell$ 104
5.4 Hypothesis Testing for a Particular Symmetry 104
5.4.1 Asymmetric Variation Test 104
5.4.2 Permutation Variant of the Asymmetric Variation Test 107
5.5 Numerical Experiments 110
5.6 Applications to Real Data 114
5.6.1 Satellite Based Magnetic Field Data 114
5.6.2 Sunspots and Spörer's Law 117
5.7 Discussion 118
5.8 Proofs and Further Mathematical Details 120
5.8.1 Proofs in Section 5.2 120
5.8.2 Proofs in Section 5.3 122
5.8.3 Proofs in Section 5.4 123
5.9 Related Topics 125
5.9.1 Estimating Lipschitz Constant of f 125
5.9.2 Further Details on the Experiments and Applications 126

6 Conclusion **133**
6.1 Future Work . . . 134
6.1.1 Coordinate Transformation Invariance . . . 134
6.1.2 G-Stable Estimators . . . 135
6.1.3 Other Methods for Estimating the Level Set . . . 135
6.1.4 Other Tests of Invariance . . . 136
6.2 Open Questions . . . 137
6.2.1 Which Estimators are Sufficiently Local? . . . 137
6.2.2 Rates of Estimation of Maximal Symmetries . . . 137
6.2.3 Which Symmetries and Distributions Satisfy Proposition 60? . . . 137

References **139**

Appendix A Other Useful Lemmas **145**

List of figures

1.1 Examples of functions exhibiting the different classes of level sets. The first is the function $f_1(x) = \sin(x_2)$ which is 1-sparse because it does not depend on the first coordinate. The second is $f_2(x) = \sin(\|x\|_2)$ which is symmetric because it is invariant to the set of rotations around the origin. The third is $f_3(x) = x_1^2 + x_1x_2 - x_2^2$ with smooth level sets (SLS). Blue colours represent the minimum value and yellow the maximum for each function on $[0,1]^2$. 4

1.2 Example level sets (bold lines) for the functions in Figure 1.1, at the point $x = (0.5, 0.5)$. 4

2.1 The orthogonal decomposition of $f = \sin(r)\cos(2\theta)^2$ into a G symmetric piece and a G-anti-symmetric piece, as described in Example 14. 27

2.2 Examples of three mice vertebrae landmarks measured at varying orientations, and an overlay of the three sets of landmarks after rotational alignment. All drawn from the R-package shapes [Dryden and Mardia, 2016] using data from Mardia and Dryden [1989]. 30

2.3 An example of 6 curves $f^{(i)} = \exp(-(x - t_i)^2) + \epsilon_i(x)$ as described in Example 16. The top left plot shows the original curves are measured, the top right shows the curves after alignment by translations of the time axis. The lower plots show the pointwise sample standard deviation of the original curves (left) and the aligned curves (right). This shows how the amplitude variation is masked by the phase variation if we were to deal only with the unaligned curves. 32

3.1 The level of $x \in \mathbb{R}^2$ for a function $f(x) = f_0(\|x\|_2)$, along with the points $\{a_i\}$ and $\{v_i = g^i_{x,h} \cdot x\}$ that will be used to manifold-average the estimator f_n. Note that these need not be spaced evenly around the level set, nor cover it completely. 48

3.2 A visualisation of the distance function $d_{x,h_O}(M, [x]_f)$ for a circular $[x]_f$ and elliptical M in red that meet at x. The distance between these manifolds is given by the distance $d(u, v)$ where the supremums are achieved, shown in blue. 53

3.3 The pointwise Mean Squared Error (MSE) for the estimator $f_n^{MAVE}(x, [x]_{p_n})$ at the point $x_0 = (0.5, 0.5)$. 57

4.1 Sub-figure (a) shows a heatmap of a regression function $f : \mathbb{R}^2 \to \mathbb{R}$. Sub-figure (b) depicts the dashed closed ball of radius h around the point $x \in \mathcal{X}$ and locations of some covariates, the data inside which (filled blue circles) are used to estimate $f(x)$ with a local polynomial estimator f_n with a kernel with support contained in this ball. The data outside this ball (red squares) are not used to estimate $f(x)$. Sub-figure (c) depicts our estimator which symmetrises with respect to $G = \{\text{Rotations around } 0\}$, using multiple balls of smaller radius with centres $\phi_i(x)$ spaced around the orbit $[x]_G$, chosen using the method descried in Section 4.2.1. Note that these need not be evenly spaced, nor cover the entire orbit. The symmetrised estimator uses more of the data points to estimate $f(x)$, and excludes one point that was previously used, shown by the colour changes. 63

4.2 The orbit of $x \in \mathbb{R}^2$ under the action of $G = SO(2)$, along with the points $\{a_i\}$ and $\{u_i = g^i_{x,h} \cdot x\}$ that will be used to partially symmetrise the estimator f_n. This is the same picture as Figure 3.1 but relabelled for the case of symmetry. 68

4.3 Example of the search space of the error minimising symmetry when $\mathcal{G} = SO(3)$, as in Example 40. As δ decreases, the middle stratum gains more points in the δ-cover. 76

4.4 Estimated risk for a baseline Local Constant Estimator f_n in blue against the symmetrised $S_{\rho_{\hat{G}_\delta}} f_n$ in orange, as well as 95% Wald confidence intervals for the true risk. Plots are on a log-log scale. 80

4.5 Estimated generalisation risk for a baseline Local Constant Estimator f_n in blue against the symmetrised $S_{\rho_{\hat{G}_\delta}} f_n$ when the same data is used in both steps of the M-estimation, as well as 95% Wald confidence intervals for the true risk. Plots are on a log-log scale. 81

5.1 Variable subset selection lattice of subsets, and symmetry estimation lattice of subgroups. The symmetry here is the translation action of $\Phi_{\mathbb{T}}$ on $\mathbb{R}^3$, and we have taken a finite sub-lattice of the lattice of closed subgroups where $\mathbb{R}_i$ is the translation group along the i^{th} orthonormal basis vector of $\mathbb{R}^3$. We have also included grey and dashed edges to other symmetries such as $\{x \mapsto x + a(e_2 + e_3) : a \in \mathbb{Z}\}$ and the larger symmetry $\{x \mapsto x + (a/2)(e_2 + e_3) : a \in \mathbb{Z}\}$. 94

5.2 The list of subgroups and their elements of the dihedral group D_4 (see section 2.3.1), and the Hasse diagram of the subgroup lattice. Each dot represents a subgroup, with edge from A down to B if $A \leq B$ and there is no other group that can be ordered between them. 97

5.3 Two examples of confidence regions for the subgroup $G_0 = \langle R_h, R_\pi \rangle$ of D_4 shown in blue with excess in red. Also shown is an example of a subset that is not a confidence region, which fails because it does not include all of the subgroups of its elements. The errors when ν is the counting measure are given underneath each picture. 98
5.4 This figure describes the results of the simulations in Example 61. Subfigures (a) and (b) track the confidence region excess $\mathcal{E}(\hat{K}_\ell)$ for each test used (black for the asymmetric variation test and red for the permutation variant), in dimension 2 for (a) and dimension 4 for (b). The error bars are Wald 95% Confidence Intervals for the true expected excess. 112
5.5 Mean squared prediction errors for the three estimators, under scenarios 1 and 2. The averages over 100 simulations are shown in black for $\hat{f}$, in blue for $\hat{f}^A_{\hat{G}}$, and in red for $\hat{f}^S_{\hat{G}}$. The error bars are Wald 95% Confidence Intervals for the true mean squared predictive error. 114
5.6 Mean squared prediction errors for the three estimators, under scenarios 3 and 4. The averages over 100 simulations are shown in black for $\hat{f}$, in blue for $\hat{f}^A_{\hat{G}}$, and in red for $\hat{f}^S_{\hat{G}}$. The error bars are Wald 95% Confidence Intervals for the true mean squared predicted error. 115
5.7 A visualisation of the measurements of the Earth's magnetic field, using the 1000 samples on February 25, 2023. In subfigure (a), red dots represent the high intensity measurements that occur towards the poles, and blue dots the regions of lowest intensity towards the equator. Red numbers indicate longitude and blue numbers represent latitude. In subfigure (b) intensity high magnetic field intensity is displayed in light blue and low intensity in dark blue. 116
5.8 Locations of all sunspots during the 21st recorded solar cycle, from the rotasym R package. The colours represent the time with the cycle, with blue towards the start and red towards the end. The blue numbers in subfigure (a) indicate latitude, the red longitude in degrees. Also shown is a marginal of time against latitude in subfigure (b) for 250 points uniformly sampled from this cycle. . . 117
5.9 A replica of Figure 5.4 for dimensions $d = 6$ and $d = 10$. 126
5.10 Plot of the frequency with which $G_{\max}(f, K_\ell) \in \hat{K}_\ell$ in the 200 simulations of Example 61. The black squares represent the results when using the asymmetric variation test and the red triangles for the permutation variant. The lower and upper blue lines represent the 0.7 and 0.95 confidence region error thresholds respectively. 127

5.11 Quantile-Quantile plots of $d_{\mathcal{X}}(X_i, X_j)$ vs $d_{\mathcal{X}}(g_i \cdot X_i, g_j \cdot X_j)$ for each of the groups in Section 5.6.2. Here the sampling of the group elements is from the uniform angle regime, where g_i has an rotational angle sampled from $U([0, 2\pi])$ in radians. Also presented are p-values for Kolmogorov-Smirnov tests for equality of the distributions. 130

5.12 Quantile-Quantile plots of $d_{\mathcal{X}}(X_i, X_j)$ vs $d_{\mathcal{X}}(g_i \cdot X_i, g_j \cdot X_j)$ for each of the groups in Section 5.6.2. Here the sampling of the group elements is from the uniform angle regime, where g_i has an rotational angle sampled from $U([0, 2\pi])$ in radians. Also presented are p-values for Kolmogorov-Smirnov tests for equality of the distributions. 131

List of tables

5.1 Proportions of correct symmetry estimation. 111
5.2 Confidence regions for each solar cycle 119
5.3 Proportions of rejections for the asymmetric variation test in Example 6.1. . . 128
5.4 Proportions of rejections for the permutation variant in Example 6.1. 128

List of Common Symbols

Other Symbols

$(\mathcal{X}, d_{\mathcal{X}})$ A metric space

$B_{\mathcal{X}}(x, h)$ The open ball of radius h, $B_{\mathcal{X}}(x, h) = \{y \in \mathcal{X} : d_{\mathcal{X}}(x, y) < h\}$

Operators and Relations

$\overset{a.s.}{\to}$ Almost sure convergence

$\lceil \cdot \rceil$ The ceiling function, the smallest integer not less than the input

$\overline{A}$ The topological closure of the set A

$\overset{D}{\to}$ Convergence in distribution

$\overset{D}{=}$ Equality in distribution

$\lfloor \cdot \rfloor$ The floor function, the largest integer not greater than the input

$\gtrsim$ Asymptotically slower, $a_n \gtrsim b_n$ if there is a positive constant c with $a_n \geq cb_n$ for all n

$\mathbf{1}_A$ An indicator function, $\mathbf{1}_A(x) = 1$ if $x \in A$ and 0 otherwise

$\asymp$ Asymptotically equivalent, $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$

$\lesssim$ Asymptotically faster, $a_n \lesssim b_n$ if there is a positive constant c with $a_n \leq cb_n$ for all n

$\overset{\mathbb{P}}{\to}$ Convergence in probability

$\setminus$ Difference of sets, $A \setminus B = \{a \in A : a \notin B\}$

Spaces of Numbers or Functions

$C^k(\mathcal{X})$ Functions that are k times continuously differentiable

$\mathbb{R}$ Real numbers

$\mathbb{R}_{\geq 0}$ Non-negative real numbers

$\mathbb{R}_{>0}$ Positive real numbers

$\mathbb{Z}$ Integers, $\{\dots, -2, -1, 0, 1, 2, \dots\}$

$\mathbb{Z}_{\geq 0}$ Non-negative integers, $\{0, 1, 2, \dots\}$

$\mathbb{Z}_{>0}$ Positive integers, $\{1, 2, 3, \dots\}$

Acronyms and Abbreviations

BSE Best Symmetric Estimator

EMS (Global) Error Minimising Symmetry

LCE Local Constant Estimator

LEMS Local Error Minimising Symmetry

LPE Local Polynomial Estimator

MAVE Manifold AVeraged Estimator

PSE Partially Symmetric Estimator

Chapter 1

Introduction

Regression, where we seek to understand the relationship between measured independent variables $X = (X_1, \ldots, X_d)$ in some space $\mathcal{X}$ and response variables $Y \in \mathbb{R}$ given finite samples $\{(X^{(i)}, Y_i)\}_{i=1}^n$, is a core task in statistics. In its simplest form we might seek to find straight lines of best fit through our samples $(X^{(i)}, Y_i)$ when $X^{(i)} \in \mathbb{R}$. This is fundamentally a search for a pattern in the data: if we know how much we expect Y_i to increase with each unit of $X^{(i)}$ and the expected value at some chosen base point x, then we can extrapolate out to any new point u using this pattern and give a useful prediction of our expected response. If we expand the relationships beyond linear ones, many other patterns are possible: we see periodicity in sound waves or ripple heights, or we see that image classifications can stay constant when we rotate or reflect an image. In this thesis we will show how methods in nonparametric regression can be improved by explicitly learning these patterns and can provide dimension reduction by compressing the covariate space to a region that we can extrapolate from using the pattern.

The usual approach to regression is to assume an additive noise model of the form:

$$Y_i = f(X^{(i)}) + \epsilon_i \tag{1.1}$$

for functions f that are contained in some hypothesis class $\mathcal{F}$, such as: the class of linear functions $f(x) = x^T\beta$, the class of polynomial functions of the covariates, or the class of all L-Lipschitz functions. Here all $X^{(i)}$ and ϵ_i are independent and follow some known distributions. We can then estimate f, which will be the conditional mean of $Y_i \mid X^{(i)}$ when $\mathbb{E}(\epsilon_i) = 0$, by using least squares (for example) over the class $\mathcal{F}$:

$$\hat{f} = \operatorname*{argmin}_{\phi \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - \phi(X^{(i)}))^2 \tag{1.2}$$

When $\mathcal{F}$ can be described with a finite number of parameters, as with the class of linear functions $f(x) = x^T\beta$ for some $\beta \in \mathbb{R}^d$, this usually has an explicit closed form solution

that converges to f in expected L^2-norm at a rate of n^{-1}. When this class becomes infinite dimensional (as with the L-Lipschitz class), we typically focus on methods that are inherently local rather than global such as local polynomial estimators or wavelet projection estimators, all defined in Section 2.1. The locality of these methods means that the data used to estimate the function at each point is much more limited than in the parametric case. This leads to slower rates of convergence in this more general setting, and in particular these rates become much slower as dimension increases.

In the statistics literature there are three main methods for overcoming this problem that all seek to find more efficient representations of the covariate space. The first is to ignore the response values and represent the covariates on some lower dimensional submanifold of $\mathcal{X}$, in essence performing a non-linear analogue to principal component analysis. This approach is sensible if the covariates are highly concentrated around this manifold as in Bickel and Li [2007] but struggles when the covariate space is sampled from more uniformly and will completely miss any behaviour of the regression function off of this submanifold. The second approach is to assume that the regression function has an additive structure: where it is a sum of functions of each of the covariates separately. This speeds up the convergence of these estimators significantly but leads to asymptotic bias for functions that are not additive. The last main approach is to assume that the regression function depends only on some, rather than all, of the covariates. When this sparsity assumption is true it will lead to faster rates (even adaptively, see Lafferty and Wasserman [2008]) but many regression functions do depend on all of the covariates and so this approach fails for them.

In this thesis, we present a new method for achieving dimension reduction and speeding up non-parametric regression. This approach considers the level sets of the regression function $[x]_f = \{u \in \mathcal{X} : f(u) = f(x)\}$, which usually[1] form $(d-1)$- or d-dimensional sub-manifolds of the covariate space. Knowledge of these level sets allows us to construct a Manifold AVeraged Estimator (MAVE):

$$f_n^{MAVE}(x) = \frac{1}{m}\sum_{j=1}^{m} f_n(x_j) \tag{1.3}$$

that averages a base estimator f_n (e.g. a local polynomial estimator, or a neural network etc.) over some chosen points $x_j \in [x]_f$. If the base estimator is sufficiently local[2] (which is guaranteed for local polynomials), this significantly reduces the variance of the MAVE relative to the base estimator. This allows us to re-optimise the bias-variance trade off and achieve dimension reduction. The dimensions reduced are precisely those of the level set, which

[1] This is made precise in chapter 3, where we give sufficient conditions for this to be an almost sure statement.

[2] Here sufficiently local means that $f_n(x)$ is independent of $f_n(y)$ for points x, y that are sufficiently separated for large n, and is defined precisely in assumption set (E) later.

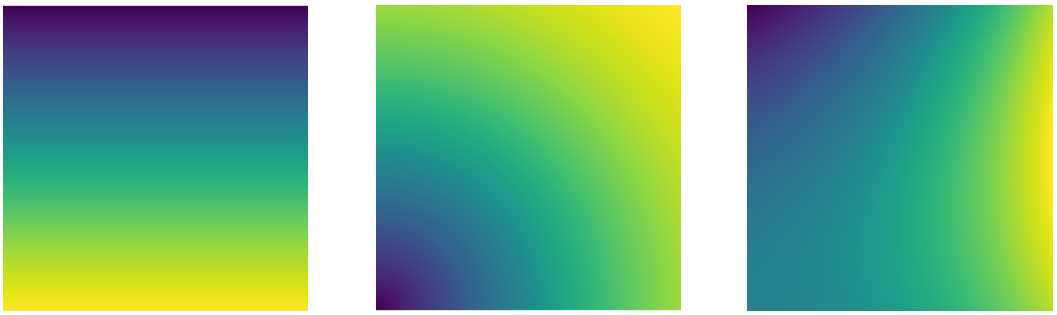
brings convergence to the same rate as for additive hypothesis classes or even parametric rates (if x is an interior point of a d-dimensional level set).

When the level sets are unknown, one might hope that they can be estimated sufficiently quickly to achieve this dimension reduction adaptively. In complete generality this is unlikely to be true because estimating the level sets requires information about the derivatives of f, which are slower to estimate than the rate we seek to improve on. Thus, in this thesis we focus on specific structures for submanifolds contained in the level sets of f. For example, we can ask if we can estimate the largest hyperplane contained in $[x]_f$. If these submanifolds can be estimated sufficiently quickly, then we can hope that we can improve the rate of our estimator by the dimension of these structures. Our estimation strategy is then straight forward: estimate the largest appropriate structure within the level sets and then use these for the MAVE.

This idea of estimating the largest hyperplane is in fact well studied from a different perspective: this is precisely the idea of sparsity (or more generally multi-index models). To see this, simply note that if f does not depend on the coordinate x_i, then the line $\{x + ae_i : a \in \mathbb{R}\}$ must be completely contained inside $[x]_f$. The benefit of the submanifold perspective is that we can think to expand to classes of non-linear structures too. By doing so we might lose the speed at which we can estimate these level sets, and it is not immediately clear how general we can make the class of submanifolds before we lose the ability to adaptively achieve dimension reduction. In this thesis, we present the framework of **symmetry**, where we show that considering the orbits of smooth proper actions[3] of locally compact groups leads to a set of possible level set sub-manifolds that is sufficiently constrained to achieve dimension reduction adaptively.

We say that a function f is invariant under the action of a group G if $f(x) = f(g \cdot x)$ for all $x \in \mathcal{X}$ and $g \in G$. If f is G-invariant, then the orbit $[x]_G = \{g \cdot x : g \in G\}$ must be completely contained inside the level set $[x]_f$. If G is acting smoothly and properly, then these orbits are smoothly embedded submanifolds and we can apply the MAVE. This provides a natural extension to sparsity and multi-index models, as the hyperplanes used to reduce the dimension are really just orbits under a translation action on the covariate space. However, we can easily consider other group actions such as rotations, or all rigid transformations (i.e., the group generated by all translations and rotations), which can now capture nonlinearity in the level sets. These group actions are patterns in the function f and they tell us how to borrow strength between estimates at different regions of the covariate space. Functions that fit into each of these classes are shown in Figure 1.1. The level sets of these functions are

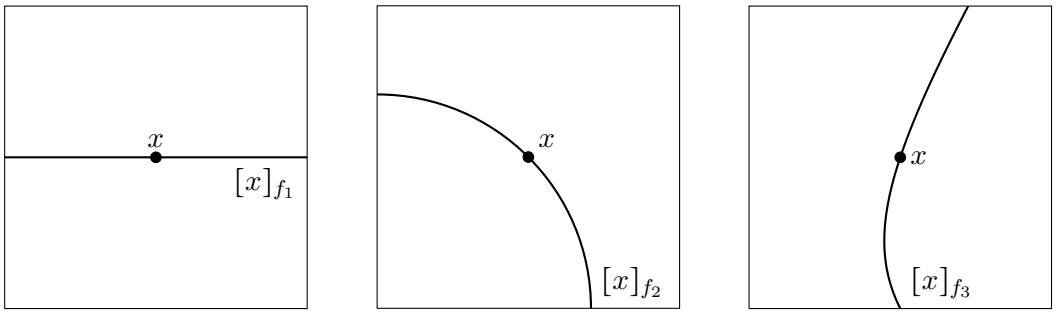
[3] A proper action is one for which the set of $g \in G$ such that $K \cap (g \cdot K') \neq \varnothing$ is compact for any two compact subsets $K, K' \subseteq \mathcal{X}$. Further definitions can be found in section 2.3.



(a) Sparse function (b) Symmetric Function (c) SLS Function

Fig. 1.1 Examples of functions exhibiting the different classes of level sets. The first is the function $f_1(x) = \sin(x_2)$ which is 1-sparse because it does not depend on the first coordinate. The second is $f_2(x) = \sin(\|x\|_2)$ which is symmetric because it is invariant to the set of rotations around the origin. The third is $f_3(x) = x_1^2 + x_1x_2 - x_2^2$ with smooth level sets (SLS). Blue colours represent the minimum value and yellow the maximum for each function on $[0,1]^2$.

shown in Figure 1.2.



(a) Sparse function (b) Symmetric Function (c) SLS Function

Fig. 1.2 Example level sets (bold lines) for the functions in Figure 1.1, at the point $x = (0.5, 0.5)$.

Thus, we have three set of structures within the level set $[x]_f$, in increasing generality:

$$\left\{\text{Hyperplanes through } x\right\} \subseteq \left\{\text{Orbits of } x\right\} \subseteq \left\{\text{Submanifolds containing } x\right\} \tag{1.4}$$

Sparsity (and the generalisation to multi-index models, see the introduction in Chapter 4) asks the question: "*what is the largest hyperplane through x contained in $[x]_f$?*". This question is known to be answerable sufficiently quickly to be useful for estimating f, allowing for adaptive removal of the dimension of the true largest hyperplane. In Chapter 3 of this thesis, we ask whether this is true of the question: "*what is the largest submanifold through x contained in $[x]_f$?*"? We show that it can have a positive answer under some relatively strong

conditions. In Chapter 4 of this thesis, we consider the intermediate case: "*what is the largest orbit of some subgroup of $\mathcal{G}$ through x contained in $[x]_f$?*". We show that estimating this largest symmetry of $[x]_f$ improves performance both asymptotically and in finite samples, and can achieve adaptive dimension reduction without the strong assumptions of Chapter 3. Lastly in Chapter 5, we consider inference for these symmetries of the level sets, giving both hypothesis tests for the invariance to particular symmetries and a method for piecing these tests together to estimate the largest applicable symmetry. In the following sections we describe briefly the content in each chapter and how we achieve these results.

Chapter 2: Background

We start with a summary of: the nonparametric regression problem; the curse of dimensionality; and classical methods, such as local polynomial regression and smoothing splines, and their properties. We cover the mathematical background from differential geometry including the definition of manifolds and how these become metric spaces. We then present some group theory material to give us the language we will use to describe the symmetries of regression functions. We give several examples of how this group theory is used in statistics both implicitly through permutation tests through to explicitly in functional data registration. We then cover some functional analysis of functions that obey symmetries; in particular, we introduce the symmetrisation operator that projects functions in $L^2(\mathcal{X})$ to the closed linear subspace of invariant functions. We describe how this operator (among other techniques) is being used in regression to improve estimators in the machine learning literature and cover some methods that attempt to discover symmetries from training data. Lastly, we mention some methods used to estimate the sub-level sets of regression functions.

Chapter 3: Regression using Level Sets

In this chapter, we give a formal treatment of the manifold averaged estimator (MAVE). We show how to deterministically construct points on the level set for smoothing the base estimator and that this construction leads to faster pointwise mean squared error decay. We consider the robustness of this estimator to misspecification of the level set and show that any introduced bias is related to a local Hausdorff distance between the true level set and the manifold used. We then show that even a naive estimate of the level set can lead to adaptive dimension reduction, though only under some strong assumptions on the regression function. We demonstrate this performance on some synthetic data.

Chapter 4: M-Estimation of Maximal Symmetries

We then move to consider estimating only the largest orbit of a subgroup of some search group $\mathcal{G}$ contained within the level set $[x]_f$. We strengthen the pointwise convergence of the MAVE from the previous chapter to give rates for the L^2-risk of the MAVE when averaging over the orbits $[x]_G$ in random design contexts. We give a metric structure on the space of subgroups of a locally compact search group $\mathcal{G}$ and show that this space is totally bounded. This allows us to define an M-estimator of the maximal symmetry of the regression function f over finite δ-covers of this subgroups space. We then show that using the orbits of this estimated symmetry leads to adaptive dimension reduction on the order of the highest dimensional orbits contained in the level sets $[x]_f$. This is then demonstrated on synthetic data for both rotation symmetries and translation symmetries.

Chapter 5: Estimation of Maximal Symmetries via Subgroup Lattices

The last chapter constructs an alternative estimator of the maximal symmetry of f that is based on hypothesis tests for $H_0(G) : f$ *is* G*-invariant*. The idea is that we can consider the space of subgroups as a *lattice*, a kind of directed acyclic graph that has additional structure. This structure allows us to reduce the problem to conducting tests along a *lattice base* from which we can extrapolate to the rest of the lattice. We show that the tests of $H_0(G)$ for the G in this lattice base lead to a confidence region within this lattice for the maximal symmetry of f, and that the unique largest element in this region gives a consistent estimate of the maximal symmetry (within the lattice considered, when the test is consistent). We present two options for hypothesis tests: one that is shown to be consistent but that requires knowledge of the noise distribution and one that does not require this at the cost of computational power and guaranteed consistency. We demonstrate these tests and the performance of the maximal symmetry estimator on synthetic data and on two collections of real data. The first examines the Earth's magnetic field and is able to reject the hypothesis that it obeys the same symmetries as an idealised bar magnet.[4] The second considers the symmetries of the distribution of sunspots. We are able to show that whilst the distribution of these sunspots appears rotationally symmetric when averaged over an entire solar cycle, the distributions conditioned on time often do not exhibit the same symmetry.

[4]This agrees broadly with the geophysical literature, which hypothesizes that this is due to a combination of non-uniform distribution of iron in the Earth's crust and solar winds.

Chapter 2

Background

In this chapter, we give the required background material for the content of this thesis. This covers four main topics: statistics; geometry; group theory; and functional analysis, as well as the connections between them. We assume that the reader has an understanding of topology, including notions of neighbourhoods and homeomorphisms, and knowledge of the adjectives: Hausdorff; second countable; compact; and locally compact. All other mathematical notation used is covered in the list of symbols before the introduction.

2.1 Nonparametric Regression

Let $\mathcal{X}$ be the closure of any open subset of $\mathbb{R}^d$ or any other orientable smooth d-dimensional manifold (defined in Section 2.2) such as the unit sphere $S^d = \{x \in \mathbb{R}^{d+1} : \|x\|_2 = 1\}$. Let μ be any Borelian probability measure (with respect to the topology on $\mathcal{X}$) that has support $\text{supp}(\mu) = \mathcal{X}$. Suppose we collect data $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$, and where:

$$Y_i = f(X_i) + \epsilon_i \tag{2.1}$$

where ϵ_i are mean zero finite variance random variables that are independent of each other and of all X_i, and f is an unknown regression function. The question in regression problems is to estimate f with some $\hat{f}$ and to do so with minimal pointwise mean square error $MSE(\hat{f}(x)) = \mathbb{E}((\hat{f}(x) - f(x))^2)$ or minimal squared L^2 risk $\mathbb{E}(\|f - \hat{f}\|_2^2)$. We usually assume that we have a *hypothesis class* $\mathcal{F} \subseteq L^2(\mathcal{X})$ of functions that f can belong to. This allows us to use the information common to functions in $\mathcal{F}$ to design better estimators of f.

One simple class is that of linear functions on, for example, $\mathcal{X} = [0,1]^d$:

$$\mathcal{F}_{\text{Linear}} = \left\{x \mapsto x^T\theta : \theta \in \mathbb{R}^d\right\} \tag{2.2}$$

Estimation of functions in this class can be done with a straightforward least squares estimator:

$$\hat{f}(x) = x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y \tag{2.3}$$

where $\mathbf{X} = [X_1, \dots, X_n]^T$ is the design matrix. Undergraduate statisticians learn that when $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ this estimator will have:

$$\mathbb{E}(\|f - \hat{f}\|_2^2) \leq Cn^{-1} \tag{2.4}$$

for some positive constant $C \in \mathbb{R}_{>0}$ under appropriate design conditions and for sufficiently large n. The next simplest class is to consider higher order polynomials, up to some degree k:

$$\mathcal{F}_{\text{Polynomial}}(k) = \left\{ x \mapsto \sum_{\alpha : |\alpha| \leq k} \theta_\alpha x^\alpha : \theta_\alpha \in \mathbb{R} \right\} \tag{2.5}$$

where α is a multi-index $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{Z}_{\geq 0}$ with $|\alpha| = \sum |\alpha_i|$, and $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$. We can estimate $f \in \mathcal{F}_{\text{Polynomial}}(k)$ using least squares again, but this time we have to embed our design matrix in a higher dimensional space. We do this by picking a basis $\{e_i\}_{i=1}^{\ell}$ for the polynomial space $\mathcal{F}_{\text{Polynomial}}(k)$ (usually the **monomial basis** given by $\{1, x_1, \dots, x_d, x_1^2, x_1 x_2, \dots\}$) and embedding $\mathbf{X}$ in the *Vandermonde matrix*:

$$U^k(\mathbf{X}) = \begin{bmatrix} e_1(X_1) & \cdots & e_\ell(X_1) \\ \vdots & \ddots & \vdots \\ e_1(X_n) & \cdots & e_\ell(X_n) \end{bmatrix} \tag{2.6}$$

We can then estimate f again using least squares:

$$\hat{f}(x) = U^k(x)\big(U^k(\mathbf{X})^T U^k(\mathbf{X})\big)^{-1} U^k(\mathbf{X})^T Y \tag{2.7}$$

which has the same rate convergence as in equation (2.4) (in terms of n).

These classes are called **parametric** because they can be completely described by a finite number of parameters in the vectors θ. For such classes one can prove[1] that this is the best attainable rate, in the sense that:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_{\text{Polynomial}}(k)} \mathbb{E}\big(\|f - \hat{f}\|_2^2\big) \geq \inf_{\hat{f}} \sup_{f \in \mathcal{F}_{\text{Linear}}(k)} \mathbb{E}\big(\|f - \hat{f}\|_2^2\big) \geq cn^{-1} \tag{2.8}$$

[1] Essentially as a consequence of the Cramér-Rao lower bound, adapted to cover biased estimators.

for some $c > 0$. We can broaden these classes significantly to include all functions that are Lipschitz, for example,

$$\mathcal{F}_{\text{Lipschitz}}(L) = \left\{ f \in L^2(\mathcal{X}) : |f(x) - f(y)| \leq d_{\mathcal{X}}(x, y) \right\} \tag{2.9}$$

This class contains subsets of the polynomial (and linear class) that have derivatives bounded by L. However, it also contains many other non-polynomial functions such as $x \mapsto L \sin(\|x\|_2)$. Because of these extra functions, this class cannot be identified with a finite number of parameters and so we call them **nonparametric**. Computing least squares estimators is much more computationally challenging for these classes. Moreover, one can establish a similar lower bound to equation (2.8):

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_{\text{Lipschitz}}(L)} \mathbb{E}\left(\|f - \hat{f}\|_2^2 \right) \geq c n^{-\frac{2}{2+d}} \tag{2.10}$$

for some positive constant $c \in \mathbb{R}_{\geq 0}$ (using Fano's Lemma, see Tsybakov [2008]). This rate is slower than the parametric rate for $d > 0$, representing the difficulty of estimating within this broader class. Another family of classes we will consider are the **Hölder Classes**:

$$\mathcal{F}_{\text{Hölder}}(L, \beta) = \left\{ f \in L^2(\mathcal{X}) \cap C^k(\mathcal{X}) : \sum_{\alpha : 0 \leq |\alpha| \leq k} \|\partial^\alpha f\|_\infty + \sum_{\alpha : |\alpha| = k} \sup_{x \neq y \in \mathcal{X}} \frac{|(\partial^\alpha f)(x) - (\partial^\alpha f)(y)|}{d(x,y)^{\beta - k}} \leq L \right\} \tag{2.11}$$

for each $L, \beta \in \mathbb{R}_{>0}$ and where $k = \lceil \beta \rceil - 1$. This class has a similar lower bound:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\left(\|f - \hat{f}\|_2^2 \right) \geq c n^{-\frac{2\beta}{2\beta + d}} \tag{2.12}$$

where $\mathcal{F} = \mathcal{F}_{\text{Hölder}}(L, \beta)$. When $\beta = 1$ this is just the Lipschitz class but when $\beta > 1$ the derivatives of f allow us to better estimate f, and as $\beta \to \infty$ (for fixed d) this rate exponent converges to the parametric rate of n^{-1}. Other classes include the Sobolev classes and Besov classes. These classes all have slower rates than the parametric classes, and all exhibit the same dependence on dimension in these lower bounds. We will soon describe estimators that achieve this worst case risk decay with this rate when $\mathcal{X} = [0, 1]^d$, which show that these are so called **minimax** rates of estimation for the problem of estimating over these classes.

This dependence on dimension means that we suffer from a *curse of dimensionality*, which means that the number of samples required to bring the worst case risk below a specified $\varepsilon > 0$ grows exponentially in d. This is markedly different from the classical parametric situation where there is no dependence on dimension and represents the main problem we seek to address in this thesis.

2.1.1 Local Polynomial Estimators

The first technique we will examine adapts the polynomial regression above to non parametric hypothesis classes. The key idea is that if we want to estimate $f(x)$ then we should try to estimate a Taylor series for f at the point x. We can do this by centering our polynomial estimator above and excluding all data points outside some local ball of radius h, called the **bandwidth**, denoted $B_{\mathcal{X}}(x,h) = \{u \in \mathcal{X} : d_{\mathcal{X}}(x,u) < h\}$. The degree of this approximation should be the number of derivatives of f, which we take from the properties of our hypothesis class $\mathcal{F}$. This reasoning leads to an estimator of the form:

$$\hat{f}_{k,h}(x) = U^k_{x,h}(x)\big(U^k_{x,h}(\mathbf{X})^T W_{x,h} U^k_{x,h}(\mathbf{X})\big)^{-1} U^k_{x,h}(\mathbf{X})^T W_{x,h} Y \tag{2.13}$$

where $U^k_{x,h}(\mathbf{X}) = U^k((\mathbf{X}-x)/h)$ and W is a diagonal matrix with $W_{ii} = h^{-1}\mathbf{1}_{X_i \in B_{\mathcal{X}}(x,h)}$. We typically use the monomial basis here for ease of applying Taylor's theorem to control the bias. Also note that we could, of course, weight the contributions by factors that depend on a distance from x, but this is not necessary for this thesis.

This estimator has attractive properties: it is linear in Y, which aids in the theoretical analysis of this estimator; it is highly interpretable because we can say precisely how the data affects the estimates; and it is strongly local, in the sense that for points that have $d_{\mathcal{X}}(x,y) \geq 2h$, their estimates are independent after conditioning on the covariates.

The theoretical analysis of this estimator is due to Stone [1977]. There it is shown that this estimator has:

$$\text{Bias}(\hat{f}_{k,h}(x)) \leq Bh^{\beta} \tag{2.14}$$

$$\text{Var}(\hat{f}_{k,h}(x)) \leq \frac{V}{nh^d} \tag{2.15}$$

for some positive constants B and V, when $f \in \mathcal{F}_{H\ddot{o}lder}(L,\beta)$ under appropriate conditions on the design regime. The bias bound comes from the fact that the approximation error of the Taylor polynomial of f at the point x is controlled on the order of h^{β} for regression functions in the Hölder class. The variance bound comes from the fact that the expected number of data points with $X_i \in B_{\mathcal{X}}(x,h)$ is on the order of $n \times \text{Vol}(B_{\mathcal{X}}(x,h)) \asymp (nh^d)^{-1}$ when $X_i \overset{iid}{\sim} U([0,1]^d)$ or under many other design regimes.

Now the pointwise mean squared error of $\hat{f}_{k,h}(x)$ is the sum of the variance and the squared bias, so we can use the bounds above to see:

$$\mathbb{E}\big((\hat{f}_{k,h}(x) - f(x))^2\big) \leq B^2 h^{2\beta} + \frac{V}{nh^d} \tag{2.16}$$

for all $f \in \mathcal{F}_{Hölder}(L, \beta)$ and sufficiently large n under some conditions on the design regime. We can then optimise this bound over our choice of $h \in \mathbb{R}_{>0}$, leading to an optimal bandwidth selection of $h = an^{-1/(2\beta+d)}$, and thus we get the bound:

$$\sup_{f \in \mathcal{F}} \mathbb{E}\big((\hat{f}(x) - f(x))^2\big) \leq Cn^{-\frac{2\beta}{2\beta+d}} \tag{2.17}$$

for sufficiently large n, where $\mathcal{F} = \mathcal{F}_{Hölder}(L, \beta)$ and C is a positive constant that depends only on L, β, and d. Thus, this estimator achieves the minimax rate of equation (2.12). This strategy does require us to know β, both for the computation of the optimal bandwidth and for the selection of the degree of the polynomials that we fit.

We will later also need to understand the convergence of the polynomial coefficients. Suppose for now that the noise is Gaussian and homoskedastic with variance σ^2, and that $\mathcal{X} = [0,1]^d$. Let $f(\mathbf{X}) = [f(X_1), \ldots, f(X_n)]$. If we define the estimate:

$$\hat{\theta} = \big(U^k_{x,h}(\mathbf{X})^T W_{x,h} U^k_{x,h}(\mathbf{X})\big)^{-1} U^k_{x,h}(\mathbf{X})^T W_{x,h} Y \tag{2.18}$$

then we have:

$$\hat{\theta} \sim N\big(A_{n,x,h} f(\mathbf{X}),\ \sigma^2 (U^k_{x,h}(\mathbf{X})^T W_{x,h} U_{x,h}(\mathbf{X}))^{-1}\big) \tag{2.19}$$

where $A_{n,x,h} = (U^k_{x,h}(\mathbf{X})^T W_{x,h} U^k_{x,h}(\mathbf{X}))^{-1} U^k_{x,h}(\mathbf{X})^T W_{x,h}$. Asymptotically, if:

$$\tfrac{1}{nh^d}\big(U^k_{x,h}(\mathbf{X})^T W_{x,h} U^k_{x,h}(\mathbf{X})\big) \to Q \tag{2.20}$$

for some positive definite matrix Q, then the distribution above allows us to show that:

$$\mathbb{E}(\|\hat{\theta} - A_{n,x,h} f(\mathbf{X})\|_2^2) \leq \tfrac{C}{nh^d} \tag{2.21}$$

for sufficiently large n. This also holds for many other choices of noise distribution, including all sub-Gaussian distributions because this MSE depends only on the first two moments of $\hat{\theta}$. Note that the constant depends on the minimal eigenvalue of Q which depends on the design regime, but it does not depend on f. This assumption is usually ensured by conditions on the design regime and is satisfied by most regimes that are "sufficiently uniform", such as a rectangular design with points $X^{(i)}$ spaced evenly along each coordinate axis. If f restricted to $B(x,h)$ is equal to a polynomial $p(u) = U^k_{x,h}(u)\theta_{p,x,h}$ on this ball, then:

$$A_{n,x,h} f(\mathbf{X}) = (U^k_{x,h}(\mathbf{X})^T W_{x,h} U^k_{x,h}(\mathbf{X}))^{-1} U^k_{x,h}(\mathbf{X})^T W_{x,h} U^k_{x,h}(\mathbf{X}) \theta_{p,x,h} = \theta_{p,x,h} \tag{2.22}$$

This is a property known as the "reproducing polynomial effect". This is the analysis that leads to the bias above above; f is equal to its Taylor series plus a term that decays on the

required order of h^β. For such functions f, we would thus have:

$$\mathbb{E}(\|\hat{\theta} - \theta_{p,x,h}\|_2^2) \leq \frac{C}{nh^d} \tag{2.23}$$

Note that the coefficients in $\theta_{p,x,h}$ depend on the center x and scale h of $U_{x,h}^k$ even if the polynomial p does not.

2.1.2 Projection Estimators

Another approach utilises Fourier analysis, noting that $L^2(\mathcal{X})$ is a separable Hilbert space. This allows us to chose an orthonormal basis $\{\phi_i\}_{i=1}^\infty$ for $L^2(\mathcal{X})$, in which we can express our regression function as:

$$f(x) = \sum_{i=1}^{\infty} \langle f, \phi_i \rangle \phi_i(x) \tag{2.24}$$

if the sum of the squares of the coefficients $\theta_j = \langle f, \phi_j \rangle$ is convergent. Each coefficient θ_j in this expansion then has a natural estimator:

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(X_i) \approx \int_{\mathcal{X}} f(x) \phi_j(x) \mu(\,\mathrm{d}x) = \theta_j \tag{2.25}$$

when the covariates are sampled iid from the integrating measure in our L^2 space or follow an appropriate fixed design regime. Each estimate can be very simply analysed:

$$\mathbb{E}(\hat{\theta}_j \mid \mathcal{D}_X) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \phi_j(X_i) \tag{2.26}$$

$$\mathrm{Var}(\hat{\theta}_j \mid \mathcal{D}_X) = \mathrm{Var}\Big(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi_j(X_i)\Big) = \frac{\sigma^2}{n^2} \sum_{i=1}^{n} \phi_j(X_i)^2 \tag{2.27}$$

Thus, the conditional bias is just the error from having finite data rather than integrating over the space, and the conditional variance is just $\sigma^2 n^{-1}$ but scaled by the approximation of $\|\phi_j\|_2^2 = 1$ using the discrete integral over the finite data. When $X_i \overset{iid}{\sim} \mu$, we can take expectations further and find:

$$\mathbb{E}(\hat{\theta}_j) = \langle f, \phi_j \rangle \tag{2.28}$$

$$\mathrm{Var}(\hat{\theta}_j) = \mathbb{E}\Big(\mathrm{Var}(\hat{\theta}_j \mid \mathcal{D}_X)\Big) + \mathrm{Var}\Big(\mathbb{E}(\hat{\theta}_j \mid \mathcal{D}_X)\Big) \tag{2.29}$$

$$= \frac{1}{n}\Big(\sigma^2 + \langle f^2, \phi_j^2 \rangle - \langle f, \phi_j \rangle^2\Big) \tag{2.30}$$

Thus, each estimator is unbiased and converges in distribution to the true value. We can then use these estimators to form an **order N projection estimator** of f:

$$\hat{f}_N(x) = \sum_{j=1}^{N} \hat{\theta}_j \phi_j(x) \tag{2.31}$$

The integrated mean squared error of this estimator then can be decomposed using the orthogonality of the basis into the sum of the errors of the N coefficients used and the bias from the remaining coefficients:

$$\mathbb{E}(\|\hat{f}_N - f\|_2^2) = \sum_{j=1}^{N} \mathbb{E}\big((\hat{\theta}_j - \theta_j)^2\big) + \sum_{j=N+1}^{\infty} \theta_j^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{N}(\sigma^2 + \langle f^2, \phi_j^2\rangle - \theta_j^2)}_{\text{Variance term}} + \underbrace{\sum_{j=N+1}^{\infty} \theta_j^2}_{\text{Bias}^2 \text{ term}} \tag{2.32}$$

(note that these terms are not *precisely* the variance and squared bias of this estimator, but they do behave much like the are in the analysis that follows). If we make a concrete choice and consider the trigonometric basis for the space $L^2([0,1])$:

$$\phi_1(x) = 1 \tag{2.33}$$

$$\phi_{2j} = \sqrt{2}\cos(2\pi j x) \tag{2.34}$$

$$\phi_{2j+1}(x) = \sqrt{2}\sin(2\pi j x) \tag{2.35}$$

then we can describe a functional class for integer β and positive real L given by:

$$\mathcal{F}_{Periodic}(L, \beta) = \Big\{ \sum_{j=1}^{\infty} \theta_j \phi_j \in L^2(\mathcal{X}) : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq L \Big\} \tag{2.36}$$

where $a_j = j^\beta$ if j is even and $a_j = (j-1)^\beta$ if j is odd. This class can be shown to be precisely identified with the periodic Sobolev class (Tsybakov [2008], Proposition 1.14). When considering this functional class, we can bound the squared bias term using the monotonicity of the a terms:

$$\sum_{j=N+1}^{\infty} \theta_j^2 \leq a_{N+1}^{-2} \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq L a_{N+1}^{-2} \in O(N^{-2\beta}) \tag{2.37}$$

Thus, we have a very similar situation to the case of local polynomial regression, where the squared bias is controlled on the order $(1/N)^{2\beta}$. The variance term can similarly bounded if we assume that $\sup_{j \in \mathbb{Z}_{>0}} \langle f^2, \phi_j^2\rangle - \theta_j^2 \leq A$, in which case:

$$\frac{1}{n}\sum_{j=1}^{N}(\sigma^2 + \langle f^2, \phi_j^2\rangle - \theta_j^2) \leq \frac{(\sigma^2 + A)N}{n} \in O(1/(nN^{-1})) \tag{2.38}$$

again showing that N^{-1} behaves like the choice of h for local polynomials. Thus, an optimal choice of N in this case would be

$$N = \lfloor a n^{\frac{1}{2\beta+d}} \rfloor \tag{2.39}$$

note that here $d = 1$ only, and extensions to higher dimensional analogues take some work. Note that many other choices of basis functions are possible, including polynomial bases or wavelet bases.

One can also show that the estimator of the coefficients $(\theta_1, \ldots, \theta_N)$ (at least in the trigonometric basis) is the solution to the minimisation problem:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^N}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{N} (\theta_j^T \phi_j(X_i)) \right)^2 \tag{2.40}$$

This approach generalises quite broadly, and other solutions are called **nonparametric least squares estimators**. It also allows for regularisation approaches, where we can penalise the minimisation function with $\sum_{j=1}^{N} \lambda_j \theta_j^2$ or $\sum_{j=1}^{N} \lambda_j |\theta_j|$, much akin to parametric alternatives of ridge regression [Hoerl and Kennard, 1970] and the Lasso [Tibshirani, 1996].

2.1.3 Cubic Smoothing Splines

The last major technique in nonparametric regression is to fit a smoothing spline, which we again consider in the one dimensional case here, where $\mathcal{X} = [a, b] \subseteq \mathbb{R}$. We say that a function $m : \mathcal{X} \to \mathbb{R}$ is a natural cubic spline with knots $x_1, \ldots, x_n$ if: (1) the restriction to each interval $[x_i, x_{i+1}]$ is a cubic polynomial; (2) m has a continuous second derivative on (a, b); and (3) m is linear on the intervals $[a, x_1]$ and $[x_n, b]$. Note that (2) completely determines the linear components. It can be shown that a natural cubic spline minimises $\int_{\mathcal{X}} m''(x)^2 \mathrm{d}x$ over all absolutely continuous functions. This leads to a very natural estimator of a non-parametric regression function with data $\{(X_i, Y_i)\}_{i=1}^n$:

$$f_n = \underset{m \in \mathcal{S}^2([a,b])}{\operatorname{argmin}} \sum_{i=1}^{m} (Y_i - m(X_i))^2 + \lambda \int_a^b m''(x)^2 \, \mathrm{d}x \tag{2.41}$$

where $\mathcal{S}^2([a, b])$ is the space of functions with absolutely continuous first derivatives. It can be shown that this minimiser exists for all $\lambda \in \mathbb{R}_{>0}$ and is a natural cubic smoothing spline with knots at $X_1, \ldots, X_n$. The parameter λ controls the balance between fitting the data and the smoothness of f_n. Smaller λ leads to a closer fit but higher variance, and larger λ smooths out the estimate. This hyper-parameter can be fitted automatically with cross-validation. This approach, as with projection estimators, has a computational benefit over local polynomial fitting because it finds a global solution in one step rather than computing $f_n(x)$ at each point x separately. This approach can be extended to higher dimensions, as with thin plate splines [Duchon, 1977] where the second derivative term is replaced by the Laplacian of m.

2.1.4 Sparsity

When the dimension of the covariate space is high, the slow convergence of minimax estimators can be problematic. One way to improve on this is to exploit lower dimensional structure in the regression function f. The primary way to do this has been through the notion of

(covariate) sparsity. We say that a function $f : \mathcal{X} \to \mathbb{R}$ (for a d-dimensional subset $\mathcal{X} \subseteq \mathbb{R}^d$) is s-sparse if there is a set of indices $\mathcal{J} \subseteq \{1, \dots d\}$ of size s such that $f(x) = f(x + ae_j)$ for all $x \in \mathcal{X}$, for all $a \in \mathbb{R}$ with $x + ae_j \in \mathcal{X}$, for all $j \notin \mathcal{J}$.

Any oracle that knows the set $\mathcal{J}$ would be expected to perform with a rate that depends on the dimension s rather than d. This can be achieved by projecting the covariate vectors X_i to the lower dimensional space of $\mathcal{X} \cap \text{Span}(e_j : j \in \mathcal{J})$ and doing the regression of Y_i against these projections. The performance of this estimator would thus depend only on the dimension of this smaller space.

When the set of important variables is unknown we might hope that estimating the set $\mathcal{J}$ can be done sufficiently quickly to recoup this performance benefit. One estimator that succeeds in this is the RODEO [Lafferty and Wasserman, 2008], which replaces the bandwidth parameter in a local polynomial estimator f_n with a bandwidth vector $h = (h_1, \dots, h_d)$, so the local ball would be a local ellipsoid with radii determined by these bandwidths. The key idea is to shrink the bandwidth in the directions that correspond to important directions and leave bandwidth in the other directions large. This is achieved by a careful consideration of the derivatives of $f_n(x; h)$ with respect to each bandwidth component—unimportant directions might be expected to have $\partial f_n(x; h)/\partial h_j$ smaller in magnitude than for the directions that are important. In the RODEO, directions that have this derivative larger than some well chosen threshold are excluded from the estimate of $\mathcal{J}$. It is shown that using the projection determined by this threshold almost achieves the same rate as the oracle (i.e., the exponent differs by a small negative ϵ).

Other notions of sparsity exist: for example, the function f may be sparse in a Fourier basis, where it may be equal to a finite sum $\sum_{j \in \mathcal{J}} \theta_j \phi_j$ as would be encouraged by a Lasso type penalty of on a nonparametric least squares estimator. To differentiate these ideas, we often refer to the notion above as covariate sparsity instead of just sparsity.

2.1.5 Additive Models

Another classical approach to improving on the curse of dimensionality is to assume that the regression function f has an additive structure:

$$f(x_1, \dots, x_d) = \sum_{j=1}^{d} f_j(x_j) \tag{2.42}$$

where each term in the sum is a function $f_j : \mathbb{R} \to \mathbb{R}$ in a smoothness class, for example $f_j \in \mathcal{F}_{\text{Hölder}}(L, \beta)$. If we knew the other functions f_j, then it would be easier to learn with a rate of L^2 risk decay of $n^{-2\beta/(2\beta+1)}$ rather than $n^{-2\beta/(2\beta+d)}$ for the overall function. Of course

in practice we do not know the other functions, and so we would implement an iterative approach that fits each f_i in sequence, removing the estimates from the observations at each step until we reach some convergence criterion. This approach is very attractive for statistical reasons: the class of Hölder additive functions $\mathcal{F}_{\text{Additive}} \subseteq \mathcal{F}_{\text{Hölder}}(L, \beta)$ is very broad in the sense that most functions in the general class can be well approximated by additive functions, and the convergence to this best additive approximation is fast [Stone, 1985]. However, this approximation error will introduce bias for functions that are not additive. Estimation of additive functions can be done through backfitting [Breiman and Friedman, 1985] or after projection pursuit [Friedman and Stuetzle, 1981].

2.1.6 Other Dimension Reduction Techniques

As a final note, "dimension reduction" also sometimes refers to the problem of reducing $\mathcal{X}$ to a submanifold on which μ is concentrated [Bickel and Li, 2007]. This problem bears a resemblance to the problems we will tackle later in that all seek to find an optimal representation of the covariates in a simpler space, but we will not require that the sampling measure μ is concentrated. Instead, we will seek to find representations that reduce dimension based on the structure of the regression function rather than covariates. This distinction is paralleled by the difference between doing Principal Component Analysis (PCA) then linear regression on the principal components compared to Partial Least Squares (PLS) [Wold, 1966]. In the first case, only the covariate distribution is considered in the dimension reduction, and it is argued in Abdi [2003] that this causes problems if dimensions in which f varies highly are ignored.

2.2 Manifolds and Differential Geometry

In statistics, we typically deal with points in linear spaces: real numbers, vectors in $\mathbb{R}^d$, or functions in L^p spaces. Sometimes the objects we wish to describe live in non-linear sets, for example the set of probability density functions on $\mathbb{R}$ cannot be added because their sum would integrate to 2, and we certainly cannot scale these densities by any real number other than one. Similarly the space of compass bearings, $[0°, 360°)$, cannot simply be added because we can go out of this interval. Another problem is that we would also like $\epsilon°$ to be "close" to $(360 - \epsilon)°$ for small $\epsilon > 0$ because these directions would be close on the compass, but the Euclidean metric on the interval would mark these as very different.

Instead, these more complicated spaces require us to consider the field of differential geometry. The common piece of these spaces we will describe is that look *locally* (i.e., on a small enough neighbourhood of each point) like a vector space but can globally take on much more complicated structures. We call such spaces manifolds, which we will briefly define and

explore some properties of. This field has numerous applications in statistics, particularly in directional statistics [Mardia and Jupp, 2009], object oriented data analysis [Marron and Dryden, 2021], shape analysis [Huckemann et al., 2010], and many more.

A d**-dimensional smooth manifold** is a Hausdorff and second countable topological space that is covered by a collection of open sets U_a with homeomorphisms $\phi_a : U_a \to \mathbb{R}^d$ such that the transition maps $\phi_b \circ \phi_a^{-1}$ are infinitely differentiable for all a, b. Examples are $\mathbb{R}^d$ itself, the open subset $(0,1)^d$, or the sphere $S^d = \{x \in \mathbb{R}^{d+1} : \|x\|_2 = 1\}$, well known in directional statistics [Mardia and Jupp, 2009]. The first two examples only require one open set (themselves) and homeomorphism (the identity map). The sphere requires two open sets (for which there are many choices) and so in this we must ensure that the transition maps for the chosen sets and maps are indeed smooth. Some topological spaces cannot be given a manifold structure: for example $\{(a,0) : a \in \mathbb{R}\} \cup \{(0,b) : b \in \mathbb{R}_{\geq 0}\}$, which has a singularity at the point $(0,0)$ that cannot be mapped smoothly to an open subset of $\mathbb{R}^d$. We will also sometimes consider manifolds with boundaries, such as the (closed) unit hypercube $[0,1]^d$, though only the interior is a smooth manifold. When referring to such manifolds with boundary as smooth manifolds, we will typically mean their interior unless otherwise stated.

At each point x in a smooth manifold $\mathcal{M}$ we can attach a copy of $\mathbb{R}^d$ called the **tangent space**, written $T_x\mathcal{M}$, where each element corresponds to an equivalence class of smooth curves $\gamma : [-1,1] \to \mathcal{M}$ with $\gamma(0) = x$ and with equal derivatives at 0. Each curve γ can be given a length by $L(\gamma) = \int_{-1}^{1} \|\gamma'(t)\|_2 \, \mathrm{d}t$, where the norms are taken in the tangent space at $\gamma(t)$. This in turn gives a metric space structure on $\mathcal{M}$, where $d_{\mathcal{M}}(x,y) = \inf_\gamma L(\gamma)$ where the infimum is taken over all piecewise smooth curves γ with $\gamma(-1) = x \in \mathcal{M}$ and $\gamma(1) = y \in \mathcal{M}$.

We call a subset $\mathcal{M}' \subseteq \mathcal{M}$ an **immersed submanifold** if the inclusion map $\phi : \mathcal{M}' \to \mathcal{M} : x \mapsto x$ has an everywhere injective derivative. An immersed submanifold is called **embedded** if ϕ is also a homeomorphism (i.e., it preserves topology). Every smooth manifold $\mathcal{M}$ can be isometrically embedded in $\mathbb{R}^k$ for some $k > d$, a result known as the Nash Embedding Theorem [Nash, 1956]. This means that curves through $\mathcal{M}$ have the same length as they would when integrated in the embedding space, and so the distance between points in the manifold is bounded from below by the distance in the embedding Euclidean space. When we embed a manifold $\mathcal{M}$ into $\mathbb{R}^k$ in this way we also have a natural embedding of the tangent spaces $T_x\mathcal{M}$ each as an affine hyperplane through the point x. We can then also define the normal space of $\mathcal{M}$ at x as the orthogonal complement of this affine space at the point x.

2.3 Symmetries and Group Theory

We now cover the mathematical background of symmetries. These are best described using the language of *group theory*, a field that was born out of the study of the symmetries of polynomial equations [Lang, 2012]. It has since found applications broadly across mathematics, including in number theory, differential equations, particle physics, and within statistics in the examples in the next section [Kleiner, 1986]. This field abstracts the notion of the sets of transformations Ψ from the introduction to study the properties of these sets because they often have similar structure. This abstraction will be important because it opens up some of the techniques we will exploit in our methods later.

A **group** is a set G with an associative binary operation (or **multiplication**) $(a, b) \mapsto ab \in G$ such that: there is an identity element $e \in G$ with $eg = ge = g$ for all $g \in G$; and there are inverses g^{-1} with $g^{-1}g = gg^{-1} = e$ for all $g \in G$. We call a group with one element, $\{e\}$, the trivial group and denote[2] it by I.

Example 1. *Consider the set of matrices that rotate vectors in* $\mathbb{R}^2$ *around the origin:*

$$SO(2) = \left\{R_\theta : \theta \in [0, 2\pi)\right\} \qquad \text{where} \quad R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{2.43}$$

and the usual matrix-multiplication product as the group multiplication. We know that this set has an identity element, given by the rotation matrix R_0 *which is the identity matrix on* $\mathbb{R}^2$*, and thus has* $R_\theta R_0 = R_0 R_\theta = R_\theta$ *for all* $R_\theta \in SO(2)$*. We also know that every rotation* R_θ *has an inverse given by* $R_{2\pi-\theta} \in SO(2)$*, because the composition of these rotations will leave any vector unchanged. Therefore, this set with matrix multiplication does indeed form a group.*

Example 2. *Consider the subset of* $SO(2)$ *given by rotations of angles with magnitude at most* $\pi/2$*. This set does* ***not*** *form a group under matrix multiplication because the product of* $R_{\pi/2}R_{\pi/2}$ *is not in this set. We could however choose an alternative multiplication, for example* $(R_\theta, R_\phi) \mapsto R_\psi$ *where* $\psi = (\theta+\psi)/2 \mod 2\pi$*. Now the set is closed under compositions because the angle of any composition will have magnitude at most* $\pi/2$*. Moreover, it clearly contains the identity and all inverses and so does indeed form a group. This illustrates that the choice of multiplication matters.*

If A is a subset of a group G, then we say that it is a **subgroup** of G if it is closed under the multiplication of G; i.e. $ab \in A$ for all $a, b \in A$. If A is any subset of G, then the intersection of all subgroups of G that contains A is called the group **generated by** A. We write $\langle A \rangle$ for

[2]Note that we also sometimes use I to mean the identity matrix, but usually mark these with subscripts of the dimension of the matrix.

this group, or $\langle a_1, a_2, \dots \rangle$ for elements $a_i \in G$, or $\langle A_1, A_2, \dots \rangle$ for the group generated by the union of subsets A_i of G.

A group G acts on a domain $\mathcal{X}$ via a map $\cdot : G \times \mathcal{X} \to \mathcal{X}$ with the properties that $e \cdot x = x$ and $g \cdot (h \cdot x) = (gh) \cdot x$ for all $g, h \in G$ and $x \in \mathcal{X}$. This means that the maps $g\cdot : \mathcal{X} \to \mathcal{X}$ are the bijections ψ from the introduction. We will usually assume that all group actions are **faithful**, which means that if $g \cdot x = x$ for all $x \in \mathcal{X}$ then $g = e$. A group action partitions the domain $\mathcal{X}$ into **orbits** of the form $[x]_G = \{g \cdot x : g \in G\}$. We call the set of orbits the **quotient space** of the action and write it as $\mathcal{X}/G$.

Example 3. *Consider the group $SO(2)$ of 2D-rotations from Example 1, and consider the domain $\mathbb{R}^2$. This group has a very natural action on these two vectors—we can just multiply them by the group elements. This would just mean $R_\theta \cdot x = R_\theta x$, which clearly has all of the desired properties of a group action. In fact, matrix multiplication forms the prototypical example of a group action. This action is faithful, because if $R_\theta x = x$ for $x = (1, 0)^T$, then $R_\theta = I_2$. The orbits of this action are circles centered at the origin, because we can rotate any vectors of the same length onto each other. Therefore, the quotient space of this action can be described by the ray $\{(a, 0)^T : a \in [0, \infty)\}$ because each circle can be uniquely defined by its intersection with this ray.*

Example 4. *Again consider the group $SO(2)$ but now let the domain $\mathcal{X}$ be $\mathbb{R}^3$. We can let this group act on this domain in many different ways. A classical choice is to pick a unit vector $u \in \mathbb{R}^3$ and then set:*

$$R_\theta \cdot x = (\cos\theta)x + (\sin\theta)(u \times x) + (1 - \cos\theta)(u \cdot x)u \tag{2.44}$$

This is known as Rodrigues' rotation formula [Rodrigues, 1840]. In this case, the group action rotates x around the chosen axis u by the angle θ. In this case the orbits are still circles but now with centres along the line through u, i.e., $[x]_{SO(2)}$ is the circle of radius $\sqrt{\|x\|_2^2 - \langle x, u\rangle^2}$ centred at $\langle x, u\rangle u$. Thus, the quotient space is identified with a set of rays orthogonal to u with roots along u.

Example 5. *Lastly, again consider the group $SO(2)$ but now let the action be given by $R_\theta \cdot x = x \in \mathbb{R}^2$ for all $R_\theta \in SO(2)$. In this case, the action is called trivial and is clearly not faithful.*

If the group G has a topology that agrees with the group multiplication, in the sense that $(g, h) \mapsto g^{-1}h$ is continuous, then we say that G is a **topological group**. When we use topological adjectives (e.g. compact, Hausdorff, etc.) for a group, we assume that the group and topology agree. We consider only **locally compact** groups in this paper, which means that every element $g \in G$ has a compact neighbourhood $g \in U \subseteq N$ for some open U.

Example 6. *Consider two topologies on the group $SO(2)$. The first we take as the topology induced by the Frobenius metric on the matrices:*

$$d_{\text{Frob}}(R_\theta, R_\phi) = 2(\cos\theta - \cos\phi)^2 + 2(\sin\theta - \sin\phi)^2 \tag{2.45}$$

In this topology it is straightforward to check that the map $R_\theta, R_\phi \mapsto R_\theta^{-1} R_\phi$ is indeed continuous and so forms a topological group. This topological group is compact, Hausdorff, and second countable.

The second topology to consider is very coarse, just generated by the single proper non trivial open set $\{R_0\}$. In this case, the inverse image of this open set under $(g, h) \mapsto g^{-1}h$ is the diagonal set $\{(R_\theta, R_\theta) : R_\theta \in SO(2)\}$, but this is not open in the product topology. Thus, this map is not continuous and so $SO(2)$ with this topology is not a topological group.

A stronger condition is that G (as a set) also admits a differential manifold structure, which must also agree in that the map above is smooth. We call such a group a **Lie group**. When we use the notion of dimension of a group we are referring to a Lie group G. As a manifold, it must necessarily be Hausdorff and second countable, and admit a metric (in the sense of distance) $d_G : G \times G \to [0, \infty)$.

Example 7. *As shown in the previous section the set $S^1 = \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$ is a smooth manifold. If we identify rotations R_θ with points $(\cos\theta, \sin\theta) \in S^1$, we can give the group $SO(2)$ the same smooth structure as the circle S^1. With this manifold structure one can check that the map $(g, h) \mapsto g^{-1}h$ is smooth and so $SO(2)$ forms a Lie group. From now when we consider this group we will consider it as a Lie group with this manifold structure.*

If G is a locally compact group, then there exists a Borel measure Γ on G such that $\Gamma(A) = \Gamma(\{ga : a \in A\})$ for all measurable A and all $g \in G$. Moreover, this measure is unique up to scaling, and we call such Γ a **(left) Haar measure** on G [Haar, 1933]. When G is compact, we can normalise any left Haar measure to give a uniform distribution on the group, $U(G)$.

Example 8. *In the case of $SO(2)$ as a Lie group, we can construct a left Haar measure Γ by taking:*

$$\Gamma(A) = \lambda(\{\theta : R_\theta \in A\}) \tag{2.46}$$

where λ is the Lebesgue measure on $[0, 2\pi)$, for any open set $A \subseteq SO(2)$. When normalised by $1/2\pi$, this is precisely the uniform distribution on the unit circle. It is left shift invariant, because

$$\Gamma(\{R_\theta R_\phi : R_\phi \in A\}) = \lambda(\{\theta + \phi \mod 2\pi : R_\phi \in A\}) = \lambda(\{\phi \mod 2\pi : R_\phi \in A\}) \tag{2.47}$$

This is a consequence of the shift invariance of the Lebesgue measure.

If a topological group acts on $\mathcal{X}$, we will usually assume that the action is continuous (in the product topology on $G \times \mathcal{X}$), and when a Lie group acts we will usually assume that the action is smooth. In the case of a locally compact Lie group acting smoothly and faithfully on $\mathcal{X}$, the orbits $[x]_G$ are themselves immersed sub-manifolds of $\mathcal{X}$. If the action is also **proper**, i.e., inverse images of compact sets under the map $(g, m) \mapsto (m, g \cdot m)$ are compact, then the orbits are in fact smoothly embedded submanifolds [Lee, 2013].

Example 9. *In the example of $SO(2)$ acting on $\mathbb{R}^2$ by matrix multiplication, we have a smooth proper faithful action of a compact (and thus locally compact) group and so must have smoothly embedded orbits. Indeed, these are each circles centred at the origin in $\mathbb{R}^d$ which are indeed smooth embedded submanifolds of $\mathbb{R}^2$.*

The Principal Orbit Type Theorem [Dieck, 1987] says that there is a "generic" topology for orbits of a smooth proper action of a compact group G, i.e., that there is an open dense subset $U \subseteq \mathcal{X}$ with orbits $[x]_G$ that are pairwise diffeomorphic for all $x \in U$, and such x are called **regular points** of the action. In particular, the dimension of these principal orbits is the reduction we can achieve by symmetrising the estimator, which we denote d^G with a superscript to distinguish from a metric on the group with a subscript d_G. For non-compact groups we may lose this guarantee of almost sure regularity, but because we will only consider proper actions there is still a maximal dimension orbit type (as all orbits are embedded submanifolds) and we call all points $x \in \mathcal{X}$ with this maximal dimensional orbit a **regular point** of the action too. All non-compact group actions considered (usually just translations of linear spaces) will have their set of regular points form a dense subset of $\mathcal{X}$ so we will refer to these orbits as principal type orbits too.

Example 10. *Continuing the same example of $SO(2)$ acting naturally on $\mathbb{R}^2$, we have a smooth proper action of a compact group. Thus, this theorem says that almost all of the orbits will have the same dimension; in this case it is the dimension of the circles that most of the orbits take. The singular exception to this is the circle of radius 0, which is a 0 dimensional submanifold instead.*

When the base space $\mathcal{X}$ has a metric $d_{\mathcal{X}}$ it descends to the quotient space of a group action if the action is proper. This has a simple form:

$$d_{\mathcal{X}/G}([x]_G, [y]_G) = \inf_{g,h \in G} d_{\mathcal{X}}(g \cdot x, h \cdot y) \tag{2.48}$$

When the action is not proper, this infimum can be 0 for all points x, y even if they are in distinct orbits so this no longer forms a strict metric. As an example, consider the action of $\mathbb{R}_{>0}$ by scaling on $\mathbb{R}^d$: all orbits meet at the origin and so the right hand side above is 0. In these cases it is often possible to construct alternative metrics on the quotient space, in this

case an option would be:

$$d_{\mathbb{R}^d/\mathbb{R}_{>0}}([x]_{\mathbb{R}_{>0}}, [y]_{\mathbb{R}_{>0}}) = \arccos\left(\frac{\langle x,y\rangle}{\|x\|_2\|y\|_2}\right) \tag{2.49}$$

This comes from picking a unique representative of each orbit (in this case on the unit sphere) and using a distance between these.

2.3.1 Important Examples of Groups

We now consider several other important groups that will be used in the examples in this thesis. These fall into three main categories: discrete symmetries that we often describe as permutations; continuous symmetries that are described by matrices; and a few other exotic examples.

Discrete Symmetries

One of the most common groups encountered are *permutation groups*, defined as follows. Let A be any set and consider the set of bijections $\phi : X \to X$ which we denote $\mathrm{Sym}(A)$. This set of bijections, with the multiplication of function composition, form a group because: it contains the identity $e : x \mapsto x$ as this is bijective; the composition of two bijections is bijective; and the inverse of any bijection is bijective. When A is finite these permutations groups are given the discrete topology and have order $|A|!$. Permutation groups are interesting because of their subgroups; Cayley's theorem says that every group can be expressed as a subgroup of $\mathrm{Sym}(A)$ for some set A. Usually we are interested in subgroups that preserve specific structures.

One example of this is the **dihedral group** of the square embedded in $\mathbb{R}^2$, given by $D_4 = \{I, R_{\pi/2}, R_\pi, R_{3\pi/2}, R_h, R_v, R_/, R_\backslash\}$ where:

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad R_{\pi/2} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad R_\pi = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \quad R_{3\pi/2} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$
$$R_h = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \quad R_v = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad R_/ = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad R_\backslash = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

This group of matrices has the property that if $A = \{a_1, a_2, a_3, a_4\} \subseteq \mathbb{R}^2$ form the vertices of a square centred at the origin then $g \cdot A = \{g \cdot a_i : a_i \in A\} = A$, thus preserving the square. This group can be expressed as a subgroup of the permutation group $\mathrm{Sym}(A)$ by considering the reordering of the vertices after the reflections and rotations above, e.g., R_h swaps the position of a_1 with a_2 and of a_3 with a_4. Dihedral groups of other regular polygons are described

very similarly, as are the symmetry groups of the five platonic solids (the only regular 3D polyhedra).

Continuous Symmetries

This thesis will later focus primarily on continuous groups because these offer orbits with positive dimension, which leads to improvements on the curse of dimensionality. Most of these groups will be described as subgroups of the **general linear group** $GL(\mathcal{X}) = \{T : T \text{ is an invertible linear operator } \mathcal{X} \to \mathcal{X}\}$ when $\mathcal{X}$ is a linear space such as $\mathbb{R}^d$. These groups all act by applying the transformation to the points unless specified otherwise. When $\mathcal{X}$ is a d-dimensional vector space over $\mathbb{R}$ we sometimes write $GL(d)$, and identify the linear operators in $GL(d)$ with their matrix representations $A = (Te_1, Te_2, \ldots, Te_d)$ in some chosen basis $\{e_j\}_{j=1}^d$.

The **special linear group** written $SL(d)$ is the subgroup of $GL(d)$ consisting of all A with $\det(A) = 1$. The **orthogonal group** written $O(d)$ consists of all A such that $A^T A = I$. The **special orthogonal group** written $SO(d)$ consists of the intersection of the special linear and orthogonal groups on $\mathcal{X}$ and is the prototypical example of a compact group in this paper. It has appeared in statistics in nearly all of the cited work so far [Mardia and Jupp, 2009, Kendall, 1984]. The group $SO(2)$ is homeomorphic (i.e., topologically equivalent) to the complex unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$, and in fact this circle forms a group under complex multiplication that is also algebraically equivalent[3] to $SO(2)$. We frequently use these notations interchangeably. There are infinitely many subgroups of $SO(3)$ that are equivalent to S^1, corresponding to rotations of the hyperplane orthogonal to a unit vector u. To distinguish these as subgroups of $SO(3)$ we write S^1_u.

The group of **affine transformations** of $\mathbb{R}^d$ consists of all invertible transformations for which the image of a line is also a line. This group contains all inveritble linear transformations, and so is a supergroup of $GL(d)$ that also contains translations $x \mapsto x + v$. Every affine transformation can be represented by a matrix $A \in GL(d)$ and a translation $v \in \mathbb{R}^d$ and takes $x \mapsto Ax + v$. The affine group is not a proper action because of the scalings—the inverse image of a neighbourhood of 0 is the whole space. It does have a subgroup, the **Euclidean Group**, consisting of all rotations, translations, and reflections of $\mathbb{R}^d$ (i.e., all isometries of this space) which does have a proper action. Within this there is also the **Special Euclidean Group** of translations and rotations, i.e., the group of all rigid motions of Euclidean space.

[3] i.e., isomorphic, which means that there is a bijection $\psi : SO(\mathbb{R}^2) \to S^1$ with $\psi(R_1 R_2) = \psi(R_1)\phi(R_2)$ for all $R_1, R_2 \in SO(2)$.

Other Groups

There are many other groups of interest in particular situations. Importantly, it is sometimes worth changing the multiplication from the standard one to a more exotic one to allow us to access other properties of groups for their study. One example of this is the set of (non-degenerate) covariance matrices, i.e., positive definite (PD) matrices in the set $\mathrm{Sym}_{>0}(d)$. This set have an obvious group multiplication given by matrix multiplication: any product of PD matrices is also PD and so is the matrix inverse of any PD matrix. However, this is not the only possible multiplication. Consider the multiplication $\oplus$ defined by:

$$A \oplus B = \exp_{I_d}(\log_{I_d} A + \log_{I_d} B) \tag{2.50}$$

where $\exp_{I_d}$ and $\log_{I_d}$ denote the matrix exponential map $\exp_I(X) = I + \sum_{k=1}^{\infty} X^k/k!$ and the logarithm map (given by the inverse of $\exp_I$), each based at the identity matrix $I_d \in \mathrm{Sym}_{>0}(d)$. It is straightforward to check that this operation is associative from the properties of these maps, and we are guaranteed that $A \oplus B \in \mathrm{Sym}_{>0}(d)$ because this contains the range of $\exp_{I_d}$. For inverses, note that

$$\begin{aligned} A \oplus A^{-1} &= \exp_{I_d}(\log_{I_d} A + \log_{I_d} A^{-1}) && (2.51)\\ &= \exp_{I_d}(\log_{I_d} A - \log_{I_d} A) && (2.52)\\ &= I_d && (2.53) \end{aligned}$$

where A^{-1} is the usual inverse in the matrix sense. Thus, $\mathrm{Sym}_{>0}(d)$ forms a group with the multiplication $\oplus$, which has the additional property of being commutative, $A \oplus B = B \oplus A$, unlike the usual matrix multiplication. This is useful because it allows us to explore additive models for a covariance matrix Σ given covariates $X_1, \ldots, X_d \in \mathbb{R}$:

$$\Sigma = M \oplus m_1(X_1) \oplus \cdots \oplus m_d(X_d) \oplus \xi_x \tag{2.54}$$

where M is a mean covariance matrix (in the Fréchet sense), each function $m_i : \mathbb{R} \to \mathrm{Sym}_{>0}(d)$, and $\xi_x \overset{iid}{\sim} \nu$ is a noise term taking values in $\mathrm{Sym}_{>0}(d)$ too. Estimation of these functions can then be done analogously to the case of additive models in regression (see Section 2.1.5). This is precisely the approach of Lin et al. [2023] where this is applied to a problem in diffusion tensor imaging.

2.4 Invariant Functions

We now describe the structure of the functions we will exploit for faster regression. This comes from an invariance to a group acting on the covariant space. We will start by describing the space of such functions here and give important tools for analysis later.

Let G be any group acting on $\mathcal{X}$. We say that a function $f \in L^2(\mathcal{X})$ is **G-invariant** if there is a function f_0 with $\|f_0 - f\|_2 = 0$ and:

$$f_0(x) = f_0(g \cdot x) \tag{2.55}$$

for all $g \in G$ and $x \in \mathcal{X}$. If $\mathcal{F}$ is a set of functions in $L^2(\mathcal{X})$, we let $\mathcal{F}_G$ be the subset of G-invariant functions in $\mathcal{F}$. For any group G the space $L^2_G(\mathcal{X})$ is a linear subspace of $L^2(\mathcal{X})$. This subspace is also closed in the topology of $L^2(\mathcal{X})$. It also forms an algebra as the pointwise product of any two G-invariant functions must also be G-invariant.

If G is acting continuously on $\mathcal{X}$, then we also have the following equivalence between G- and $\overline{G}$-invariant functions (with the bar indicating the topological closure of G).

Lemma 11. *Suppose that $\mathcal{G}$ is a locally compact group acting continuously on $\mathcal{X}$. Any continuous $f \in L^2(\mathcal{X})$ is G-invariant if and only if f is $\overline{G}$-invariant, for all $G \leq \mathcal{G}$.*

Proof of Lemma 11. Since $G \subseteq \overline{G}$ the forwards direction is clear. Suppose that f is either weakly or strongly G-invariant. Take $g \in \overline{G}$ and a net $g_\bullet$ in G converging to g, and note that

$$f(g \cdot X) = f((\lim g_\bullet) \cdot X) = f(\lim g_\bullet \cdot X) = \lim f(g_\bullet \cdot X) = \lim f(X) = f(X) \tag{2.56}$$

almost surely, giving the result. □

Let $\mathcal{G}$ be any group acting on $\mathcal{X}$ and f be any regression function. We say that G is a **maximal invariant subgroup** of $\mathcal{G}$ if f is G_0-invariant, and every subgroup of $\mathcal{G}$ that f is invariant to is a subgroup of G_0. We will show that pair of f and $\mathcal{G}$ determines a unique maximal invariant subgroup. This will be done using the following rules for invariance the proofs are found in the proofs section of Chapter 5.

Proposition 12 (Rules for Invariance)**.** *Suppose that $\mathcal{G}$ is a group acting measurably on $\mathcal{X}$, and let $f : \mathcal{X} \to \mathbb{R}$ be a regression function. Suppose that G, G' and H are subgroups of $\mathcal{G}$. Then:*

1. *If f is G-invariant, then f is H-invariant for any $H \leq G$;*
2. *If f is not G-invariant , then f is not H-invariant for any $H \geq G$;*
3. *If f is both G- and H-invariant, then f is $\langle G, H\rangle$-invariant; and*
4. *If f is G-invariant but not H-invariant, then f is not G' invariant for any $G' \leq \mathcal{G}$ where $\langle G, H\rangle = \langle G, G'\rangle$.*

Moreover, if $\mathcal{G}$ is a closed group acting continuously and f is continuous then the groups $\langle G, H\rangle$ and $\langle G, G'\rangle$ can be replaced by their topological closures above.

These rules will be used extensively in Chapter 5. They all essentially follow from the definition of invariance. Using these rules we can prove the following result that guarantees the existence and uniqueness of a maximal invariant subgroup, and we can utilise them in the estimation algorithm in Section 5.3.

Proposition 13. *For any group $\mathcal{G}$ acting on $\mathcal{X}$, every function $f : \mathcal{X} \to \mathbb{R}$ has a unique maximal invariant subgroup $G_{\max}(f, \mathcal{G})$. If f is continuous and $\mathcal{G}$ acts continuously on $\mathcal{X}$ then $G_{\max}(f, \mathcal{G})$ is closed.*

The proof is to essentially consider the closure of the group generated by all of the subgroups that f is invariant to (i.e., $\overline{\langle G_i \leq \mathcal{G} : f \text{ is } G\text{-invariant} \rangle}$.

2.4.1 Compact Symmetries

When G is compact and μ is a G-invariant probability measure, $L^2(\mathcal{X})$ admits an orthogonal decomposition:

$$L^2(\mathcal{X}) = L^2_G(\mathcal{X}) \oplus A_G(\mathcal{X}) \tag{2.57}$$

See Elesedy and Zaidi [2021] for a proof. We term the functions in $A_G(\mathcal{X})$ **G-anti-symmetric**. There also exists an orthogonal projection onto $L^2_G(\mathcal{X})$, denoted $S_G : L^2(\mathcal{X}) \to L^2_G(\mathcal{X})$. This projection has an explicit form:

$$(S_G f)(x) = \mathbb{E}\big(f(g \cdot x)\big) \tag{2.58}$$

where $g \sim U(G)$. The orthogonal decomposition in equation (2.57) can therefore be expressed as:

$$f = S_G f + f^{\perp}_G, \qquad \text{where } f^{\perp}_G = f - S_G f \in A_G(\mathcal{X}). \tag{2.59}$$

Note that by definition $S_G f = 0$ for any $f \in A_G(\mathcal{X})$. When μ is not G-invariant, this operator is no longer an orthogonal projection, but it is still idempotent and its image is still $L^2_G(\mathcal{X})$.

Example 14. *Consider the case of $G = \{z \in \mathbb{C} : |z| = 1\}$ acting by rotations on $\mathbb{R}^2$ with μ a standard Gaussian measure. The function f on this disc given by:*

$$f(x) = \sin(\|x\|_2)\cos(2\arctan(x_2/x_1))^2 = \sin(r)\cos(2\theta)^2 \tag{2.60}$$

in Euclidean and polar coordinates respectively, has an orthogonal decomposition shown in Figure 2.1.

2.4.2 Non-Compact Symmetries

When G is not compact, we can no longer use the operator in equation (2.58) because a Haar measure may not be normalisable (as with the Lebesgue measure on $\mathbb{R}^d$). We can

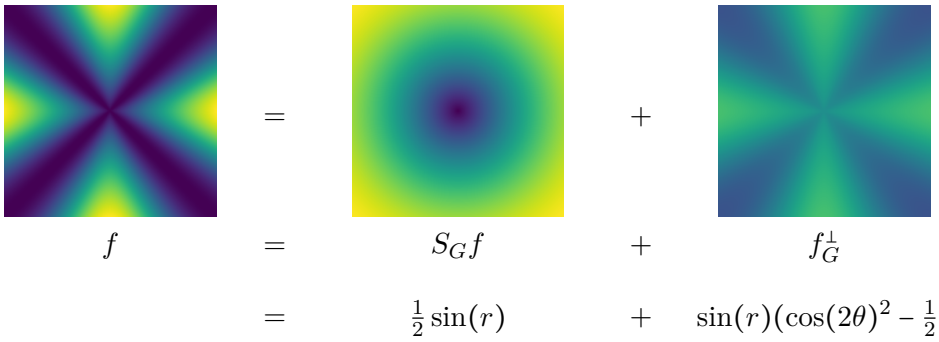


Fig. 2.1 The orthogonal decomposition of $f = \sin(r)\cos(2\theta)^2$ into a G symmetric piece and a G-anti-symmetric piece, as described in Example 14.

however pick any other distribution ρ on G, and generalise to the **Partial Symmetrisation Operator** $S_\rho : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ given by

$$S_\rho f(x) = \mathbb{E}(f(g \cdot x)) \tag{2.61}$$

where $g \sim \rho$. This operator is not a projection, because it is not idempotent, but it is still a contraction in the L^2 norm when the integrating measure μ is G-invariant[4], and in the L^∞ norm for all measures μ. It is also trivially the identity on $L^2_G(\mathcal{X})$.

2.5 Examples of Symmetry in Statistics

Symmetry has a long history in statistics, though sometimes implicitly. The obvious examples are examining the symmetry of distributions, as this is sometimes a feasible way to test normality hypothesis, and in shape statistics. In this section, we now present a number of examples both of these explicit forms and of the implicit occurrences too. Some show that traditional problems can be expressed in the language of symmetries, and some show that symmetries can be exploited for some statistical or computational gain. We also describe the current efforts to test for symmetries and to estimate symmetries of objects, including regression functions.

2.5.1 Symmetry in Distributions

The main examination of symmetry in statistics is as a property of distributions. One original focus of this idea was to distinguish Gaussians from non-Gaussians. If we consider only the first two moments of a real valued random variable X, then we can always fit a Gaussian with the same moments. If we also consider the standardised third moment $\mathbb{E}((X - \mathbb{E}(X))^3)/(\mathbb{E}((X - \mathbb{E}(X))^2)^{3/2}$ (called the **skew**), then we can distinguish a Gaussian

[4]This is usually impossible for finite measures with non-compact symmetries but in other contexts it is possible, as with the Lebesgue measure on $\mathbb{R}^d$ being translation invariant.

X from a non-Gaussian. Any Gaussian must have no skew because of the symmetry around its mean, and so the asymmetry quantified by the skew can be used to construct tests for normality ($\sqrt{b_1}$ test of Shapiro et al. [1968]). There are also myriad ways in which we examine symmetry of distributions.

A prime example of this is in time series, where we often consider the property of *stationarity*. Let $\{X_t\}_{t\in\mathbb{T}}$ be a stochastic process indexed by times $t \in \mathbb{T}$ which is either $\mathbb{R}$ or $\mathbb{Z}$. We say that X is **strictly stationary** if:

$$\left(X_{t_1}, X_{t_2}, \ldots, X_{t_n}\right) \overset{D}{=} \left(X_{t_1+a}, X_{t_2+a}, \ldots, X_{t_n+a}\right) \tag{2.62}$$

for all times $a, t_1, \ldots, t_n \in \mathbb{T}$ and all $n \in \mathbb{Z}_{>0}$. This is precisely a statement about translation invariance on the time domain. Let G be the group $\mathbb{T}$ under addition and let $\nu_{t_1,\ldots,t_n}$ be the distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$. We can define an action of G on the stochastic process $\{X_t\}_{t\in\mathbb{T}}$ by:

$$g \cdot \{X_t\}_{t\in\mathbb{T}} = \{X_{t+g}\}_{t\in\mathbb{T}} \tag{2.63}$$

With this group action we have the equivalence: X_t is strictly stationary if and only if the law of every finite coordinates $\{X_t\}_{t\in\mathbb{T}}$ is G-invariant.

This is closely related to the notion of **seasonality** in time series, where the distribution is equal when shifted by integer multiples of a specific period $a \in \mathbb{T}$. In this case, the distributions above are no longer invariant to all translations $\mathbb{T}$ but instead they are invariant to the subgroup $\mathbb{T}_a = \{ak : k \in \mathbb{Z}\}$.

Another example is in spatial distributions where we might examine the distribution of locations of events such as sunspots as in García-Portugués et al. [2020]. In this example, we consider distributions on the unit sphere $S^2 = \{x \in \mathbb{R}^3 : \|x\|_2\}$ and note that the observations of sunspots often occur evenly around longitudinal circles even if unevenly in the north-south direction. Consider the group $SO(2)$ acting on S^2 by matrix multiplication. A natural question is: *"can we identify if the distribution μ on S^2 is invariant to this action?"*. This question has been studied by many authors, particularly Baringhaus [1991] and García-Portugués et al. [2020] who each construct tests based on U-statistics. In the simplest form, if $X_i \overset{iid}{\sim} \mu$ on the sphere S^2, then to test for the rotational symmetry of μ around some axis $\theta \in S^2$ we can decompose each X_i as:

$$X_i = (X_i^T\theta)\theta + (I - \theta\theta^T)X_i \tag{2.64}$$

the second component measures the distribution of X_i around the axis θ, and if μ is rotational invariant this component (after normalisation) would be uniformly distribution around the hypersphere S^1 that is orthogonal to θ. Thus, a test that detects a deviance from uniformity

of the data $(I-\theta\theta^T)X_i/\|(I-\theta\theta^T)X_i\|_2$ can reject the rotation invariance of μ about the axis θ. Other recent tests have also been proposed, one based on optimal transport [Huang and Sen, 2023] and another based on an energy distance[5] statistic between the original distribution and its symmetrisation [Chiu and Bloem-Reddy, 2023].

2.5.2 Symmetry in Hypothesis Tests

Permutation tests [Good, 2005] are based on a symmetry that exists under a null hypothesis and does not under the alternative. Specifically, suppose that we have collected independent equal sized two datasets $\mathcal{D}_1 = \{X_i\}_{i=1}^n$ and $\mathcal{D}_2 = \{Y_i\}_{i=1}^n$ with $X_i, Y_i \in \mathcal{X}$ and consider the hypotheses $H_0 : \mathcal{D}_1 \overset{D}{=} \mathcal{D}_2$ vs the alternative $H_1 : \mathcal{D}_1 \overset{D}{\neq} \mathcal{D}_2$. Suppose that we have a statistic $T(\mathcal{D}_1, \mathcal{D}_2)$ that converges to a finite constant E in probability under the null distribution and diverges to ∞ in probability under the alternative (e.g. if $X_i, Y_i \in \mathbb{R}$, the energy distance[5] between $\mathcal{D}_1$ and $\mathcal{D}_2$).

Under the null hypothesis, we can interchange the datapoints between the distributions without affecting the distribution, i.e. the permuted datasets $\mathcal{D}_1^{(j)} = \{Z_i^{(j)}\}_{i=1}^n$ where $Z_i^{(j)}$ is X_i with probability 1/2 and Y_i otherwise and $\mathcal{D}_2^{(j)} = \{X_i + Y_i - Z_i^{(j)}\}_{i=1}^n$ will have the same distributions as $\mathcal{D}_1$ and $\mathcal{D}_2$. Therefore, we can estimate the p-value of the null hypothesis by:

$$p = \tfrac{1}{B}\left|\left\{j \in \{1, \dots, B\} : T(\mathcal{D}_1, \mathcal{D}_2) \leq T(\mathcal{D}_1^{(j)}, \mathcal{D}_2^{(j)})\right\}\right| \tag{2.65}$$

for some choice of permutations $B \geq 100$. The permuted datasets will have $T(\mathcal{D}_1^{(j)}, \mathcal{D}_2^{(j)}) \xrightarrow{\mathbb{P}} E$ for each j under both hypotheses because the permuted datasets are each drawn from the same mixture distribution. Under the null hypothesis, these will be roughly evenly spaced around $T(\mathcal{D}_1, \mathcal{D}_2)$ which can be shown to control the size of the test exactly. However, under the alternative hypothesis $T(\mathcal{D}_1, \mathcal{D}_2)$ will be much larger than the $T(\mathcal{D}_1^{(j)}, \mathcal{D}_2^{(j)})$, so p will be small and the test will have power.

This permutation test relies on a permutation symmetry that exists under the null hypothesis only. Let G be the group $\{0,1\}^n$ of binary vectors under addition mod 2 and let G act on the space $\mathcal{X}^n \times \mathcal{X}^n$ as:

$$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \cdot \left(\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \right) = \left(\begin{bmatrix} a_1 X_1 + (1-a_1)Y_1 \\ \vdots \\ a_n X_n + (1-a_n)Y_n \end{bmatrix}, \begin{bmatrix} (1-a_1)X_1 + a_1 Y_1 \\ \vdots \\ (1-a_n)X_n + a_n Y_n \end{bmatrix} \right) \tag{2.66}$$

This precisely describes the permutations between the datasets above, where the group element $(a_1, \dots a_n)$ is sampled uniformly. Now with this group action we can rephrase the

[5] See Székely et al. [2004] for a definition of energy distances.

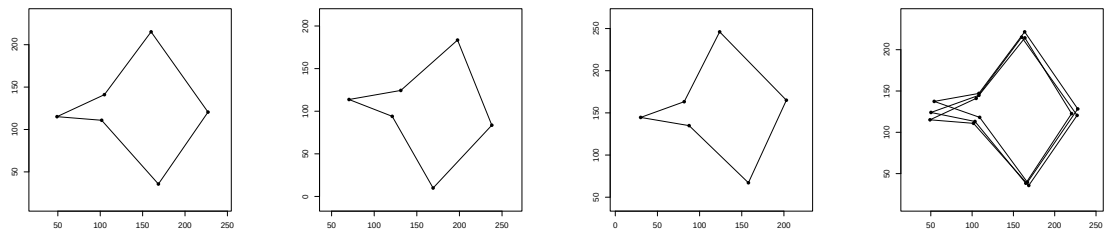


Fig. 2.2 Examples of three mice vertebrae landmarks measured at varying orientations, and an overlay of the three sets of landmarks after rotational alignment. All drawn from the R-package shapes [Dryden and Mardia, 2016] using data from Mardia and Dryden [1989].

null hypothesis in terms of invariance: $\mathcal{D}_1 \overset{D}{=} \mathcal{D}_2$ if and only if the distribution of $(\mathcal{D}_1, \mathcal{D}_2)$ is G-invariant. This means that permutation tests are a test of G-invariance of this joint distribution.

2.5.3 Symmetry in Spaces

Symmetries are often invoked explicitly when we want to describe the *shape* of an object. The idea is that we should separate the information about the object from the way in which we observe it; if we look at the same object from different angles (as with protein samples in Cryo-EM imaging, as well as in many other contexts) we would introduce variance into the observations that do not exist in the samples. To overcome this, we decompose the observation X into its orbit $[X]_\Psi$, which captures the underlying information in the object, and its position $\psi_X \in \Psi$. To make sense of the position, we pick a representative element $\tilde{X} \in [X]_\Psi$ in some deterministic way that depends only on the orbit[6] and then pick $\psi_X \in \Psi$ such that $X = \psi_X(\tilde{X})$. This means that we can decompose the variance of X into the variance of its orbit and variance of its position, which often leads to better estimators and inferential methods as in the following examples.

Example 15 (Shape Spaces)**.** *Consider the problem of identifying the average shape of an 2D-object described by a collection of k landmark points collected in a $2 \times k$ matrix. An example of these landmarks is shown in Figure 2.2 where six landmarks are observed for the T2-vertebrae of three different mice.*

If we merely averaged these objects as matrices, we would often include the information in the orientation, scale, and position of the landmarks rather than identifying the true "shape" of the object. As pioneered by Kendall [1984], we thus want to remove this extra information by considering these objects equivalent if they can be transformed into each other using bijections

[6] I.e., $\tilde{X} = \tilde{Y}$ whenever $[Y]_\Psi = [X]_\Psi$.

in an accepted class Ψ such as:

$$\Psi_{\text{Affine}}(\mathbb{R}^2) = \left\{\textit{All rotations, reflections, positive scalings and translations of } \mathbb{R}^2\right\} \tag{2.67}$$

$$\Psi_{\text{Rigid}}(\mathbb{R}^2) = \left\{\textit{All rotations, reflections, and translations of } \mathbb{R}^2\right\} \tag{2.68}$$

These transformation groups can act on the landmark space by applying the transformations to the columns of X:

$$\psi \cdot \begin{bmatrix} X_1 & X_2 & \cdots & X_k \end{bmatrix} = \begin{bmatrix} \psi(X_1) & \psi(X_2) & \cdots & \psi(X_k) \end{bmatrix} \tag{2.69}$$

*for any $\psi \in \Psi_{\text{Rigid}}(\mathbb{R}^2) \subseteq \Psi_{\text{Affine}}(\mathbb{R}^2)$. Using the transformations in $\Psi_{\text{Affine}(\mathbb{R}^2)}$ results in **Affine Shape Space**, which is defined as:*

$$\mathbb{R}^{2\times k}/\Psi_{\text{Affine}}(\mathbb{R}^2) = \left\{[X]_{\Psi_{\text{Affine}}(\mathbb{R}^2)} : X \in \mathbb{R}^{2\times k}\right\} \tag{2.70}$$

*Using the rigid transformation results in Kendall's **Reflection-Size-and-Shape Space**:*

$$\mathbb{R}^{2\times k}/\Psi_{\text{Rigid}}(\mathbb{R}^2) = \left\{[X]_{\Psi_{\text{Rigid}}(\mathbb{R}^2)} : X \in \mathbb{R}^{2\times k}\right\} \tag{2.71}$$

These differ in that the size-and-shape space retains the scaling of the landmarks. Both of these spaces are useful in many downstream tasks. For example, consider a classification problem: given data $\{(X^{(i)}, Y_i)\}_{i=1}^n$ with $X_i \in \mathbb{R}^{2\times k}$ and $Y_i \in \{0,1\}$ how should we predict a label Y_0 on a new landmark collection $X^{(0)}$? A fully general model might assume:

$$Y_i \sim \text{Bernoulli}\big(f(X^{(i)})\big) \tag{2.72}$$

for some function $f : \mathbb{R}^{2\times k} \to [0,1]$, for all $i \in \mathbb{Z}_{\geq 0}$. Sometimes we are very confident that the orientation and position of the object will not affect the classification, so we could instead assume the simpler model:

$$Y_i \sim \text{Bernoulli}\big(\tilde{f}([X^{(i)}]_{\Psi_{\text{Rigid}}(\mathbb{R}^2)})\big) \tag{2.73}$$

for some function $\tilde{f} : \mathbb{R}^{2\times k}/\Psi_{\text{Rigid}}(\mathbb{R}^2) \to [0,1]$, for all $i \in \mathbb{Z}_{\geq 0}$. This model is simpler in the sense that $\tilde{f}$ can vary less than f could; if f is assumed to be in some hypothesis class $\mathcal{F}$ then the function $\tilde{f} \circ [\cdot]_{\Psi_{\text{Rigid}}(\mathbb{R}^2)}$ lives in the subspace of $\Psi_{\text{Rigid}}(\mathbb{R}^2)$-invariant functions within $\mathcal{F}$.

Example 16 (Functional Data Registration)**.** *Consider the situation of functional data analysis [Ramsay and Silverman, 2005], where our data objects f now live in infinite dimensional function spaces such as $L^2(\mathbb{R})$. Suppose that we have collected a number of function samples $\{f^{(i)}\}_{i=1}^n$ which each follow roughly the same form $f^{(i)}(x) = a\exp(-(a_i x - t_i)^2) + \epsilon_i(x)$ for some small magnitude noise process $\epsilon_i(x)$, some fixed shift times $t_i \in \mathbb{R}$ and scales $a_i \in \mathbb{R}_{>0}$,*

and a common scale term $a \in \mathbb{R}_{>0}$.

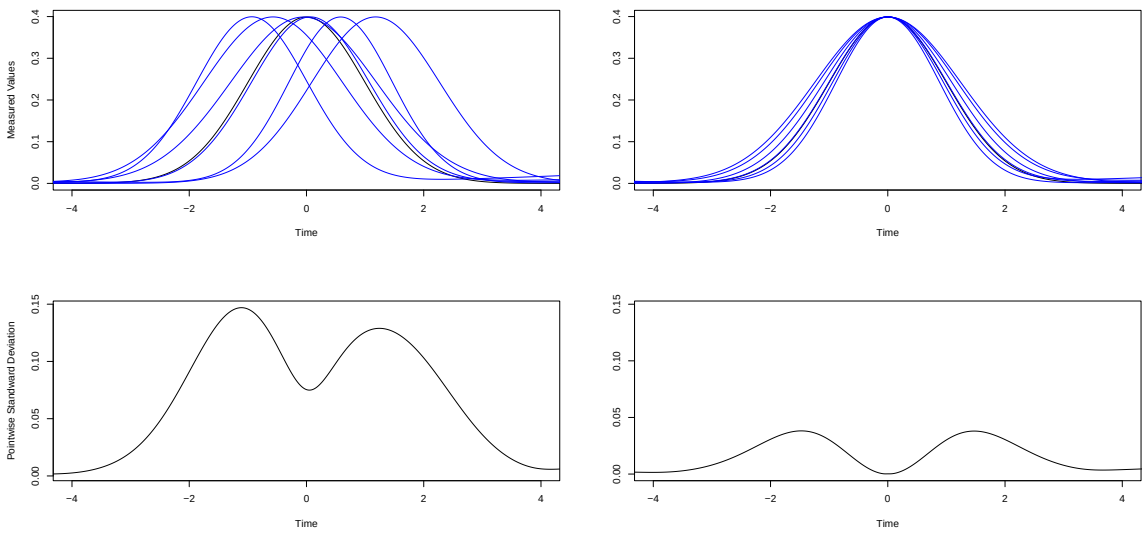


Fig. 2.3 An example of 6 curves $f^{(i)} = \exp(-(x - t_i)^2) + \epsilon_i(x)$ as described in Example 16. The top left plot shows the original curves are measured, the top right shows the curves after alignment by translations of the time axis. The lower plots show the pointwise sample standard deviation of the original curves (left) and the aligned curves (right). This shows how the amplitude variation is masked by the phase variation if we were to deal only with the unaligned curves.

If we want to describe the variation of these functions, it makes little sense to use the direct mean if the central locations t_i *are spaced out more than the bump around them spreads. Instead, we can decompose* $f^{(i)}$ *into a* ***phase*** $\phi_i(x)$ *and* ***amplitude*** $\tilde{f}_i$*:*

$$f^{(i)}(x) = \tilde{f}_i(\phi_i(x)) \tag{2.74}$$

and then think about the mean and variance of these components instead. In this case, we would have:

$$\phi_i(x) = x - t_i \tag{2.75}$$

$$\tilde{f}_i(x) = \exp(-x^2) + \epsilon_i(x) \tag{2.76}$$

and so we can give more meaningful descriptors of the mean phase shift $n^{-1}\sum_{i=1}^{n} t_i$ *and the variance around this mean and correctly identify that the amplitude variance comes only from the noise.*

In practice, we assume that the phase transformations belong to restricted class of transforma-

tions of the functional object's domain. Typical choices include the set of translations, strictly increasing linear transformations, and strictly increasing diffeomorphisms. These are all groups that are acting on the domain. The amplitude component can then be identified with the orbit of the function under the action of the phase transformations Ψ under composition, which lives in the quotient space $L^2(\mathbb{R})/\Psi$. This identification comes from noting that $\tilde{f}_i$ is the unique representative from the orbit $[f^{(i)}]_\Psi$ with $\tilde{f}_i = \phi_i^{-1} \cdot f^{(i)}$. Often this amplitude component is the component of importance, we are typically more interested in the shape of a COVID case number curve rather than which date it starts on, for example. This means that as with shape spaces we register the functions using the functions in Ψ and consider the amplitudes in the quotient space $L^2(\mathbb{R})/\Psi$.

2.5.4 Symmetries in Density Estimation

Much like nonparametric regression, nonparametric density estimation problems suffer from a curse of dimensionality. In these problems we have a sample $\{X_i\}_{i=1}^n$ drawn independently from a distribution with density $f : \mathbb{R}^p \to [0, \infty)$ and we seek to estimate f within some class of possible densities which we will denote $\mathcal{F}$. When $\mathcal{F}$ is the space of log-concave density functions and $d \geq 2$ the lower bound on the rate of estimation of f under squared Hellinger loss is:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\Big(\int_{\mathbb{R}^d} (f^{1/2} - \hat{f}^{1/2})^2 \Big) \geq c_d n^{-\frac{2}{d+1}} \tag{2.77}$$

for some positive constants c_d, as shown in Kim and Samworth [2016]. One approach to overcome this curse of dimensionality is to assume that the super-level sets $K_f(a; \mu) = \{u \in \mathbb{R}^d : f(u - \mu) \geq f(a - \mu)\}$ of the density f centered at μ are **homothetical**: that they are all scaled versions of some fixed convex $K \subseteq \mathbb{R}^d$: there exists a μ such that for all $a \in \text{range}(f)$ there exists an α with $K_f(a; \mu) = \alpha K_f$. If K_f is known then Xu and Samworth [2021] give an estimator $\hat{f}_n$ that has convergence:

$$\mathbb{E}\Big(\int_{\mathbb{R}^d} (f^{1/2} - \hat{f}_n^{1/2})^2 \Big) \leq C n^{-4/5} \tag{2.78}$$

for some positive constant C independently of the dimension d, and then give consideration of the behaviour for unknown K_f and for other divergences between densities. This homothetical super-level set condition imposes a non-trivial symmetry on the density function f. The authors take this as an invariance of the level sets under the action of scaling, however it also imposes an invariance condition on the density function f. Assume that K_f has interior containing 0 in the topology of $\mathbb{R}^d$. Define a bijection $\psi : \mathbb{R}^d \to \mathbb{R}^d$ by setting $\psi(u) = (\inf\{t \in [0, \infty) : u \in tK\})u$ for all $u \in S^d$, and enforcing positive homogeneity $\psi(au) = a\psi(u)$ for all $a \in \mathbb{R}_{\geq 0}$ and $u \in S^d$. This ensures that $\psi(K_f)$ is the unit ball in $\mathbb{R}^d$. Let $SO(d)$ act on $\mathbb{R}^d$ via:

$$g \cdot_\psi x = \psi^{-1}(g \cdot \psi(x)) \tag{2.79}$$

where $\cdot$ denotes the usual action of $SO(d)$. Then f (after centering around μ) is invariant to this new $\cdot_\psi$ action of $SO(d)$. Note that this action is not necessarily smooth, but will be if ψ is a diffeomorphism. With this reframing we see that homotheticity is another example of using symmetry as a dimension reduction technique.

2.5.5 Symmetries of Regression Functions

Symmetry has recently gain significant traction in the machine learning community, where the symmetries of image analysis are often obvious and exploitable. For example, the classification of an image of a handwritten digit in the MNIST dataset [LeCun et al., 1998] should not depend on the precise position of the digit in the pixel grid. This is equivalent to the statement that the classification function of any particular digit is invariant to the group of translations at the pixels (that loop around the edges to preserve the information). This idea was for a long time implicitly captured by convolutional neural networks, and recent efforts have moved to make this framing more explicit [Kondor and Trivedi, 2018, Bronstein et al., 2021] which we will discuss in the next section. Here we will consider cases where the symmetries are considered implicitly, examples of which have a long history in statistics.

Example 17 (Sparsity in Linear Regression)**.** *Consider the normal linear model $Y_i = X_i\beta + \epsilon_i$ where $X_i \sim \mu_i$ are row vectors in $\mathbb{R}^p$ valued and $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ independently of all the X_i. When the dimension of the β is large but many entries are 0, we often get estimators that perform better when we make use of this information. This comes from a shift from the variance in these coordinates to the bias we experience if we make errant assumptions about which entries are 0. Specifically, we say that $\beta \in \mathbb{R}^p$ is s-sparse if there exists a set of indices $\mathcal{J} \subseteq \{1,\dots,p\}$ with $|\mathcal{J}| = p - s$ and $\beta_j = 0$ for all $j \notin \mathcal{J}$.*

This condition on β is an invariance condition of the regression function $f(x) = x^T\beta$. Consider the action of $\mathbb{R}^s$ on $\mathbb{R}^p$ given by:

$$(a_1,\dots,a_s)\cdot(x_1,\dots,x_p) = \left(x_1 + \sum_{j=1}^{s} a_j \mathbf{1}_{\mathcal{J}_j^c=1}, \cdots, x_p + \sum_{j=1}^{s} a_j \mathbf{1}_{\mathcal{J}_j^c=p}\right) \tag{2.80}$$

where $\mathcal{J}^c$ are the indices in $\{1,\dots,p\}$ but not in $\mathcal{J}$, ordered in a vector with s-components. This action shifts the vector x along the basis vectors with indices not in $\mathcal{J}$. In this case:

$$f(a\cdot x) = \left(x_1 + \sum_{j=1}^{s} a_i \mathbf{1}_{\mathcal{J}_i^c=1}, \cdots, x_p + \sum_{j=1}^{s} a_i \mathbf{1}_{\mathcal{J}_j^c=p}\right)\beta \tag{2.81}$$

$$= \sum_{k=1}^{p}\left(x_k + \sum_{j=1}^{s} a_j \mathbf{1}_{\mathcal{J}_j^c=k}\right)\beta_k \tag{2.82}$$

$$= x\beta + \sum_{k=1}^{p}\sum_{j=1}^{s} a_j \mathbf{1}_{\mathcal{J}_j^c=k}\beta_j \tag{2.83}$$

Now the terms in the final sum all vanish because $\mathbf{1}_{\mathcal{J}_j^c=k}\beta_k = 0$ *by definition—if the indication function is not* 0 *then* $\beta_k = 0$*. Thus, the regression function in this case is invariant to this group action.*

Example 18 (Sparsity in Nonparametric Regression)**.** *Consider now the nonparametric version of the linear regression example above. We say that a (continuously differentiable) regression function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* s *sparse if it depends only on the covariate indices in* $\mathcal{J} \subseteq \{1, \dots, p\}$ *with* $|\mathcal{J}| = p - s$*. This can be formulated as* $\frac{\partial f}{\partial x_j}(x) = 0$ *for all points* $x \in \mathbb{R}^d$*. It can equally be framed as a translation invariance as above, where the action is given in Example 17. In this case, the sparsity means that we know that* $f(x + a_j e_j) = f(x)$ *(a consequence of the mean value theorem; as otherwise there would have to be a point with non-zero partial derivative) for all* $j \in \mathcal{J}^c$*. Thus, again we have that* f *is* s*-sparse if and only if* f *is invariant to the action of* $\mathbb{R}^s$ *determined by* $\mathcal{J}^c$*.*

Example 19 (Multi-index Models)**.** *In the same context as Example 18, the notion of sparsity can be generalised to sparsity in some new choice of basis for* $\mathbb{R}^d$*; the idea is that the function itself can be sparse even if we have chosen the wrong measurement frame. In this case we consider the class of* ***multi-index models*** *[Hristache et al., 2001] where:*

$$f(x) = h(Rx) \tag{2.84}$$

for some rank $d - s$ *orthogonal projection matrix* $R : \mathbb{R}^d \to \mathbb{R}^{d-s}$*, and some link function* $h : \mathbb{R}^{d-s} \to \mathbb{R}$*. In this case the invariance of* f *is described by any translation in the null space of* R *as:*

$$f(x + u) = h(Rx + Ru) = h(Rx + 0) = f(x) \tag{2.85}$$

for all $u \in \ker R$*. This kernel forms a group under addition (as all linear spaces do) that acts on* $\mathbb{R}^d$ *by translation* $u \cdot x = x + u$*. Therefore, in this case* f *obeys a multi-index model parametrised by* R *if and only if* f *is* $(\ker R)$*-invariant.*

Example 20 (Regression on Graphs)**.** *A graph is a (finite) set of nodes* N *and a set of edges* $E \subseteq N \times N$ *that capture relationships between nodes. These nodes and edges are often given weights that measure the intensity of these relationships, captured by a function* $w : N \times N \to [0, 1]$ *with* $w(a, b) = w(b, a)$ *and* $w(a, b) = 0$ *if and only if* $(a, b) \notin E$*. Graphs are frequently used to model causal relationships [Pearl, 2009], social networks [Otte and Rousseau, 2002], and molecules [Gaudelet et al., 2021] among many other use cases. In the case of molecules the nodes are individual atoms, edges denote covalent bonds, and weights might contain information such as the distance between the atoms. Graphs are usually stored as an adjacency matrix with entries* $A_{ij} = w(a_i, a_j)$ *for* $a_i, a_j \in N$*.*

Often we are interested in predicting the value of a response from a covariate graph; for

example, the toxicity or efficacy of a candidate drug molecule given a large library of previous measurements. We know that the regression function $f : A \mapsto f(A)$ must obey a permutation symmetry as follows. Let $\mathrm{Sym}(|N|)$ *be the group of permutations σ of the integers* $1, \ldots, |N|$ *(i.e., σ is a bijection of this set) and let it act on the space of adjacency matrices by permuting the rows and columns:*

$$\sigma \cdot \begin{bmatrix} A_{11} & \cdots & A_{k1} \\ \vdots & \ddots & \vdots \\ A_{k1} & \cdots & A_{kk} \end{bmatrix} = \begin{bmatrix} A_{\sigma(1)\sigma(1)} & \cdots & A_{\sigma(k)\sigma(1)} \\ \vdots & \ddots & \vdots \\ A_{\sigma(k)\sigma(1)} & \cdots & A_{\sigma(k)\sigma(k)} \end{bmatrix} \tag{2.86}$$

Then f must be $\mathrm{Sym}(|N|)$*-invariant because this action is simply a relabelling of the nodes; it cannot change the physical properties of the molecule.*

2.6 How to Use Symmetry in Regression

Given all of these examples of symmetries in regression, the obvious question to ask is: *"how can we exploit a known symmetry to improve our estimators?"*. There are four main approaches that have been used to exploit a known symmetry in regression in the recent literature. Suppose that f is known to be G-invariant for some group G acting on the covariate space $\mathcal{X}$. We can improve our estimate of f by using:

1. **(Data Augmentation)** where we add samples of the form $(g_{ij} \cdot X_i, Y_i)$ for some chosen number of samples $g_{ij} \in G$;

2. **(Data Projection)** where we explicitly consider the quotient space $\mathcal{X}/G$ and doing regression either using the dataset $\{([X_i]_G, Y_i)\}_{i=1}^n$ or $\{(\tilde{X}_i, Y_i)\}_{i=1}^n$ where $\tilde{X}_i$ are representatives of each X_i in a specified fundamental domain;

3. **(Kernel Symmetrisation)** where we replace a kernel K used in kernel ridge regression or local polynomial estimation with a G-invariant kernel K_G; and

4. **(Feature Averaging)** where we estimate f using the original dataset and then smooth this estimate over the orbits using a (partial) symmetrisation operator, giving an estimate of the form $S_\rho f_n$.

Each of these methods comes with various advantages and disadvantages both statistical and computational. For example, data augmentation can be useful in a large neural network where we can now bring n up to the order of the number of parameters; however, it can be harmful when considering confidence intervals in a linear regression setting because the apparent larger sample size can lead to liberal confidence intervals for the parameters [Huang et al., 2022]. It is attractive because of its simplicity; it only requires sampling from the group and avoids computation of quotient spaces.

Data projection is the tool most often used in sparse and multi-index models because the fundamental domains are particularly easy to describe as hyperplanes spanned by the important directions with projections as simple as forgetting variables or an orthogonal projection matrix. In these cases analysis of the resulting estimator can be tackled because the quotient space has a known form with similar properties to the original space (i.e., it is still a vector space). Thus, usual theorems can be readily applied to show that we can achieve appropriate dimension reduced rates. However, in general the quotient space can have complicated topological structure—at best it might look like a manifold with singularities (a space known as an orbifold or V-manifold [Satake, 1956]) and at worst it can lose all sense of the original structure of $\mathcal{X}$. This means that the fundamental domains and projections to them are sometimes difficult to construct in general and the resulting estimators can be difficult to analyse, which means that this method is not well studied in full generality.

Kernel Symmetrisation benefits because there is only one extra step relative to the original estimator; you need to take K and compute $K_G = S_G K$ (where the integration is with respect to the first coordinate). This is usually easier than constructing the quotient space for data projection. Analysis of kernel symmetrised ridge regression has been recently undertaken. Elesedy [2021] showed that the performance improvement over the unsymmetrised estimator is strictly positive almost surely, but their bound decays as $O(n^{-1})$ so does not guarantee dimension reduction. This is done by examining the specific closed form solution of the kernel ridge regressor. Bietti et al. [2021] showed that if G is finite then the risk of a kernel symmetrised ridge regressor converges to $1/|G|$ of the risk of the original estimator. Their analysis centres on the proportion of spherical harmonics that are G-invariant, which can be enumerated combinatorially when G is finite. They note in their paper that this technique seems difficult to extend to the case of continuous groups where we can achieve dimension reduction. Tahmasebi and Jegelka [2023] has even more recently shown that if G is a positive dimensional compact Lie group acting isometrically and the orbits have positive dimension then a kernel symmetrised ridge regressor will achieve dimension reduction, with worst risk that decays with a rate that depends on $d - \dim[x]_G$ rather than d. This result considers the spectral behaviour of the Laplace-Beltrami operator Δ_f (which is a generalisation of the usual Laplacian of f on $\mathbb{R}^d$ to a function f on a smooth manifold). The key is that the projection of these eigenspaces onto $L^2_G(\mathcal{X})$ have measurably smaller entropy than the original spaces when f is G-invariant, which translates to faster learning of f.

Feature averaging can also be shown to give an almost sure performance improvement, as originally done in Lyle et al. [2020]. This essentially follows from the fact that S_G is an orthogonal projection when G is compact and the sampling distribution of the covariates is

G-invariant. In this case, we have:

$$\|S_G f_n - f\|_2 = \|S_G f_n - S_G f\|_2 \leq \|S_G\|_{\text{Op}} \|f_n - f\|_2 = \|f_n - f\|_2 \tag{2.87}$$

almost surely for any estimator f_n. The fact that S_G is an orthogonal projection ensures that its operator norm is 1 (as it is self-adjoint and all eigenvalues are either 1 or 0). This also leads to a neat characterisation of the asymptotic bias when f is not G-invariant, as it must be equal to the anti-symmetric part of f. One problem of this and the kernel symmetrisation approach is that the symmetrisation operator S_G is only defined for compact groups with a uniform distribution. Extensions using a partial symmetrisation operator for non-compact G have not been well studied.

2.7 Symmetry Discovery

Because of the interest in G-invariant and G-equivariant models in the machine learning community, there have been a few efforts to find the symmetries that f is invariant to. In the machine learning literature this is called **symmetry discovery**. Suppose that $\mathcal{G}$ is a group that acts continuously on the covariate space. The following three methods, as well a couple of slight variants, are the only known methods at the time of writing. These algorithmic approaches suffer for our problem later in three key ways: they only consider parametric classes of subsets of $\mathcal{G}$ rather than all subgroups of $\mathcal{G}$; they do not have easily accessible statistical properties; and they have a much higher tolerance for symmetries that only hold locally (which can lead to asymptotic bias).

2.7.1 Auto-Augment

Auto-augment [Cubuk et al., 2019] represented a shift in the data augmentation world from finding some extra data that probably has the same labels in an image classification task (as is done with generative adversarial models, for example) to searching for transformations of the image space that can be applied to the data set to give new data. In this approach, only a finite set of transformations in pythons image library (PIL) were considered, and the transformations to include are found by training a large recurrent neural network. This makes statistical properties of the method hard to access, though it is validated empirically on several well known test datasets in the machine learning community (CIFAR-10 and CIFAR-100 [Alex, 2009], and SVHN [Netzer et al., 2011]). This strategy was later extended to other tools such as Fast-Auto-Augment [Lim et al., 2019].

2.7.2 Augerino

The next step to learn data-augmentations was Augerino [Benton et al., 2020]. This approach considers searching over a parametric class of subsets of $\mathcal{G}$. Let $\theta \in \mathbb{R}^{\dim \mathcal{G}}$ define a box in the tangent space of $\mathcal{G}$ at the identity e (this space is sometimes called the Lie Algebra) with orthonormal basis vectors t_i:

$$A_\theta = \left\{ a \in T_e\mathcal{G} : a = \sum a_i t_i, |a_i| \leq \theta_i \right\} \tag{2.88}$$

These boxes define subsets of $\mathcal{G}$ through the exponential map:

$$G_\theta = \left\{ \exp_e(a) : a \in A_\theta \right\} \tag{2.89}$$

The optimal subset is then estimated by minimising the regularised loss:

$$L(\theta, w) = \mathbb{E}\big(\mathbb{E}_{g \in U(G_\theta)}\big(f_n(gX; w)\big)\big) + R(\theta) \tag{2.90}$$

where $X \sim \mu$ is the measure on our covariate space and w contains the weights for the neural network f_n. One suggested choice of the regularisation component is $R(\theta) = -\lambda \|\theta\|_2$ to encourage more symmetry. This technique can only lead to a estimated subset rather than a subgroup. The only lower dimensional subgroups available to estimate are of the form:

$$G_i = \left\langle \exp_e(\alpha_i t_i) : \alpha_i \in \mathbb{R} \right\rangle \tag{2.91}$$

or the compositions $\langle G_i : i \in \mathcal{I} \rangle$ for some index set $\mathcal{I}$. Since these groups are highly dependent on the original choice of the t_i, we miss many of the possible subgroups of $\mathcal{G}$ that f may be invariant to. For example, if $f : \mathbb{R}^2 \to \mathbb{R}$ is invariant to translations along the axis $(1, 1)$ only, then these would be missed if we picked the usual orthonormal basis vectors for $T_0\mathbb{R}^2 \simeq \mathbb{R}^2$.

2.7.3 SCALE

The latest effort to learn the appropriate transformation is SCALE [Chatzipantazis et al., 2023]. This method largely resembles Augerino but it adapts the distribution of the random variable g. Augerino uses the uniform distribution on the subsets G_θ, but SCALE instead samples g using a compositional approach as follows:

1. Fix a K-vector of probabilities $(\pi_i, \ldots, \pi_K) \in (0, 1)^K$ and K different distributions U_i on the group $\mathcal{G}$;
2. Sample $g_i \sim U_i$ for $i = 1, \ldots,$;
3. Sample a K-vector $I_i = 1$ with probability π_i and 0 with probability $1 - \pi_i$; and
4. Compute $g = \prod_{i=1}^K g_i^{I_i}$.

The probabilities π_i and the distributions U_i can then be treated as learnable parameters. If we constrain U_i to the form $U_i = U(\exp_e(-\theta_i t_i), \exp_e(\theta_i t_i))$ (where θ_i and t_i are as in the Augerino section), then we can optimise the loss function above in terms of π and θ. This offers some more flexibility in the way in which the data would get augmented, in particular it utilises the compositional structure of the group $\mathcal{G}$. It still encounters the same problem if missing diagonal subgroups that are not contained in the distributions U_i. This can be alleviated somewhat by increasing K and including more of them, but there will still usually be infinitely many that we cannot access in this method for any fixed K.

2.8 Estimation of Level Sets

In regression, there has been some interest in the estimation of **superlevel sets** $A(f, \tau) = \{u \in \mathcal{X} : f(u) > \tau\}$ for regression functions f [Cavalier, 1997, Cuevas et al., 2006, Reeve et al., 2023, Müller et al., 2023]. This problem is closely related to the problem of identifying the superlevel sets of an unknown density function [Tsybakov, 1997]. This has many applications for clustering the samples into bands with similar response values, such as in precision medicine where we might seek to separate a population based on likely treatment outcomes.

In the following chapter we will examine another related problem: the estimation of the **level set** $[x]_f = \{u \in \mathcal{X} : f(u) = f(x)\}$. Note that some of the earlier literature [Tsybakov, 1997, Cavalier, 1997] uses the term level set to refer to a superlevel set, but this is largely inconsistent with the mathematical literature and the more recent approaches. This mainly differs from the estimation of superlevel sets in the loss function used to judge the estimate. In the case of superlevel sets, two common measures are the symmetric difference:

$$d_\Delta(A, B) = \mu\big(\{x \in A : x \notin B\} \cup \{x \in B : x \notin A\}\big) \tag{2.92}$$

and the Hausdorff metric:

$$d_{\text{Haus}}(A, B) = \max\big(\sup_{a \in A} \inf_{b \in B} d_{\mathcal{X}}(a, b), \sup_{a \in A} \inf_{b \in B} d_{\mathcal{X}}(a, b)\big) \tag{2.93}$$

used in Cavalier [1997], Tsybakov [1997]. Recent approaches [Müller et al., 2023] focus on one sided error $\mu(\{x \in A : x \notin B\})$ (where A will be the true level set and B the estimate usually) along with probabilistic guarantees on the event $\{B \subseteq A\}$.

For estimating level sets, we must shift focus from the entire sets to only their boundary. We will later use the Hausdorff metric on these boundaries. Estimates of the superlevel set with low error in any of these metrics will likely have a boundary that is close to the boundary of the true level set, but this is not guaranteed in the worst case for any of these

metrics without strong conditions on the smoothness and curvature of the boundary. Indeed as the boundary forms a measure 0 subset it can deviate significantly without creating large changes under theses metrics. This is partly the reason for the assumptions made in Cavalier [1997] on the shape of the level sets. As of writing there has been no work that studies the performance of level set estimators in regression under this Hausdorff loss.

It is worth examining Müller et al. [2023] from a methodological perspective because of some parallels with the work in Chapter 5. In this paper the superlevel sets are estimated by testing whether each sampled covariate point X_i is contained in the true superlevel set $A(f, \tau)$ for the fixed $\tau \in \mathbb{R}$. This is done with a supermartingale test that exploits the *isotonicity*, i.e., non-decreasing with respect to the partial ordering $x \preceq y$ if and only if $x_i \leq y_i$ for all i, of the regression functions considered.

This isotonic property also allows for a clever combination of the p-values of the hypotheses $H_0^{(i)} : X_i \in A(f, \tau)$ because if $X_i \preceq X_j$ then $H_0^{(j)}$ is true whenever $H_0^{(i)}$ is, or conversely $H_0^{(i)}$ is false whenever $H_0^{(j)}$ is. This is done by constructing an appropriate directed acyclic graph (DAG) with nodes X_i and edges from X_j to X_i if $X_i \preceq X_j$ with no X_k distinct from X_i and X_j such that $X_i \preceq X_k \preceq X_j$. They then apply a testing procedure that takes advantage of the logical structure of the hypotheses. Specifically, rejection at a node automatically informs rejection of later nodes (in the direction of the edges). This means that the we do not need to test all nodes for their inclusion in $A(f, \tau)$; we can test the roots first and only move forward if we do not reject at the significance level α_i for the hypothesis $H_0^{(i)}$. The estimated level set is taken as the *upper hull* of points $\hat{A} = \{x \succeq X_i : H_0^{(i)} \text{ is not rejected}\}$, where the rejections are given by this nested testing procedure. By careful consideration of the family wise error rate the significance levels α_i can be chosen to maximise the risk $\mathbb{E}(\mu(A(f, \tau) \setminus \hat{A}))$ while ensuring that $\mathbb{P}(\hat{A} \subseteq A(f, \tau)) \geq 1 - \alpha$ for the chosen significance level $\alpha \in (0, 1)$. This nested testing approach bears some similarity to the approach we will use in Chapter 5 later.

Chapter 3

Regression Using Level Sets

In this chapter, we will show that knowledge of a regression function's level set $[x]_f = \{u \in \mathcal{X} : f(u) = f(x)\}$ allows for faster estimation of $f(x)$ than the global minimax rate for the Hölder hypothesis class for f considered. We show that a naive plug-in estimator for the level set can also achieve faster rates, but this requires some strong conditions. This then leads to the natural question of whether we can construct meaningful hypothesis classes for the level sets that allow for fast rates without such strong assumptions on the regression function. The later chapters will give an answer to this question.

3.1 Introduction

In nonparametric regression, we collect data $\{(X^{(i)}, Y_i)\}_{i=1}^n$ and seek to estimate the regression function f in the model $Y_i = f(X^{(i)}) + \epsilon_i$ for independent and identically distributed noise variables ϵ_i with mean 0 and finite homoskedastic variance σ^2. In this chapter, the covariates $X^{(i)}$ will be deterministically chosen points in the unit hypercube $\mathcal{X} = [0,1]^d$ and the response variables Y_i are real valued. The regression functions considered will be constrained to the Hölder hypothesis class $\mathcal{F}(L, \beta)$ defined in section 2.1. As noted there, the minimax rate for the estimation of f under L^2 loss is known to be determined by the lower bound:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}\big(\|\hat{f} - f\|_2^2\big) \geq c n^{-\frac{2\beta}{2\beta+d}} \tag{3.1}$$

which is achieved by local polynomial estimators (among others) as shown in Stone [1982]. The dependence on dimension causes this rate to slow dramatically for higher d, a curse well known in the non-parametric statistical literature.

As discussed in section 2.1, the main routes to improve on this curse (when μ is not concentrated around a lower dimensional manifold) are consideration of additive models for f and sparse models for f. In this section we develop a new route that considers the level

sets of f, denoted by $[x]_f = \{u \in \mathcal{X} : f(u) = f(x)\}$. Knowledge of the level set $[x]_f$ should clearly help with the estimation of $f(x)$. In particular, it is clear that knowledge of $[x]_f$ tells us about the directions that ∇f must vanish in: at all of the directions tangent to $[x]_f$ at x (assuming that the level set is a smooth submanifold of $\mathcal{X}$).

In Section 3.3, we quantify how much this knowledge can improve the estimation of $f(x)$. We do this by providing upper bounds on the worst case risk of the **Manifold Averaged Estimator** (MAVE) denoted f_n^{MAVE}. This estimator uses a base estimator f_n of f (e.g., a local polynomial estimator) and constructs points in a local neighbourhood of x intersected with $[x]_f$. The estimate of $f(x)$ is then taken as the average of the base estimator's values on this collection of points. If the level set is well specified, this does not inflate the bias of the MAVE at x. If the base estimator is sufficiently local (see assumption E1 later), as with local polynomial estimators, then this can cut the variance of the estimate dramatically. This allows us to re-optimise any hyper-parameter choices in the base estimator. We show that under these new optimal choices that:

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}\big((f_n^{\text{MAVE}}(x; [x]_f) - f(x))^2\big) n^{\frac{2\beta}{2\beta + d - \dim[x]_f}} (\log n)^{d - \dim[x]_f} \leq C_1 \tag{3.2}$$

where C_1 is a positive constant. Note that the powers of the terms containing n depend on f so these must occur inside of the supremum on the left. Thus, the MAVE achieves the desired dimension reduction—it removes all dimensions in the level set of f at x (up to log-factors). Usually[1] level sets are dimension $d-1$ sub manifolds of $\mathcal{X}$, which brings this rate down to the rate (up to log factors) for well specified additive models. At other points we might have dimension d level sets where f is constant, where we achieve the parametric rate (up to log factors). When the function is s-sparse, the true level sets contain a dimension $d-s$ hyperplane and so the MAVE achieves the best possible rate for such sparse functions. More importantly, it also allows for non-linear level sets that go far beyond just the case of sparse functions.

This analysis requires an oracle with knowledge of the true level set of the unknown regression function. In the same section we examine the robustness of this process to misspecification of the level set, and bound the risk in terms of a local Hausdorff distance between the true level set and the submanifold used. This leads to the natural question: if we estimate the level set $[x]_f$ and use this in the MAVE, does it lead to a dimension adaptive estimator of f?

Previous efforts to estimate the level sets of regression functions [Cavalier, 1997, Reeve et al., 2023, Müller et al., 2023] have usually come from the problem of estimating super-level

[1] See section 3.2.1: in many cases this is an almost sure statement.

sets, i.e., sets $A(f,\tau) = \{x \in \mathcal{X} : f(x) \geq \tau\}$ for which $[x]_f$ is the boundary of $A(f, f(x))$. As such, little is known about the performance of these estimators in terms of the local Hausdorff loss we consider, and it is unclear how to use their loss functions to achieve a similar result. Because of this, we consider in section 3.4 perhaps the most naive estimator of $[x]_f$, the plug in estimator $[x]_{p_n}$ where p_n is an independent local polynomial estimator of f. Under a strong assumption on the regression function f—that it is equal to a polynomial on some neighbourhood of x—we can show that we achieve the oracle's rate (up to log factors) using $f^{MAVE}(x, [x]_{p_n})$. We then show this in some numerical experiments both in situations that satisfy this strong condition and others.

3.1.1 Main Contributions

We now state the results of this chapter in more technical language. All terms defined later are italicised. Let f be a regression function in the Holder class $\mathcal{F} = \mathcal{F}(L,\beta)$ for $\beta > d$. Let f_n be a minimax optimal estimator of f over this hypothesis class that satisfies assumption set (E) under an appropriate design regime.

1) We define the Manifold Averaged Estimator (MAVE) for $f(x)$ as the average of the estimate $f_n(v_i)$ for points v_i constructed from the level set $[x]_f = \{v \in \mathcal{X} : f(v) = f(x)\}$ that are evenly rectangularly spaced in the tangent space $T_x[x]_f$;

2) We show that if this level set is well specified and a smooth submanifold of $\mathcal{X}$ then this estimator achieves asymptotic dimension reduction—that $\sup_{f\in\mathcal{F}} \mathbb{E}((f_n^{MAVE}(x) - f(x))^2) \leq Cn^{-\frac{2\beta}{2\beta+d-\dim[x]_f}}$;

3) We show that this estimator has some robustness to misspecification of the level set, and give the bound $MSE(f_n^{MAVE}(x, M)) \leq 2Bh^{2\beta} + \frac{V}{nh^d}\mathbb{E}(m^{-1}) + 2L^2\mathbb{E}\big(d_{x,h_O}(M, [x]_f)^2\big)$;

4) We consider the behaviour of the MAVE when using estimated level sets of the form $[x]_{p_n}$ for a local polynomial estimate p_n of $f(x)$. When these estimates are independent (e.g., from data splitting) and when f agrees with a polynomial on some neighbourhood of a *regular point* x, we show that $MSE(f_n^{MAVE}(x, [x]_{p_n})) \leq C_* n^{-\frac{2\beta}{2\beta+1}} (\log n)^{(d-1)/\alpha}$.

3.2 Background

The results in this chapter rely on understanding the material in the differential geometry section of Chapter 2, and the estimating level sets section of Chapter 2 is helpful for the context of later parts in section 3.4. Alongside this, we need some specific definitions and results from the field of differential topology that consider the geometry of level sets of functions.

3.2.1 Geometry of Level Sets

If $f : [0,1]^d \to \mathbb{R}$ is a sufficiently smooth function, in this case at least d-times continuously differentiable, then many of the level sets are forced to take a specific structure of a $d-1$ dimensional submanifold of $[0,1]^d$. This is formalised in the following two theorems. Intuitively, these relate the behaviour of the gradient $\partial f/\partial x = [\partial f/\partial x_1, \ldots, \partial f/\partial x_d]$ to the behaviour of the level set. Clearly this gradient when viewed as a linear operator can only have rank 1 or 0, and so at least $d-1$ directions appear in its null space. If it is rank 1 for all $u \in [x]_f$, for any $x \in [0,1]^d$ then we say that the level set is **regular** and that the value $f(x)$ is a **regular value** of f. This property means that the kernel of $\partial f/\partial x$ can be identified with the tangent space of $[x]_f$ which gives the dimension in the following theorem. We say that other values of $f(x)$ are **critical values**, and the level sets with critical values are **critical level sets**.

Theorem 21 (Regular Level Set Theorem)**.** *Suppose that $f : [0,1]^d \to \mathbb{R}$ is a differentiable function and suppose that c is a regular value of f. Then the level set $[x]_f = \{u \in [0,1]^d : f(u) = c\}$ is a smooth embedded submanifold of $[0,1]^d$ of dimension $d-1$.*

A proof can be found in Lee [2013], essentially as a consequence of the implicit function theorem. The name *regular* comes from the following theorem due to Morse [1939] (in the one dimensional codomain case) and later Sard [1942] (in the higher dimensional codomain case). This result states that almost all values f can take are regular.

Theorem 22 (Morse–Sard Theorem)**.** *Suppose that $f : [0,1]^d \to \mathbb{R}$ is a k-times continuously differentiable function for $k \geq d$. Then the set of critical values has Lebesgue measure 0.*

Note that this does not imply that almost all input values lie in regular orbits: if f is constant on a set $U \subseteq [0,1]^d$ of positive Lebesgue measure, then these points fall into a critical level set of dimension d. Other critical level sets can take any dimension between 0 and d. This theorem does restrict the measure of the lower dimensional critical level sets—there will usually only be a countable number of these critical values (except in some pathological cases of uncountable but Lebesgue measurable sets of critical values) and so the countable union of the lower dimensional manifolds must also be μ-negligible.

3.3 Manifold AVeraged Estimator (MAVE)

Let f be any regression function and take any point $x \in \mathcal{X} = [0,1]^d$ such that the level set $[x]_f$ is a smooth submanifold of $\mathcal{X}$. Suppose that we have collected data $\{(X_i, Y_i)\}_{i=1}^n$ according to the nonparametric regression model:

$$Y_i = f(X_i) + \epsilon_i \tag{3.3}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We suppose in this section that we know the smooth level set $[x]_f$. Our goal in this section is to construct an estimator that uses this known level set to better estimate $f(x)$.

Suppose that we have an estimator f_n of f that we will call the **base estimator**. This could be anything from a local polynomial regressor to a nonparametric least squares estimator, to a neural network. Local polynomial estimators satisfy the assumption set (E) used to derive theoretical properties later, so this is a good choice and what we will use in simulations later. We will for this chapter consider only the Hölder class with $\beta > d$ as suggested by the Morse-Sard theorem, with a brief discussion of other classes in section 3.6.

Given this base estimator, we will give a deterministic construction of **averaging points** $\{v_i\}_{i=1}^m \subseteq [x]_f$. The number of points m will depend on two bandwidths: $h, h_O \in \mathbb{R}_{>0}$ and we will give optimal choices for these later. With these averaging points we can then define our **Manifold AVeraged Estimator (MAVE)** as:

$$f_n^{\text{MAVE}}(x; [x]_f) = \frac{1}{m} \sum_{i=1}^{m} f_n(v_i) \tag{3.4}$$

The bandwidth h will control the locality of the base estimator and the bandwidth h_O will control for much of the level set we average over.

3.3.1 Constructing Spaced Points

To construct the set $\{v_i\}$ we need to examine the geometry of the level set $[x]_f$. Let $W_x^{[x]_f}$ be the largest hypercube[2] in $T_x[x]_f$ that is contained in the orthogonal projection of $[x]_f$ onto $T_x[x]_f$, where this projection is taken in the ambient space $\mathcal{X} = [0,1]^d$. Let $r(x, [x]_f)$ be the the minimum of: (1) half of the side length of the hypercube $W_x^{[x]_f}$; or (2) the bandwidth h_O. If the orbit is 0-dimensional, then we take $r(x, [x]_f) = 1$. We call $r(x, [x]_f)$ the **shadow radius** of $[x]_f$ at x. We construct the averaging points $\{v_i\}_{i=1}^m$ with η spacing as follows:

(i) Pick any rectangular grid[3] $\{a_i\}_{i=1}^m$ in $W_x^{[x]_f}$ such that $d(a_i, a_j) \geq \eta$ whenever $i \neq j$; and

(ii) Orthogonally (with respect to $T_x[x]_f$) project this grid onto $[x]_f$, giving a set $\{v_i\}_{i=1}^m \subseteq [x]_f$.

This is shown in Figure 3.1. The rectangular grid is usually easy to specify, with points $a_i = x + \sum \alpha_i \eta w_i$ where w_i are the basis vectors of $T_x[x]_f$ and α_i are integers in the range:

$$\left\{ -\lfloor r(x, [x]_f)/\eta \rfloor \}, \ldots, 0, \ldots, \lfloor r(x, [x]_f)/\eta \rfloor \right\} \tag{3.5}$$

[2]Note that this is well defined as one must exist, as a consequence of being in the interior of the level set, and the side lengths here are bounded above by the domain $\mathcal{X}$.

[3]i.e., there is a coordinate basis in which the points are evenly spaced along each dimension.

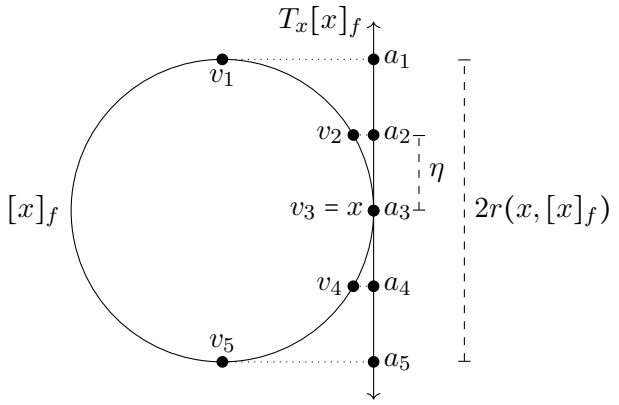


Fig. 3.1 The level of $x \in \mathbb{R}^2$ for a function $f(x) = f_0(\|x\|_2)$, along with the points $\{a_i\}$ and $\{v_i = g^i_{x,h} \cdot x\}$ that will be used to manifold-average the estimator f_n. Note that these need not be spaced evenly around the level set, nor cover it completely.

The projection is also usually easy because it is the solution to $f(a_i + z) = f(x)$ for $z \in \ker(\partial f/\partial x)^\perp$, and in practice we will only have to solve this for polynomial f. Note that the kernel is either rank 1 or 0 and so z is in at most a one dimensional space. In this case, the kernel and its normal space is easy to compute and the equation can be solve exactly for low degrees ($k \le 5$, as the differentiation removes one degree) and numerically with arbitrary precision for higher degrees.

We now give a lower bound on m when the points are required to be $2h$ separated to ensure the independence of the samples $f_n(v_i)$ under the strict locality assumption later. We will later relax this but show that this bound also captures the proportion of samples that contribute to variance reduction even as we take $\eta \to 0$.

Proposition 23. *For all $x \in \mathcal{X}$ and $h, h_O > 0$ pick $\{v_i\}_{i=1}^m$ with $2h$ spacing as above. Then we have $m \ge \max(1, (r(x, [x]_f)(2h)^{-1})^{\dim[x]_f})$, and $d_{\mathcal{X}}(v_i, v_j) \ge 2h$ whenever $i \ne j$. Moreover, for sufficiently small h_O, i.e. $h_O \le r(x, [x]_f)$, this reduces to:*

$$m \ge \max\left(1, \left(\frac{h_O}{h}\right)^{\dim[x]_U}\right) \tag{3.6}$$

Proof. This is simply the bound that results from counting the possibilities of the α_i above when $\eta = 2h$. □

3.3.2 Theoretical Properties

We now derive rates of convergence of the MAVE under some assumptions on the base estimator. Suppose that our base estimator satisfies the following properties:
Assumption set (E):

(E1) (Strict Locality) Conditioned on $\mathcal{D}_X$, $f_n(x)$ and $f_n(y)$ are independent whenever $d(x,y) \geq 2h$;

(E2) (Default Value) If the ball $B_{\mathcal{X}}(x,h)$ contains no X_i, then $f_n(x) = 0$;

(E3) (Optimality) There exist positive constants $B, V \in \mathbb{R}_{>0}$, such that for all $f \in \mathcal{F}$, for μ-almost all x, and for all sufficiently large n:

(a) (Bias) If $B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X$ is non-empty, then $|\mathbb{E}(f_n(x)) - f(x)| \leq Bh^\beta$; and

(b) (Variance) If $B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X$ is non-empty, then $\mathrm{Var}(f_n(x)) \leq \frac{V}{|B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X|}$.

These assumptions on the base estimator are natural. (E1) is somewhat restrictive but is satisfied by local polynomial estimators using a kernel K supported only on the interior of the unit ball that is scaled by a bandwidth h, and with a deterministic value for the case where there is no data in the h-ball around x. When the deterministic default value is 0, the local polynomial estimator satisfies (E2) trivially. Any other estimator can be made to satisfy this too by multiplication by the indicator function $\mathbf{1}_{|\mathcal{D}_X \cap B_{\mathcal{X}}(x,h)|>0}$. Condition (E3) is required to ensure that the derived estimators are also rate optimal and essentially says that the estimator f_n performs well when there is data in the h-ball and that the sampling regime ensures that this happens frequently. These assumptions are satisfied even for simple estimators such as the Nadaraya–Watson estimator when $\beta \leq 2$ in the context of our statistical problem, and more generally by local polynomial estimators for higher order β as shown in section 2.1.

Remark 24. *Assumption (E1) could be weakened to allow for sufficiently quickly decaying covariance which would allow for Gaussian kernels or k-nearest neighbour estimators. Such a weakening is offered with condition (E1'), which is satisfied for a Gaussian kernel in local polynomial estimation, or when the bandwidths are chosen adaptively. An interesting question would be whether such a condition also applies to kernel ridge regressors or neural networks.*

(E1') (Weak Locality) $\mathrm{Cov}(f_n(x), f_n(y) \mid \mathcal{D}_X) \leq \exp(c - an)$ *for some positive constants* $a, c \in \mathbb{R}_{>0}$.

Under these conditions, we know that the estimates $f_n(v_i)$ are independent, which gives:

$$\mathrm{Var}(f_n^{\mathrm{MAVE}}(x; [x]_f)) = m^{-1}\mathrm{Var}(f_n(v_i)) \tag{3.7}$$

but since $f(v_i) = f(x)$ by definition of the level set, we also know that

$$\begin{aligned}
\mathrm{Bias}(f_n^{\mathrm{MAVE}}(x; [x]_f)) &= \left| m^{-1} \sum_{i=1}^{m} f_n(v_i) - f(x) \right| && (3.8)\\
&\leq m^{-1} \sum_{i=1}^{m} \mathrm{Bias}(f_n(v_i)) && (3.9)\\
&\leq Bh^\beta && (3.10)
\end{aligned}$$

This reduction in variance means that we can take h much smaller than we could normally, which would be with a rate $h \propto n^{-1/(2\beta+d)}$. The following theorem gives one such option for faster bandwidths. Note that this is true for any choice of level set $[x]_f$ whether regular or critical.

Theorem 25. *Suppose n is large enough that the base estimator satisfies assumption set (E), and set*

$$h = an^{-\frac{1}{2\beta+d-\dim[x]_f}} \tag{3.11}$$

Then for any h_O which satisfies

$$h_O \geq a_O(\log n)^{-d} \tag{3.12}$$

for a chosen positive constant a_O the manifold averaged estimator satisfies:

$$\sup_{f\in\mathcal{F}} \mathbb{E}\big((f_n^{MAVE}(x;[x]_f) - f(x))^2\big) n^{\frac{2\beta}{2\beta+d-\dim[x]_f}} (\log n)^{\dim[x]_f - d} \leq C_1 \tag{3.13}$$

for a positive constant C_1.

Proof. We have the usual bias-variance decomposition of the MSE:

$$\mathbb{E}\big((f_n^{\text{MAVE}}(x;[x]_f) - f(x))^2\big) = \text{Bias}(f_n^{\text{MAVE}}(x;[x]_f))^2 + \text{Var}(f_n^{\text{MAVE}}(x;[x]_f)) \tag{3.14}$$

Now we can use the independence of the estimates $f_n(v_i)$ (which follows from the strict locality of f_n and the $2h$-spacing from the construction of the v_i) to bound the variance:

$$\begin{aligned}
\text{Var}(f_n^{\text{MAVE}}(x;[x]_f)) &= \text{Var}\Big(m^{-1}\sum_{i=1}^m f_n(v_i)\Big) && (3.15)\\
&= m^{-2}\text{Var}\Big(\sum_{i=1}^m f_n(v_i)\Big) && (3.16)\\
&= m^{-1}\sup_{v\in B(x,h_O)} \text{Var}(f_n(v)) && (3.17)\\
&\leq m^{-1}\frac{V}{nh^d} && (3.18)
\end{aligned}$$

As for the bias term, we have:

$$\begin{aligned}
\text{Bias}(f_n^{\text{MAVE}}(x;[x]_f)) &= \Big|m^{-1}\sum_{i=1}^m f_n(v_i) - f(x)\Big| && (3.19)\\
&\leq m^{-1}\sum_{i=1}^m \text{Bias}(f_n(v_i)) && (3.20)\\
&\leq Bh^\beta && (3.21)
\end{aligned}$$

For n large enough that $h_O \le r(x, [x]_f)$, we can then use the bound on m given in Proposition 31 to find:

$$\mathbb{E}\big((f_n^{\text{MAVE}}(x;[x]_f) - f(x))^2\big) \le B^2 h^{2\beta} + \frac{V}{nh^d}\min\Big(1, \big(\tfrac{h}{h_O}\big)^{\dim[x]_f}\Big) \tag{3.22}$$

$$\le B^2 h^{2\beta} + \frac{V}{nh^{d-\dim[x]_f} h_O^{\dim[x]_f}} \tag{3.23}$$

When $h = an^{-\frac{1}{2\beta+d-\dim[x]_f}}$ the first term decays at the rate $n^{-\frac{2\beta}{2\beta+d-\dim[x]_f}}$, and the second decays at the rate $n^{-\frac{2\beta}{2\beta+d-\dim[x]_f}} h_O^{-\dim[x]_f}$ for any choice of h_O. When $h_O \ge a_O(\log n)^{-d}$, and since these bounds are uniform in f, we get the desired result. □

3.3.3 Robustness to the Choice of Manifold

The previous result relies heavily on the fact that we pick points that are precisely on the level set $[x]_f$. In practice, we do not know this level set and so we now consider the performance of this estimator over other choices of manifold to estimate over. Suppose that M is any submanifold of $[0,1]^d$ and suppose that $\{v_i\}$ are points in M picked via the method in Section 3.3.1. We allow M to be a random variable so long as it is independent of the data used to estimate each $f_n(v_i)$. Then the bias of the MAVE has:

$$\big|\mathbb{E}(\tfrac{1}{m}\sum_{i=1}^m f_{n,h}(v_i) \mid \{v_i\}) - f(x)\big| = \big|\tfrac{1}{m}\sum_{i=1}^m (\mathbb{E}(f_{n,h}(v_i) \mid v_i) - f(v_i) + f(v_i) - f(x))\big| \tag{3.24}$$

$$\le \underbrace{\tfrac{1}{m}\sum_{i=1}^m \big|\mathbb{E}(f_{n,h}(v_i) \mid v_i) - f(v_i)\big|}_{\text{Average Bias of LPE}} + \underbrace{\tfrac{1}{m}\sum_{i=1}^m \big|f(v_i) - f(x)\big|}_{\text{Average Bias of Level Set Est.}} \tag{3.25}$$

If $M = [x]_f$, we would have the second term vanish and the results of the previous section would hold. As for the variance, we still have:

$$\operatorname{Var}\big(\tfrac{1}{m}\sum_{i=1}^m f_{n,h}(v_i) \mid \{v_i\}\big) = m^{-2}\sum_{i,j=1}^m \operatorname{Cov}\big(f_{n,h}(v_i), f_{n,h}(v_j) \mid \{v_i\}\big) \tag{3.26}$$

$$= m^{-2}\sum_{i=1}^m \operatorname{Var}\big(f_{n,h}(v_i) \mid \{v_i\}\big) \tag{3.27}$$

$$\le m^{-1}\sup_{x\in\mathcal{X}} \operatorname{Var}\big(f_{n,h}(x)\big) \tag{3.28}$$

because of the $2h$-spacing of v_i and v_j and the independence condition on the base estimator regardless of M—its only affect is through the number of points m. Now using the pointwise bounds on the bias and variance from assumption set (E), we see that the pointwise risk of

the MAVE is:

$$\mathbb{E}\Big(\big(\tfrac{1}{m}\sum_{i=1}^{m} f_{n,h}(v_i) - f(x)\big)^2\Big) = \mathbb{E}\Big(\mathbb{E}\big((\tfrac{1}{m}\sum_{i=1}^{m} f_{n,h}(v_i) - f(x))^2 \mid \{v_i\}_{i=1}^m\big)\Big) \tag{3.29}$$

$$= \mathbb{E}\Big(\mathrm{Bias}(\tfrac{1}{m}\sum_{i=1}^{m} f_{n,h}(v_i) \mid \{v_i\}_{i=1}^m)^2 + \mathrm{Var}(m^{-1}\sum_{i=1}^{m} f_{n,h}(v_i) \mid \{v_i\}_{i=1}^m)\Big) \tag{3.30}$$

$$\leq \mathbb{E}\Big(\Big(Bh^\beta + m^{-1}\sum_{i=1}^{m}\big|f(v_i) - f(x)\big|\Big)^2 + \tfrac{V}{nh^d m}\Big) \tag{3.31}$$

$$= 2B^2h^{2\beta} + \tfrac{V}{nh^d}\mathbb{E}(m^{-1}) + 2\mathbb{E}\Big(\Big(\tfrac{1}{m}\sum_{i=1}^{m}\big|f(v_i) - f(x)\big|\Big)^2\Big) \tag{3.32}$$

using the sum of squares inequality. Thus, understanding the pointwise risk requires control of the first two moments of the error term

$$Z = \tfrac{1}{m}\sum_{i=1}^{m}\big|f(v_i) - f(x)\big| \tag{3.33}$$

that depends on M. When $\beta \geq 1$ we can use the Lipschitz bound for each term in the sum, but this would be far from optimal. First replace x with the nearest point $u_i = \mathrm{arginf}_{u\in[x]_f} d(v_i, u)$ in the level set $[x]_f$, for which by definition $f(x) = f(u_i)$. This gives:

$$Z \leq \tfrac{1}{m}\sum_{i=1}^{m} L d(v_i, u_i) \leq L \sup_{v\in M\cap B(x,h_O)} d(v, [x]_f) \tag{3.34}$$

where $M \cap B(x, h_O)$ is the estimated level set we average over, and d_{h_O}. For notational convenience we define: $d_{x,h_O}(M, [x]_f) = \sup_{v\in M\cap\overline{B(x,h_O)}} d(v, [x]_f)$. See Figure 3.2 for a visualisation of this distance between these manifolds.

Therefore, if we use the assumptions set (E), we have the following bound on the MSE:

$$MSE(f_n^{MAVE}(x, M)) \leq 2Bh^{2\beta} + \tfrac{V}{nh^d}\mathbb{E}(m^{-1}) + 2L^2\mathbb{E}\big(d_{x,h_O}(M, [x]_f)^2\big) \tag{3.35}$$

If M is of the same dimension as $[x]_f$ (usually $d-1$ for regular level sets of sufficiently smooth f), then $\mathbb{E}(m^{-1})$ is usually unaffected by the choice of M; if $h_O \to 0$ and $h/h_O \to 0$ then $\mathbb{E}(m^{-1}) \asymp (h_O/h)^{\dim M}$ because $M \cap B(x, h_O)$ will converge its tangent space on this ball. Therefore, the robustness to the choice of M depends almost entirely on $\mathbb{E}\big(d_{x,h_O}(M, [x]_f)^2\big)$.

3.3.4 Increasing the Number of Averaging Points

So far we have ensured that the averaging points $\{v_i\}$ are $2h$-separated so that we can use the independence of the estimates $f_n(v_i)$. We now examine what happens if we reduce the spacing of these points so that we average over more of them. This of course introduces

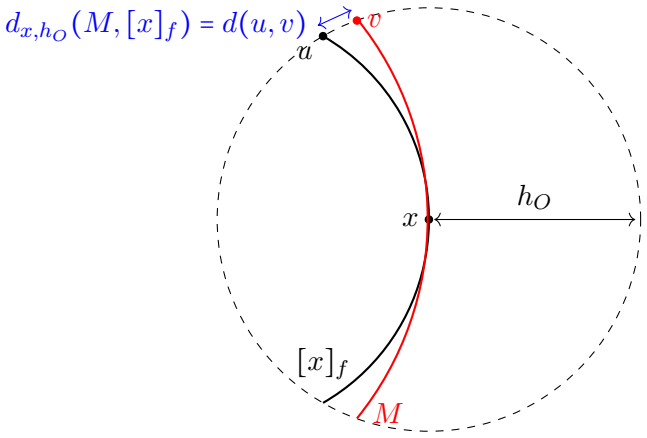


Fig. 3.2 A visualisation of the distance function $d_{x,h_O}(M,[x]_f)$ for a circular $[x]_f$ and elliptical M in red that meet at x. The distance between these manifolds is given by the distance $d(u,v)$ where the supremums are achieved, shown in blue.

correlation that we have to quantify.

Suppose that $\{v_i\}$ is now constructed with η spacing, but the independence radius of assumption (E2) is still $2h$. Then

$$\mathrm{Var}\big(m^{-1}\sum_{i=1}^{m} f_n(v_i)\big) = m^{-2}\sum_{i,j=1}^{m}\mathrm{Cov}(f_n(v_i),f_n(v_j)) \le m^{-2}\sum_{i,j=1}^{m}\rho_{ij}V' \tag{3.36}$$

using the Cauchy-Schwarz inequality and the bound $\mathrm{Var}(f_n(v)) \le \sup_{v\in\mathcal{X}} V/|B(v,h)\cap\mathcal{D}_X| = V'$ from assumption (E3) (note that all such balls will contain $O(nh^d) = O(n^{2\beta/(2\beta+1)})$ points with a rectangular design regime when $\beta \ge d$). Here ρ_{ij} is the correlation between $f_n(v_i)$ and $f_n(v_j)$, which is clearly bounded above by 1 for all i and j, and is equal to 0 if $d(v_i,v_j) > 2h$. Thus, we can bound:

$$m^{-2}\sum_{i,j=1}^{m}\rho_{ij} \le m^{-1}\sum_{i=1}^{m} m^{-1}\sum_{j=1}^{m}\mathbf{1}_{d(v_i,v_j)\le 2h} = m^{-1}\sum_{i=1}^{m}|\{v_j : v_j \in B(v_i,2h)\}| \tag{3.37}$$

As $\eta \to 0$, the number of points v_j in the ball $B(v_i,2h)$ is bounded above by the number of points a_j in the ball $B(a_i,2h)$, which has:

$$|\{a_j : a_j \in B(a_i,2h)\}| \to \frac{\mathrm{Vol}_{T_{x[x]_f}}(B(a_i,2h)\cap B(x,h_O))}{\mathrm{Vol}_{T_{x[x]_f}}(B(x,h_O))} \tag{3.38}$$

where the volume is the volume of the intersection with tangent space (measured with respect to Lebesgue measure on this linear tangent space). This gives the bound:

$$m^{-2}\sum_{i,j=1}^{m}\rho_{ij} \leq c_0 m^{-1}\sum_{i=1}^{m}\frac{\mathrm{Vol}_{[x]_f}(B(v_i,2h)\cap B(x,h_O))}{\mathrm{Vol}_{[x]_f}(B(x,h_O))} \leq c_1 m^{-1}\sum_{i=1}^{m}\left(\frac{h}{h_O}\right)^{\dim[x]_f} \leq c_1\left(\frac{h}{h_O}\right)^{\dim[x]_f} \tag{3.39}$$

for positive constants c_0 and c_1. This bound means that we can take the spacing $\eta \to 0$ and maintain the same rates as Theorem 25: the variance term will still decay at the rate of $n^{-1}h^{d-\dim[x]_f}h_O^{\dim[x]_f}$ even if the averaging points are not $2h$ spaced. This result says that the locality of f_n allows for dimension reduction as long as the averaging points are sufficiently evenly spaced out.

3.4 Estimation of the Level Sets

We now seek to answer the question: under what conditions can we estimate $[x]_f$ fast enough to ensure that $MSE(f_n^{MAVE}(x,\widehat{[x]_f}))$ is faster than the minimax rate for the Hölder class $\mathcal{F}(L,\beta)$? We can answer this question for a particularly simple plug in estimator of $[x]_f$, and achieve the rates of the oracle that knows $[x]_f$ up to a poly-logarithmic factor. This does rely on a particularly strong condition, but one that is natural for the Hölder class $\mathcal{F}(L,\beta)$.

3.4.1 Local Polynomial Level Set Estimation

Perhaps the simplest estimator of $[x]_f$ is the plug in estimator $[x]_{f_n}$ for an estimator f_n of f. In our case we will use a local polynomial estimator with bandwidth h_O, denoted p_n to distinguish it from the estimator f_n used in the averaging. Suppose that f agrees with a polynomial $p(v) = U_{x,h_O}(v)\theta_{p,x,h_O}$ on a ball $B(x,H)$ for some positive H, i.e. $f(v) = p(v)$ for all $v \in B(x,H)$. By definition, we know that if $h_O < H$ then for all $v \in B(x,h_O)$:

$$|p_n(v) - p(v)| = |U_{x,h_O}(v)(\theta_{p,x,h_O} - \hat{\theta}_{p,x,h_O})| \leq \sqrt{\ell}\|\theta_{p,x,h_O} - \hat{\theta}_{p,x,h_O}\|_2 \tag{3.40}$$

Set $u_i = \arg\inf_{u\in[x]_p\cap B(x,h_O)} d(v_i,u)$. Then

$$|p(v_i) - p(u_i)| \leq |p(v_i) - p_n(v_i)| + |p_n(v_i) - p(u_i)| \tag{3.41}$$

$$\leq |p(v_i) - p_n(v_i)| + |p_n(x) - p(x)| \tag{3.42}$$

$$\leq 2\sqrt{\ell}\|\theta_{p,x,h_O} - \hat{\theta}_{p,x,h_O}\|_2 \tag{3.43}$$

because $v_i \in [x]_{p_n}$ and $u_i \in [x]_p$. Now since $v_i - u_i$ is orthogonal to the tangent space of $[x]_p$ at u_i (a necessary condition to be a minimum of the distance), and assuming that the level

set is regular, we know that $\frac{\partial p}{\partial x}\big|_{u_i}(v_i - u_i)$ is non-zero, and that:

$$|p(v_i) - p(u_i)| \approx \left|\frac{\partial p}{\partial x}\big|_{u_i}(v_i - u_i)\right| \times d(v_i, u_i) \tag{3.44}$$

by considering the first order Taylor expansion of p around u_i. As $\|v_i - u_i\|_2 \to 0$ this approximation will be the dominant term, and so we must have $d(v_i, u_i) \leq C|p(v_i) - p(u_i)| \leq C\|\beta_{x,h_O} - \hat{\beta}_{x,h_O}\|_2$ for some positive constant C (that will depend on f and p) for sufficiently small distances. Therefore we have:

$$\mathbb{E}\big(d_{x,h_O}([x]_{p_n}, [x]_f)^2\big) = \mathbb{E}\big(d_{x,h_O}([x]_{p_n}, [x]_p)^2\big) \leq C^2\mathbb{E}\big(\|\theta_{p,x,h_O} - \hat{\theta}_{p,x,h_O}\|_2\big) \leq \frac{C_0}{nh_O^d} \tag{3.45}$$

using the result of convergence of the polynomial coefficients in section 2.1. Thus, we have the following result for the pointwise mean squared error of the MAVE using this estimate of the level set $[x]_f$.

Theorem 26. *Suppose that f agrees with a polynomial p on some ball $B(x, H)$ for some positive H. Suppose that $x \in (0,1)^d$ is a point on a regular level set of f. Let f_n be a local polynomial estimator of degree $k = \lceil\beta\rceil - 1$ that uses a rectangular kernel and the bandwidth $h = an^{-\frac{2\beta}{2\beta+1}}$, and let p_n be a local polynomial estimator of degree k and that uses the bandwidth $h_O = a_O(\log n)^{-\alpha}$ for positive a, a_O, α. Suppose that f_n and p_n are independent, using iid copies of $\mathcal{D}$. Then for sufficiently large n such that assumption set (E) holds and than $h_O < H$ we have:*

$$MSE(f_n^{MAVE}(x, [x]_{p_n})) \leq C_* n^{-\frac{2\beta}{2\beta+1}}(\log n)^{(d-1)/\alpha} \tag{3.46}$$

for a positive constant C_*.

Proof. We know from the previous analysis that:

$$MSE(f_n^{MAVE}(x, [x]_{p_n})) \leq 2Bh^{2\beta} + \tfrac{V}{nh^d}\mathbb{E}(m^{-1}) + 2L^2\mathbb{E}\big(d_{x,h_O}([x]_{p_n}, [x]_f)^2\big) \tag{3.47}$$

Recall that $r(x, M)$ is the shadow radius of M at the points x. From Proposition 31, we know:

$$m = m(x, h, M) \geq \max\Big(1, \big(\tfrac{r(x,[x]_{p_n})}{2h}\big)^{\dim[x]_{p_n}}\Big) \tag{3.48}$$

Thus, we have the upper bound:

$$\mathbb{E}(m^{-1}) \leq \mathbb{E}\Big(\min\Big(1, \big(\tfrac{2h}{r(x,[x]_{p_n})}\big)^{\dim[x]_{p_n}}\Big)\Big) \tag{3.49}$$

Since $d_{x,h_O}([x]_{p_n}, [x]_p) \to 0$, $h/h_O \to 0$, and $h_O \to 0$, we know that $r(x, M)$ will eventually scale with h_O exactly and the minimum will be the second term. Thus we would have:

$$\mathbb{E}(m^{-1}) \leq \left(\tfrac{2h}{h_O}\right)^{\dim[x]_p} \tag{3.50}$$

Since $[x]_p$ is a regular level set of an infinitely differentiable function, we know that it has dimension $d-1$ by the regular level set theorem. This, plus the bound on $\mathbb{E}\big(d_{x,h_O}([x]_{p_n}, [x]_f)^2\big) \leq \frac{C_0}{nh_O^d}$ gives:

$$MSE(f_n^{MAVE}(x, [x]_{p_n})) \leq C_1 h^{2\beta} + \tfrac{C_2}{nhh_O^{d-1}} + \frac{C_3}{nh_O^d} \tag{3.51}$$

$$\leq C_1 n^{-\frac{2\beta}{2\beta+1}} + C_2 n^{-\frac{2\beta}{2\beta+1}} (\log n)^{(d-1)/\alpha} + C_3 n^{-1} (\log n)^{d/\alpha} \tag{3.52}$$

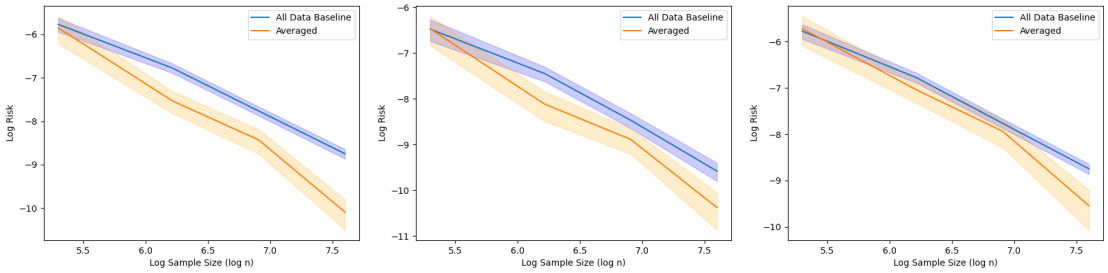
The middle term will dominate the rate, giving the result. □

The condition that f agrees with a polynomial is quite strong. If we knew that this were true, then we would simply use a polynomial regressor with the data in $B_{\mathcal{X}}(x, H)$ and achieve the optimal parametric rate of n^{-1}. This result says that we will come much closer to this rate using the MAVE with a plug in estimator of the level set, relative to a minimax optimal nonparametric LPE which would only converge at the rate $n^{-2\beta/(2\beta+d)}$. Methods that improve on this by adaption to the true maximal smoothness of f will eventually take bandwidths with the same behaviour and achieve nearly parametric rates. It is interesting to see that focusing on the dimension via the level sets of f leads to nearly the same rates in this case. This assumption is natural for the Hölder class—these are functions that look locally very much like their Taylor series of degree k. We will examine the robustness of this estimator to violations of this assumption in the numerical experiments.

3.5 Numerical Experiments

We will consider the same setup for three examples of functions all within the Hölder class $\mathcal{F}(L, 3)$. This allows us to examine the pointswise MSE under the conditions of Theorem 26 as well as a mild violation and a major violation of these assumptions. We examine the performance of the MAVE using the optimal choice of bandwidth against a baseline LPE with optimal choice of bandwidth. All fitted polynomials are degree 2 as determined by the class. We generated data from the model:

$$Y_i = f_j(X_i) + \epsilon_i \tag{3.53}$$



(a) MSE for f_1 (b) MSE for f_2 (c) MSE for f_3

Fig. 3.3 The pointwise Mean Squared Error (MSE) for the estimator $f_n^{MAVE}(x, [x]_{p_n})$ at the point $x_0 = (0.5, 0.5)$.

with X_i chosen on the rectangular grid and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. The three functions examined are $f_j : \mathbb{R}^2 \to \mathbb{R}$:

$$f_1(x) = \sin\left(2\pi \|x\|_2\right) \tag{3.54}$$

$$f_2(x) = \sin\left(2\pi \sqrt{(x_1^2 + 0.8x_2^2)}\right) \tag{3.55}$$

$$f_3(x) = \sin\left(2\pi \|x\|_2\right) + 0.4\cos(x_1) \tag{3.56}$$

These exhibit differing behaviours: f_1 has level sets that are described by circles; f_2 also has varietal level sets but now they are elliptical; and f_3 has level sets that deviate from this structure. The performance is judged by the MSE at the point $x_0 = (0.5, 0.5)$, i.e.:

$$R_i(f_n^{MAVE}, x_O) = \mathbb{E}_{f_i}\left((f_n^{MAVE}(x_0, [x_0]_{p_n}) - f_i(x_0))^2\right) \tag{3.57}$$

where the subscript f_i corresponds to data generated using this regression function in the non-parametric regression model. The sample sizes used range are $n = 200, 500, 1000, 2000$. The noise standard deviation is $\sigma = 0.1$. The MSE at x_0 for these estimators are plotted in Figure 3.3, on a log / log scale. We see that the MAVE using the estimated level set is nearly always an improvement on the baseline. It performs best when the level sets are most easily estimated.

3.6 Discussion

This method opens up new ways to reduce the dimension of the regression problem in many nonparametric contexts. It can achieve dimension reduction for sparse functions as well as for many functions with non-linear level sets. The key now is to find the optimal way to estimate $[x]_f$ with respect to the loss function $d_{x,h_O}([x]_f, \hat{A})$. The plug-in estimator could be improved in many ways, one of which we suggest below. This problem is likely very difficult

in full generality because derivatives are statistically harder to estimate that the original function[4], so we can instead focus on specific hypothesis classes that impose more structure on $[x]_f$ explicitly rather than on f (as we have done in this chapter). Some natural classes to consider are:

- functions with level sets that are described by the zero-set of a polynomial equation, i.e. $[x]_f = \{u \in \mathcal{X} : p(u) = 0\}$ for some polynomial p;
- functions with level sets that have bounded curvature;
- functions with level sets that described by the orbits of a group G acting on $\mathcal{X}$;

This last example is the focus of the rest of this thesis. In the following chapters we will show that this class is simple enough to estimate the true level sets using a straight forward M-estimator, and that this leads to adaptive dimension reduction.

It is important to note that this idea cannot lead to an improvement in the rate of estimation uniformly over our functional hypothesis class; it cannot contradict the well established minimax rates. The analysis presented in this chapter avoids this because it only applies pointwise to regular points of the regression function—if we try to estimate at an isolated critical point then the level set is 0 dimensional and we have no improvement in the rate. This means that the worst case risk of the MAVE, i.e., $\sup_{f\in\mathcal{F}} MSE(f_n^{MAVE}(x,[x]_f))$, is still on the order of $n^{-\frac{2\beta}{2\beta+d}}$ (which is why the rates depend on f in the statement of Theorem 25). Even in these cases the process of estimating level sets and using the MAVE on these level sets is a novel method of re-balancing the bias-variance tradeoff in nonparametric regression. In the same way that not all functions are additive but additive models are still useful for generic nonparametric regression, level set based models can be useful in general even when the level sets of the regression function are not precisely in a hypothesis class of level sets considered.

This method also demonstrates an interesting relationship between the smoothness of the functional class and the dimension of its domain: having enough derivatives allows for the dimension reduction to work. As mentioned in section 3 we have so far considered only classes with many derivatives, but it also holds for some other classes with minor adaption. The main roadblocks are:

1) we lose the guarantee of regular level sets; and

2) the number of datapoints in the balls $B(v_i, h)$ can decrease with n rather than increase.

[4] See Stone [1982], the minimax rate of estimating the k^{th} derivative of a function in the Hölder class is $n^{-\frac{2(\beta-k)}{2\beta+d}}$.

The first is only a minor issue—we would only consider removing the dimensions of the level set rather than assuming that there are $d-1$ of them. This would make the search for unknown level sets computationally harder but not statistically harder as there are only finitely many possible dimensions to consider. This is part of the analysis required in the following chapter. The second is more of a problem: if $\beta \leq \dim[x]_f/2$ and $h \asymp n^{-1/(2\beta+d-\dim[x]_f)}$ then $nh^d \to 0$, and so the assumptions in set (E) would fail for most standard estimators and most design regimes. This can be solved for specific estimators such as a local constant estimator when $\beta = 1$ by averaging only over the estimates $f_n(v_i)$ for which $B(v_i, h)$ contains at least one datapoint X_j. For most design regimes (e.g., rectangular grids or $U([0,1]^d)$) we would then still expect that we are averaging over $mnh^d \asymp n^{-2\beta/(2\beta+d-\dim[x]_f)}$ datapoints leading to the fast rate for this case even when $\beta \leq \dim[x]_f$.

Chapter 4

M-Estimation of Maximal Symmetries

This chapter is based on the paper Christie and Aston [2024a]. We consider structured classes of level sets where they all obey the same global structure: $[x]_f = \{g \cdot u : g \in G, u \in U\}$ for some closed set $U \subseteq \mathcal{X}$. The class of level set structures is thus indexed by the groups G considered, and so the level set estimation problem is one of *symmetry estimation*. We show that estimation in this class is fast enough to achieve adaptive dimension reduction under much weaker assumptions on f than the previous chapter.

Fully generic estimation of level sets as in the previous chapter may be too slow to achieve dimension adaptive estimation, so we now focus in on specific class of level sets where estimation is sufficiently fast. This is exactly what is done with methods that utilise covariate sparsity. If f depends only on some subset of the covariates then it must have a level set that contains the hyperplane spanned by the remaining basis vectors. The dimensions of this hyperplane are the dimensions removed from the rate of estimation of f. However this is far from the broadest class of possible level sets. In particular the MAVE allows us to consider non-linear level sets, so in this chapter we generalise the notion of covariate sparsity to that of symmetry. The level sets will now contain the orbits of group actions, and the class of possible level sets will be the set of subgroups of a search group $\mathcal{G}$. Thus the level set estimation becomes one of symmetry estimation.

4.1 Introduction

Symmetries are ubiquitous across statistics in fields ranging from time series obeying seasonality to rotation invariant protein volume estimation [Jiang and Tang, 2017]. In regression, symmetry is taking hold across literature in neural networks [Kondor and Trivedi, 2018, Bronstein et al., 2021], and generalises several structures that are well studied such as covariate

sparsity and multi-index models (see Examples 27 and 28 below). Lie groups were initially conceived to study the properties of differential equations [Hawkins, 2012]; now in statistics we have an opportunity to use these mathematical tools to better understand problems in regression and beyond.

The key idea is that if G is a set of differentiable bijections $\phi : \mathcal{X} \to \mathcal{X}$ of the d-dimensional covariate space $\mathcal{X}$, and the regression function $f : \mathcal{X} \to \mathbb{R}$ is G-invariant, i.e., it obeys the invariance rule $f(x) = f(\phi(x))$ for all $\phi \in G$ and $x \in \mathcal{X}$, then we can use estimates of each $f(\phi(x))$ for $\phi \in G$ to construct a lower variance estimator of $f(x)$ without introducing bias. We call the set $\{\phi(x) : \phi \in G\}$ the **orbit** of x under G, written $[x]_G$, and when this set is smooth and has positive dimension[1] d^G we might hope that the new estimate has a worst case risk decay with d replaced by $d - d^G$. This is a special case of the structure in the previous section, as the orbits are contained in the level sets of any G-invariant f.

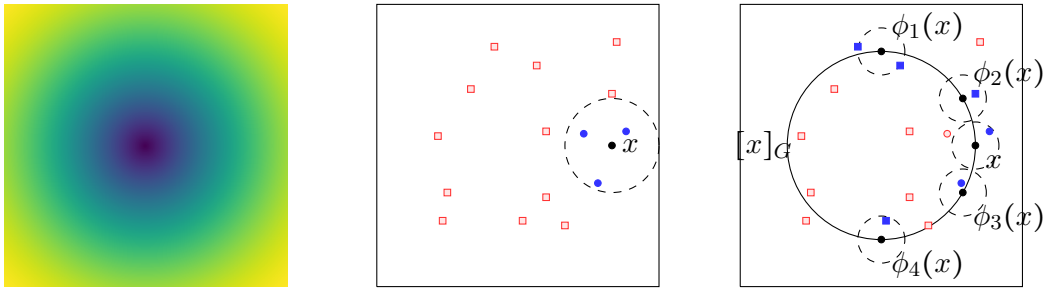
To construct an estimator with the rate corresponding to the lower dimensional space, we apply a *symmetrisation operator* to any existing estimator f_n satisfying some optimality conditions akin to local polynomial regressors. These operators average f_n over the orbits $[x]_G$ using some chosen distribution on the set G. For example, $\tilde{f}_n(x) = m^{-1} \sum_{i=1}^m f_n(\phi_i(x))$ is an estimator that averages using the uniform distribution on a set of transformations $\{\phi_i\}_{i=1}^m \subseteq G$ for a chosen m. This set could be chosen deterministically, as in Section 4.2.1 (a slightly modified MAVE), or could be sampled uniformly from G (or a compact neighbourhood of $e \in G$) as in Section 4.2.4.

In an ideal world, where each estimate $f_n(\phi_i(x))$ is guaranteed to be independent, this would cut the variance by a factor of m. When f is G-invariant, it does so without introducing bias. This is the situation in Figure 4.1. The first results of this chapter show that this is true even when dependence considerations are taken into account, and quantifies the asymptotic bias introduced when f is not G-invariant. This allows us to re-optimise the bias-variance trade-off and achieve faster convergence to f, if it is G-invariant.

However, if the symmetry G is not known, we must instead seek to estimate it from the data. Specifically, if G is a large set of possible transformations, then can we estimate the maximal subset $G_0 \subseteq G$ for which f is G_0-invariant? If G_0 can be estimated sufficiently quickly, then an estimate that uses this symmetry could adapt to the symmetry of f, and provide rates of the form:

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}(\|f_{n,\hat{G}} - f\|_2^2) n^{\frac{2\beta}{2\beta+d-d_{G_0}}} \leq C \tag{4.1}$$

[1] the principal orbit theorem ensures that almost all orbits $[x]_G$ have the same dimension in many of the cases we consider, see Section 2.3.



(a) Regression Function

(b) Local Polynomial Estimation

(c) Symmetrised LPE

Fig. 4.1 Sub-figure (a) shows a heatmap of a regression function $f : \mathbb{R}^2 \to \mathbb{R}$. Sub-figure (b) depicts the dashed closed ball of radius h around the point $x \in \mathcal{X}$ and locations of some covariates, the data inside which (filled blue circles) are used to estimate $f(x)$ with a local polynomial estimator f_n with a kernel with support contained in this ball. The data outside this ball (red squares) are not used to estimate $f(x)$. Sub-figure (c) depicts our estimator which symmetrises with respect to $G = \{\text{Rotations around } 0\}$, using multiple balls of smaller radius with centres $\phi_i(x)$ spaced around the orbit $[x]_G$, chosen using the method descried in Section 4.2.1. Note that these need not be evenly spaced, nor cover the entire orbit. The symmetrised estimator uses more of the data points to estimate $f(x)$, and excludes one point that was previously used, shown by the colour changes.

where $f_{n,\hat{G}}$ is an estimator symmetrised with the estimated transformations. This situation is the focus of this chapter: we give a (two-step) M-estimator[2] of G_0 and show that using this estimated symmetry can achieve adaptive rates of the form of equation (4.1).

This is done through understanding the set G as a *group*, the mathematical objects used to describe symmetries. We can do this because if f is G-invariant in the sense above then f is also invariant to the sets of inverses $\{\phi^{-1} : \phi \in G\}$ and compositions $\{\phi \circ \psi : \phi, \psi \in G\}$. This description is used to reduce the search space from the set of subsets of G to the much smaller set of closed, connected subgroups of G, and to keep the language in this chapter consistent with the machine learning and mathematical literature. We have aimed to keep the level of group theory background required to a minimum in this chapter, introducing any concepts needed. Most of the proofs rely only on elementary probability theory and linear analysis, and on the geometric properties of the orbits. We now demonstrate how this generalises many usual structures in nonparametric regression across several examples.

Example 27 (Covariate Sparsity)**.** *Suppose that $\mathcal{X} = \mathbb{R}^p$ and let $\mathcal{G}$ be the set of translations $\{\phi_g : x \mapsto x + g\}$. Suppose that the regression function $f : \mathcal{X} \to \mathbb{R}$ is s-sparse, i.e., there are $(p - s)$ coordinate positions i that f does not depend on, so $f(x + ae_i) = f(x)$ for all $a \in \mathbb{R}$ and*

[2] In the sense of Van der Vaart [2000], chapter 5.

standard basis vectors e_i for these positions i. This is precisely the definition of invariance to the subgroups of $\mathcal{G}$ given by the $\mathbb{R}$-span of ϕ_{e_i}.

Example 28 (Multi-Index Models)**.** *In the context of the previous example, there are of course many other subgroups of $\mathcal{G}$ that f could be invariant to, for example translations by the hyperplane orthogonal to the vector $\mathbf{1}_d = (1, \ldots, 1)$. A function invariant to this subgroup must be of the form $f(x_1, \ldots, x_d) = f_A(x_1 + \cdots + x_d)$, as the function's value at off diagonal inputs are completely determined by value at the projection to the line through $\mathbf{1}_d$. Thus functions that are invariant to such subgroups are examples of the multi-index structure $f(x) = h(Tx)$ for a orthogonal projection $T : \mathbb{R}^d \to \mathbb{R}^m$ as described in Hristache et al. [2001], in this case with $T = \mathbf{1}_d^T$.*

Example 29 (Compact Sparsity)**.** *The previous example is not an example of a **compact** group action, but is very related to this example where we consider the compact domain $[0,1]^d$ rather than $\mathbb{R}^d$. Now consider the action of the d-torus $\mathbb{T}^d = \{\phi_g : g \in [0,1)^d\}$ given by translations modulo the integer grid:*

$$\phi_g(x) = (g_1 + x_1 - \lfloor g_1 + x_1 \rfloor, \ldots, g_d + x_d - \lfloor g_d + x_d \rfloor) \tag{4.2}$$

for all $x \in [0,1)^d$, and $\phi_g(x) = x$ for all other x. In this case continuous functions that are invariant to the span of ϕ_{e_i} are also sparse, in the sense that they do not depend on the variable in the i^{th} position. The only difference to the previous examples is now the group $\mathbb{T}^d$ is compact (note that the topology on this set identifies the boundaries, so it is homeomorphic to a product of two circles), which can make the search for the maximal subgroup computationally easier.

Example 30 (3D Rotations)**.** *Suppose that $\mathcal{X} = \overline{B_{\mathbb{R}^3}(0,1)}$, the closed unit ball in $\mathbb{R}^3$. Let $\mathcal{G} = SO(3) = \{$rotations around 0$\}$. Any function f that is invariant to $\mathcal{G}$ depends only on $\|x\|_2$, such a function is not covariate sparse but does only depend on a lower dimensional projection—a non-linear one in this case. If f is invariant to the subgroup of rotations around the x axis, denoted S_x^1, then f depends only on x_1 and $\sqrt{x_2^2 + x_3^2}$, a two dimensional non-linear projection. This also generalises to higher dimensions, where the groups $SO(d)$ refer to rotations around the origin in d-dimensions.*

4.1.1 Main Contributions

We now state the results in this chapter in technical language. All terms with definitions contained in the background material are italicised here. Suppose that $\mathcal{X}$ is a closed subset of $\mathbb{R}^d$ with non-trivial Lebesgue volume, or any d-dimensional *manifold*, and let $\mathcal{G}$ be a locally compact *Lie group acting smoothly, faithfully*, and *properly* on $\mathcal{X}$. Let $\mathcal{F}(L, \beta)$ be the *Hölder class* of functions on $\mathcal{X}$. Suppose that we have collected independent and identically

distributed (iid) data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ with $Y_i = f(X_i) + \epsilon_i$ where $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$, and with X_i and ϵ_i independent. Suppose that f_n is a rate optimal estimator over the hypothesis class $\mathcal{F}(L, \beta)$, such as a local polynomial regressor of degree $\ell = \lceil \beta \rceil - 1$. Let $S_\rho : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ be the *partial symmetrisation operator* given by the orbit average $S_\rho f = \mathbb{E}(f(g \cdot x))$ when $g \sim \rho$ for any distribution ρ on the group $\mathcal{G}$. For each $f \in \mathcal{F}$, let $G_{\max}(f, \mathcal{G})$ be the unique *maximal invariant subgroup* of f. In this chapter:

1) we construct **Partially Symmetrised Estimators** $S_{\rho(G)} f_n$ using specific deterministic discrete distributions $\rho(G)$ on each closed subgroup $G \leq \mathcal{G}$ for which we obtain pointwise bounds for all $x \in \mathcal{X}$:

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}\big((S_{\rho(G)} f_n(x) - S_{\rho(G)} f(x))^2\big) \leq C_{G,x} n^{-\frac{2\beta}{2\beta + d - d^G}} \tag{4.3}$$

   for positive constants $C_{G,x}$ that do not depend on f or n, and where d^G is the dimension of a principal orbit of G;

2) we show that when $\Omega \subseteq \mathcal{X}$ is compact and μ-measurable we also obtain the integrated risk bound

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}\big((S_{\rho(G)} f_n(X) - S_{\rho(G)} f(X))^2 \mid X \in \Omega\big) \leq C_{G,\Omega} n^{-\frac{2\beta}{2\beta + d - d^G}} \tag{4.4}$$

   where $C_{G,\Omega} = \mathbb{E}(C_{G,X} \mid X \in \Omega) < \infty$ for $X \sim \mu$;

3) we show that equation (4.3) holds when $\rho(G)$ is replaced by the uniform distribution $U(G)$ for compact groups G, and with a the Monte-Carlo estimate with $\{g_i\}_{i=1}^n$ sampled uniformly from the group, when μ is G-invariant. This means in practice we do not need the explicit distributions ρ_G for these groups. As a special case, when f is G-invariant this recovers the rates of Tahmasebi and Jegelka [2023] as a corollary, though in a slightly different context;

4) we develop a metric structure on the set of closed subgroups $K(\mathcal{G}) = \{\overline{G} : G \leq \mathcal{G}\}$, and show that the asymptotic bias of the partially symmetrised estimators $S_{\rho(G)} f_n$ is locally Hölder continuous in this metric, when the distributions $\rho(G)$ each has compact support contained in G;

5) we show that when $\Omega \subseteq \mathcal{X}$ is compact closure of an open subset we can compute an **Error Minimising Symmetry** $\hat{G}$, an estimator of $G_{\max}(f, \mathcal{G})$, and that the **Best Symmetric Estimator** $S_{\rho(\hat{G})} f_n$ achieves the adaption property:

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}\big((S_{\rho(\hat{G})} f_n(X) - f(X))^2 \mid X \in \Omega\big) n^{\frac{2\beta}{2\beta + d - d^{G_{\max}(f,\mathcal{G})}}} \leq \sup_{G \in K(\mathcal{G})} 2C_{G,\Omega} + 2 + \frac{2\beta(2C_{G,\Omega} + 4L)}{\mu(\Omega)(2\beta + d)} \tag{4.5}$$

achieving the goal described in equation (4.1). This requires that the action of $\mathcal{G}$ is Lipschitz with respect to the chosen distance on the group.

Each result here assumes mild and standard conditions on the base estimator f_n and on the group action, which are given explicitly in the assumption sets (E) and (Q) later. The key assumptions of the problem are that the distribution μ is not concentrated around the singularities of the action, that β is at least $\dim[x]_{\mathcal{G}}/2$, and that we have a baseline estimator f_n that is sufficiently local, in the same way local polynomial estimators using kernels with compact support are, and minimax optimal for the class $\mathcal{F}(L,\beta)$.

4.2 Estimation using a Particular Symmetry

We now turn to the statistical problem of estimating f from within the Hölder function class $\mathcal{F}(L,\beta) \subseteq L^2(\mathcal{X})$ using an iid sample $\{(X_i,Y_i)\}_{i=1}^n$ with $\mathbb{E}(Y_i \mid X_i) = f(X_i)$ and $\mathrm{Var}(Y_i \mid X_i) = \sigma^2 < \infty$. This will be done using the same idea as the MAVE from the previous chapter, but the results here are extended to the random design regime and to integrated risk using the additional structure of the orbital level sets.

Let G be any locally compact Lie group acting smoothly, properly, and faithfully on $\mathcal{X}$ and let $U \subseteq G$ be a compact set with nonempty interior containing the identity. Suppose that f_n is an estimator for f that is minimax optimal for the class $\mathcal{F}(L,\beta)$, that we will call the **base estimator**. We will first, for each point $x \in \mathcal{X}$, construct a discrete distribution ρ_m^x with support contained in U and consider the *Partially Symmetric Estimator* $S_{\rho_m^x} f_n(x)$. We will establish point-wise bounds:

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}\big((S_{\rho_m^x} f_n(x) - S_{\rho_m^x} f(x))^2\big) \leq C_{G,x} \phi_n^{\beta, d-d^G} \tag{4.6}$$

where $\phi_n^{\beta,d-d^G} = n^{-\frac{2\beta}{2\beta+d-d^G}}$ and $d^G = \dim[x]_G$ for any principal orbit $[x]_G$. We will then show that these can be integrated under some conditions on the group action. Note that this estimator will not converge to f unless f is G-invariant so these are not always true risk bounds, but they are useful for understanding the convergence of these estimators to their asymptotic bias.

For G-invariant f, this is perhaps the expected rate that would be achieved by projecting the covariates to the quotient space $\mathcal{X}/G$ and regressing the response against these orbits (of course, proving this would require a mild treatment of the topological conditions of the quotient space $\mathcal{X}/G$ and then applications of standard theorems). For the non G-invariant f, it is non-obvious that we can apply usual results because we do not observe iid data with conditional mean $S_{\rho_m^x} f$ - the distribution is very much affected by the averaging over the orbits.

4.2.1 Defining Partially Symmetrised Estimators

In this section, we give an explicit construction for Partially[3] Symmetrised Estimators, as described above. The main choice to make is the distribution ρ to be used to partially symmetrise the base estimator f_n. We will do this by constructing a set of m points $g^i \in G$ and take $\rho = U(\{g^i\})$. This must balance two goals: we want as many points as possible around $[x]_G$ in order to maximise the variance reduction, but would like the estimators $f_n(g^i \cdot x)$ to be independent for theoretical tractability. In order to achieve this, we pick the a large number of points m such that we can guarantee that $B_{\mathcal{X}}(g^i \cdot x, h)$ are pairwise disjoint for all i, j for a chosen bandwidth $h > 0$. This is done by examining the geometry of the orbit $[x]_G$, which is a smoothly embedded sub-manifold of $\mathcal{X}$ under our conditions on the group action. These g^i depend on x and h, so we add subscripts to remind us of this choice.

To construct the set $\{g^i_{x,h}\}$, first isometrically embed $\mathcal{X}$ in $\mathbb{R}^q$ for some $q \in \mathbb{N}$, and consider all points in $\mathcal{X}$ and $T_x[x]_G$ via this embedding. Since $U \subseteq G$ is a compact neighbourhood of the identity, $[x]_U = \{g \cdot x \in \mathcal{X} : g \in U\}$ is a compact sub-manifold of $[x]_G$ of the same dimension. Let $W_x^{[x]_U}$ be the largest hypercube in $T_x[x]_U$ that is contained in the orthogonal projection of $[x]_U$ onto $T_x[x]_U$. Let $R_x^{[x]_U}$ be the side length of the hypercube $W_x^{[x]_U}$. If the orbit is 0-dimensional, as with the trivial group $I = \{e\}$, then we take $R_x^{[x]_U} = 1$. We construct the group elements $g^i_{x,h}$ as follows:

(i) Pick a maximal grid $\{a_i\}_{i=1}^m$ in $W_x^{[x]_U}$ such that $d(a_i, a_j) \geq 2h$ whenever $i \neq j$;

(ii) Orthogonally (with respect to $T_x[x]_U$) project this grid onto $[x]_U$, giving a set $\{u_i\}_{i=1}^m \subseteq [x]_U$; and

(iii) Pick any[4] $g^i_{x,h} \in U$ such that $g^i_{x,h} \cdot x = u_i$.

This allows us to define the symmetrising distribution ρ_m^x as the uniform distribution on the set $\{g^i_{x,h}\}$, which has size m (which depends on x, h, and U implicitly). This process is depicted in Figure 4.2 for an example with G the special orthogonal group of 2D rotations, $SO(2)$. We use this choice of ρ to construct the partial symmetrisation operator $S_{\rho_m^x}$, and apply this to the chosen base estimator f_n to construct the **Partially Symmetrised Estimator** $S_{\rho_m} f_n$. This is only a slight modification of the MAVE from the previous chapter, where the level set $[x]_f$ is replaced with the orbit $[x]_G$.

The first step to analysing the partially symmetrised estimator requires us to prove the following lower bound on m and confirm that the points are sufficiently spaced around $[x]_U$.

[3]The name *partial* comes from the fact that $S_{\rho_m^x}$ may not project f_n to the space of G-invariant functions, as noted in Section 2.4.

[4]Note that there can be more than one option if the **stabiliser** $G_x = \{g \in G : g \cdot x = x\}$ is non trivial.

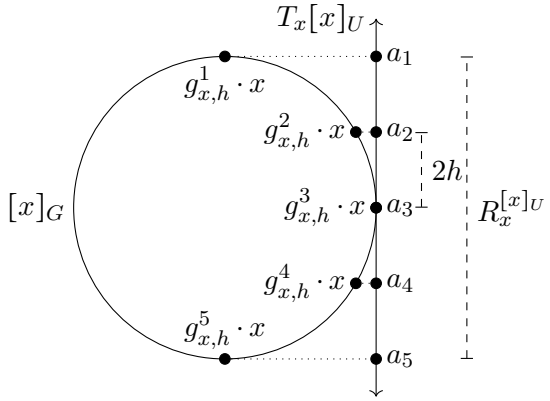


Fig. 4.2 The orbit of $x \in \mathbb{R}^2$ under the action of $G = SO(2)$, along with the points $\{a_i\}$ and $\{u_i = g^i_{x,h} \cdot x\}$ that will be used to partially symmetrise the estimator f_n. This is the same picture as Figure 3.1 but relabelled for the case of symmetry.

Proposition 31. *For all $x \in \mathcal{X}$ and $h > 0$ we have $m \geq \max(1, (R_x^{[x]_U}(2h)^{-1})^{\dim[x]_U})$, and $d_{\mathcal{X}}(g^i_{x,h} \cdot x, g^j_{x,h} \cdot x) \geq 2h$ whenever $i \neq j$.*

4.2.2 Pointwise Convergence Rates of Partially Symmetrised Estimators

Symmetrised estimators improve convergence because symmetrisation operators can have minimal effect on the bias (even none when f is G-invariant), but the variance can be dramatically reduced. This allows for a re-optimisation of the bias-variance trade-off for a faster converging estimator. This is formalised in this section, which requires us to make some assumptions on the base estimator f_n, the bandwidth h_n, and later on the group action. Recall from Section 3.3 assumption set (E):

Assumption set (E):

(E1) (Strict Locality) Conditioned on $\mathcal{D}_X$, $f_n(x)$ and $f_n(y)$ are independent whenever $d(x, y) \geq 2h$;

(E2) (Default Value) If the ball $B_{\mathcal{X}}(x, h)$ contains no X_i, then $f_n(x) = 0$;

(E3) (Optimality) There exist positive constants $B, V \in \mathbb{R}_{>0}$, such that for all $f \in \mathcal{F}$, for μ-almost all x, and for all n:

  (a) (Bias) If $B_{\mathcal{X}}(x, h) \cap \mathcal{D}_X$ is non-empty, then $|\mathbb{E}(f_n(x) \mid \mathcal{D}_X) - f(x)| \leq Bh^{\beta}$; and

  (b) (Variance) If $B_{\mathcal{X}}(x, h) \cap \mathcal{D}_X$ is non-empty, then $\mathrm{Var}(f_n(x) \mid \mathcal{D}_X) \leq \frac{V}{|B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X|}$.

Alongside these we introduce the further condition on optimality to ensure that the random deign regime we now also consider are amenable.

(E3) (Optimality) There exists a positive constant $C \in \mathbb{R}_{>0}$, such that for all $f \in \mathcal{F}$, for μ-almost all x, and for all n:

(c) (Risk) $\mathbb{E}((f_n(x) - f(x))^2) \leq C(B^2h^{2\beta} + V(n\mu(B_{\mathcal{X}}(x,h)))^{-1})$

This is clearly satisfied by fixed design regimes and estimators that satisfy (E3a) and (E3b) and so this is just a statement that the random design regimes do not introduce too much variance to the estimator.

We will also require the following conditions that relate the geometry of the (partial)-orbits $[x]_U$ to the measure μ and to the smoothness of the regression function.

Assumption set (Q):

(Q1) $\mathbb{E}\big((R_X^{[X]_U})^{-d^G}\big) < \infty$, where $X \sim \mu$; and

(Q2) $d^G < 2\beta$.

Condition (Q1) says that the distribution μ is not overly concentrated around any singularities of the action. This is satisfied in many normal situations, as shown in the example below. We also show that in some cases this is not satisfied. The second condition, (Q2), ensures that the new optimal bandwidth selection rule does not cause the variance of f_n to increase too rapidly. When $d^G \geq 2\beta$ the bound on the variance of the original estimator f_n in (E2) can in fact increase with n. This is balanced by averaging over $O(h^{-d^G})$ more points, resulting in the overall reduction in variance. However, if $d^G \geq 2\beta$ then we are saying that the number of points in each local ball $B(x,h)$ will not increase with n when $h = an^{-1/(2\beta+d-d^G)}$ for some $a > 0$, which causes many estimators to fail assumption set (E) (for example, a local polynomial estimator might fail to even exist without local data). This assumption is a slight weakening of the assumption that $\beta \geq d$ from the previous chapter, and we will discuss the effect of low β in Section 4.5.

Example 32 (Example of Assumption (Q1))**.** *Consider the action of $G = SO(3)$ on $\mathcal{X} = \mathbb{R}^3$ by rotations, and since G is compact we take $U = G$. The orbit of $x \in \mathcal{X}$ is the sphere of radius $\|x\|_2$, which means that $R_x^{[x]_G} = \sqrt{2}\|x\|_2$ for all $x \in \mathbb{R}^3$. We will consider three possible distributions of X all centered at the singularity of this action, the origin. If μ is a standard Gaussian distribution on $\mathcal{X}$, i.e., $X \sim N(0, I_3)$, then $\|X\|_2^2 \sim \chi_3^2$. The regular points of the action are the non-zero points of $\mathcal{X}$, and $d^G = 2$. Thus*

$$\mathbb{E}((R_X^{[X]_G})^{-d^G}) = \mathbb{E}((\sqrt{2}\|X\|_2)^{-2}) = 1/2 < \infty \tag{4.7}$$

using the fact that if $Y \sim \chi_\nu^2$ then $\mathbb{E}(1/Y) = 1/(\nu-2)$. If μ is the uniform distribution on the unit ball then $\|X\|_2 \overset{D}{=} z^{1/3}$ when $z \sim U([0,1])$ and $\|X\|_2^{-2} = 3 < \infty$ too. If instead the law

of X has $\|X_2\|_2^2 \sim \text{Cauchy}_+(0,1)$, where $Y \sim \text{Cauchy}_+(0,1)$ if it is the absolute value of a $\text{Cauchy}(0,1)$ variable, and is uniform in direction, then the expectation above is infinite.

These assumptions allow us to prove the following proposition bounding the point-wise convergence rate of the partially symmetrised estimator $S_{\rho_x^m} f_n(x)$. This is done by separate analysis of the point-wise bias and variance of this estimator, and the results for each of these can be found in Lemmas 45 and 46 in Section 4.6 along with the proof of this Proposition.

Proposition 33 (Point-wise MSPE of $S_{\rho_m^x} f_n$)**.** *Under assumption sets (E) and (Q), in the random design context, with the bandwidth choice $h_n = an^{-1/(2\beta+d-d^G)}$ for any positive a, we have:*

$$\sup_{f\in\mathcal{F}(L,\beta)} \mathbb{E}\big((S_{\rho_m^x} f_n(x) - S_{\rho_m^x} f(x))^2\big) \leq C_{G,x}\phi_n^{\beta,d-d^G} \tag{4.8}$$

for all $n \in \mathbb{Z}_{>0}$, where $C_{G,x}$ is given by:

$$C_{G,x} = \Big(2\big(B^2a^{2\beta} + \frac{2^{d^G}V}{c_\star(x)(R_x^{[x]_U})^{d^G}}\big) + (L+V)\big(\frac{c_\star(x)a^d(2\beta-d^G)}{16\beta}\big)^{\frac{2\beta}{d}}\Big) \tag{4.9}$$

where $c_\star(x) = \inf_{g\in U} c_{g\cdot x}$. In the fixed design regime these bounds hold for all sufficiently large n such that $|B_{\mathcal{X}}(x,h)\cap\mathcal{D}_X| > n\mu(B_{\mathcal{X}}(x,h))/2$.

4.2.3 Integrated Risk of Partially Symmetrised Estimators

The pointwise bound is true for almost all $x \in \mathcal{X}$, but not uniform because of the $R_x^{[x]_U}$ and $c_\star$ terms. The inverse moments of $R_X^{[X]_U}$ can blow up to infinity on neighbourhoods of singular points of the action. Assumption (Q1) controls this blow up, ensuring that μ is not too concentrated around the singularities of the action. This allows us to integrate the pointwise bound to gain integrated rates over some Borel subset $\Omega \subseteq \mathcal{X}$. When $\Omega \subseteq \mathcal{X}$ is compact, we can take $c_\star$ to be a constant in x over Ω, because its minimum over $\mathcal{X}$ works in the required bound. Thus we can obtain the integrated risk in the following proposition. Note that from here we sometimes drop the superscript x on ρ_m^x when considering $S_{\rho_m^x}\phi(x)$ as a function of $x \in \mathcal{X}$.

Theorem 34 (IMSPE of $S_{\rho_m} f_n$)**.** *Suppose that Ω is a compact Borel subset of $\mathcal{X}$ such that (Q2) holds when X is conditioned on $X \in \Omega$, and suppose that $c_x = c > 0$ for all $x \in \Omega$. Under assumption sets (E) and (Q), with a bandwidth $h_n = an^{-1/(2\beta+d-d^G)}$ we have:*

$$\sup_{f\in\mathcal{F}(L,\beta)} \mathbb{E}\big((S_{\rho_m^X} f_n(X) - S_{\rho_m^X} f(X))^2 \mid X \in \Omega\big) \leq C_{G,\Omega}\,\phi_n^{\beta,d-d^G} \tag{4.10}$$

where x is any regular point of the action, and $C_{G,\Omega} = \mathbb{E}(C_{G,X} \mid X \in \Omega) < \infty$ where $X \sim \mu$. In particular, when $\mathcal{X}$ is compact, we have:

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}\big(\|S_{\rho_m} f_n - S_{\rho_m} f\|_2^2\big) \leq C_G \, \phi_n^{\beta, d - d^G} \tag{4.11}$$

where $C_G = \mathbb{E}(C_{G,X}) < \infty$. These results also hold in the fixed design regime for sufficiently large n that $|B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X| > n\mu(B_{\mathcal{X}}(x,h))/2$

The condition that assumption (Q2) holds when conditioned on $X \in \Omega$ holds in almost all interested cases, for example, when Ω is a d-dimensional unit hypercube in $\mathbb{R}^d$ when $\mathcal{G} = \mathbb{R}^d$ acts by translation.

4.2.4 Computing Partial Symmetrisation Operators

The explicit construction of ρ_m^x is always possible given any point $x \in \mathcal{X}$, the bandwidth h, the group G, and its action on $\mathcal{X}$, however, it requires computational effort in the general case—especially when the orbits are non-linear. We now show that in some cases we can compute these averaging operators quickly by using the specific structure of the group action.

Covariate Sparsity - Linear Orbits

Recall Example 27, where we show that covariate sparsity in non-parametric regression is a special case of a translation symmetry. Specifically, in this case $\mathcal{G} = \mathbb{R}^d$ acts on $\mathcal{X} = \mathbb{R}^d$ by translation and the covariates that f does not depend on can be identified with subgroups of $\mathcal{G}$ of the form $\{ae_i : a \in \mathbb{R}\}$ for the standard basis $\{e_i\}$ of $\mathcal{X}$. When constructing symmetrisation operators we can instead exploit the uniform structure of the orbits of all subgroups of $\mathcal{G}$. Let $U = [-1,1]^d$, and see that $T_x[x]_U$ can be identified with U precisely (as there are only trivial stabilisers) for all $x \in \mathcal{X}$. Thus a $(2h)$-grid on $[x]_U$ is exactly the easily computable grid for the linear tangent space. This means that we need only one computation of the group elements g^i for the dataset $\mathcal{D}$.

Compact Groups

In the special case of a compact group G and when μ is G-invariant, we can use the properties of the full symmetrisation operator S_G to improve on the estimator $S_{\rho_m} f_n$. Specifically, we can show $S_G f_n(x) = S_G S_{\rho_m^x} f_n(x)$ by the shift invariance of the Haar measure (see the proof of the corollary below), which gives the result:

$$\|S_G f_n(x) - S_G f\|_2 = \|S_G S_{\rho_m^x} f_n - S_G S_{\rho_m^x} f\|_2 \leq \|S_{\rho_m^x} f_n - S_{\rho_m^x} f\|_2 \tag{4.12}$$

because S_G is an orthogonal projection with operator norm 1. Thus the fully symmetrised estimator $S_G f_n$ will outperform the partially symmetrised $S_{\rho_m} f_n$, almost surely. Thus as a corollary to Theorem 34, we obtain:

Corollary 35 (IMSPE of Fully Symmetrised Estimators)**.** *Suppose that $\mathcal{X}$ and G are both compact and that μ is G-invariant. Under conditions (E) and (Q), we have that:*

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}(\|S_G f_n - S_G f\|_2^2) \leq C_G \phi_n^{\beta, d - d^G} \tag{4.13}$$

for the same constant C_G as in Theorem 34.

Now we can estimate $S_G f_n$ with a simple Monte-Carlo estimator, where we average over M samples of $g_i \overset{iid}{\sim} U(G)$. This gives the estimator of f:

$$\hat{f}_{n,G}(x) = \frac{1}{M} \sum_{i=1}^{M} f_n(g_i \cdot x) \tag{4.14}$$

which is easily computable whenever we can sample uniformly from the group. The analysis of this estimator is simple when μ is G-invariant, and achieves the same rates as $S_{\rho_m} f_n$ and $S_G f_n$ when $M = n$.

Corollary 36. *Suppose $\mathcal{X}$ and G are compact, and that μ is G-invariant. Under assumption sets (E) and (Q), with $M \asymp n$, the estimator $\hat{f}_{n,G}$ satisfies:*

$$\sup_{f \in \mathcal{F}(L,\beta)} \mathbb{E}(\|\hat{f}_{n,G} - S_G f\|_2^2) \leq (12C + 12L^2 + 2C_G) \phi_n^{\beta, d - d^G} \tag{4.15}$$

where C_G is as in Proposition 34 and C is as in (E3).

4.3 Adaption to Maximal Symmetries

We can now consider the main problem of this chapter: what is the best subset $G \subseteq \mathcal{G}$ to use to (partially) symmetrise f_n to best estimate f? The first step is to note that if f is invariant to the elements of a subset $A \subseteq \mathcal{G}$, then f is also the group **generated** by A, i.e., the intersection of all subgroups of $\mathcal{G}$ that contain A, written $\langle A \rangle$. Moreover Lemma 11 ensures that f will be invariant to the topological closure $\overline{\langle A \rangle}$. Thus our best subset will be a closed subgroup of $\mathcal{G}$—a fact with simplifies our search remarkably to just the question: How can we estimate the maximal invariant subgroup of f, and can we do this quickly enough to improve the estimation of f?

Given any estimator f_n that satisfies the condition set (E), the rates in the previous section hold for the distributions $\rho_x^m(G)$ constructed point-wise in x for each closed subgroup $G \leq \mathcal{G}$.

So a simple two step M-estimator, where we minimise $\hat{\mathcal{E}}(S_{\rho_m^x(G)} f_n)$ over the closed subgroups $G \leq \mathcal{G}$, would give an estimate of the unique maximal subgroup of f, and the minimising estimator might be expected to perform with the rate $\phi_n^{\beta, d-d^{G_{\max}(f,\mathcal{G})}}$. In this section we should that it can be approximated by minimising over a well chosen finite set of subgroups (that grows with n), and that the minimiser over this set accomplishes the desired property of a risk that decays with the fast rate $\phi_n^{\beta, d-d^{G_{\max}(f,\mathcal{G})}}$, both globally when $\mathcal{X}$ is compact and locally when it is not.

This first requires three key pieces of analysis in order to do statistics when the objects are subgroups:

1) that we can construct a metric on the set of closed subgroups in the same way that we must with other object oriented data analysis tasks [Marron and Dryden, 2021];

2) that this set is totally bounded in this metric[5]; and

3) that we can bound the asymptotic bias of $f_{n,G}$ in terms of the distance between G and H (for which f is H-invariant).

We denote the set of closed subgroups of $\mathcal{G}$ as $K(\mathcal{G})$. Moreover, we will be greedy in our estimation by further reducing the problem to just connected subgroups. This is justified because if f is G-invariant, then f is also G'-invariant where G' is the connected component of the identity in G, and $d^{G'} = \dim[x']_{G'} = \dim[x]_G$ for the regular points of each actions. Therefore partial symmetrisation with respect to each of G' and G will achieve the same convergence rate $\phi_n^{\beta, d-d^G}$, though perhaps with differing constants.

With these mathematical results we can define our estimator of $G_{\max}(f, \mathcal{G})$, the Error Minimising Symmetry, and then our estimator for f, the Best Symmetric Estimator given by partial symmetrisation by the error minimising symmetry. We then prove that the Best Symmetric Estimator achieves the desired adaption to the maximal symmetry of f.

4.3.1 Metrising Subgroup Space

First, recall that if (X, d_X) is a metric space then the **Hausdorff metric** between compact subsets A, B is given by

$$d_{\text{Haus}}(A, B) = \max\Big(\sup_{g \in A} \inf_{h \in B} d_X(g, h), \sup_{h \in B} \inf_{g \in A} d_X(g, h) \Big) \tag{4.16}$$

If X is a compact space, then the set of closed subsets of X is compact under this Hausdorff metric [Henrikson, 1999]. This also means that it is totally bounded, i.e., for all $\epsilon > 0$ there

[5] recall that a subset U of a metric space (X, d_X) is **totally bounded** if it can be covered by a finite number of open ϵ balls $B_X(x_i, \epsilon)$ for any $\epsilon > 0$.

are a finite number of subsets A with open balls of radius ϵ in the Hausdorff metric that cover the set of compact subsets.

Remark 37. *Note that this is slightly different to the metric introduced in Section 3.3, which would here measure the Hausdorff distance between orbits. Instead, here we ignore the action and consider the distance between the group elements.*

All Lie groups $\mathcal{G}$ can be given a metric $d_{\mathcal{G}}(g,h)$, given by the length of the shortest geodesic from g to h under any chosen metric[6] on $\mathcal{G}$. Since $\mathcal{G}$ is locally compact groups we can take a compact neighbourhood U of the identity $e \in \mathcal{G}$. We define a finite metric $d_{\text{Haus}(U)} : K(\mathcal{G}) \times K(\mathcal{G}) \to [0,\infty)$ given by:

$$d_{\text{Haus}(U)}(G,H) = d_{\text{Haus}}(G \cap U, H \cap U) \tag{4.17}$$

Note that since $\mathcal{G}$ is Hausdorff we know $G \cap U$ is compact for all $G \in K(\mathcal{G})$, so this is well defined. Moreover, it inherits all the metric properties of d_{Haus} with base space U, for which the set of closed subsets, $\{\overline{A} \subseteq U\}$, is compact. The subgroups $K(\mathcal{G})$ can be identified with the subsets $\{G \cap U \subseteq U\} \subseteq \{\overline{A} \subseteq U\}$, which means that it is totally bounded (as a subset of a totally bounded space).

Now we need to consider how a small distance $d_{\text{Haus}(U)}(G,H)$ allows us to understand the behaviour of S_ρ (for some distribution ρ on G with $\text{supp}(\rho) \subseteq U \cap G$) over the space of H invariant functions. This allows us to bound the asymptotic bias when considering G-invariant estimators for an H invariant regression function f. The main requirements are bounds on derivatives of f and of the action of $\mathcal{G}$. In particular, we say that $\mathcal{G}$ has a $L_{\mathcal{G}}$-**Lipshitz** action on $\mathcal{X}$ if $d_{\mathcal{X}}(g \cdot x, h \cdot x) \leq L_{\mathcal{G}} d_{\mathcal{G}}(g,h)$ for all $x \in \mathcal{X}$ and $g,h \in \mathcal{G}$ (with $d_{\mathcal{G}}(g,h)$ the metric on the group $\mathcal{G}$ we use to compute Hausdorff distances). Since the action of $\mathcal{G}$ is smooth this condition only means that the derivative of $g \cdot x$ with respect to g is uniformly bounded over $x \in \mathcal{X}$. An example of this is the following.

Example 38 (Example of Lipschitz Action)**.** *Consider $\mathcal{G} = SO(3)$ acting naturally on the closed unit ball $\overline{B_{\mathbb{R}^3}(0,1)} \subseteq \mathbb{R}^3$. The Riemannian distance on $\mathcal{G}$ is given by $d_{\mathcal{G}}(g,h) = 2^{-1/2}\|\log(g^{-1}h)\|_F$, which is equal to the (minimal magnitude) angle of the rotation $g^{-1}h$ (Example 4.7 in Fletcher [2010]). Thus we have:*

$$d(g \cdot x, h \cdot x) = \|g \cdot x - h \cdot x\|_2 = \|(I - g^{-1}h)x\|_2 \tag{4.18}$$

Thinking of $g,h \in \mathcal{G}$ as matrices here. This expression is maximised when $\|x\|_2$ attains its maximum of 1, and this chord length is bounded by the angle between x and $g^{-1}h \cdot x$, so we

[6]When $\mathcal{G}$ is isomorphic to the Cartesian product of a compact group and $\mathbb{R}^k$ for some k, then there is a natural choice given by the unique bi-invariant Riemmanian metric, see Milnor [1976], Lemma 7.5.

have:

$$d(g \cdot x, h \cdot x) = \|(I - g^{-1}h)x\|_2 \leq 2^{-1/2}\|\log(g^{-1}h)\|_2 = d_{\mathcal{G}}(g,h) \tag{4.19}$$

Therefore the action on this space is 1*-Lipschitz.*

Proposition 39. *Let ρ_G^x be distributions on the groups G each with support contained in a compact neighbourhood of the identity U, and which can be distinct for each point $x \in \mathcal{X}$. Suppose that $\mathcal{G}$ has an $L_{\mathcal{G}}$-Lipschitz action on $\mathcal{X}$. Then for all $f \in \mathcal{F}(L,\beta) \cap L_H^2(\mathcal{X})$, we have:*

$$\left|(S_{\rho_G^x} f)(x) - f(x)\right| \leq L L_{\mathcal{G}}^{\alpha} d_{\mathrm{Haus}(U)}(G,H)^{\alpha} \tag{4.20}$$

for all $x \in \mathcal{X}$, where $\alpha = \min(\beta, 1)$.

4.3.2 Defining the Error Minimising Symmetries

Our goal now is to construct a finite set of closed subgroups for which the minimiser of $\hat{\mathcal{E}}(S_{\rho(G)} f_n)$ over this set obtains the rate of risk decay corresponding to the maximal symmetry of f. Recall that we need only consider the closed connected subgroups of $\mathcal{G}$, as noted at the start of section 4.3. Note that for convenience we suppress the scripts on $\rho(G) = \rho_m^x(G)$. To ensure that we capture examples for each possible rate, we first stratify the space of closed connected subgroups of $\mathcal{G}$ by the dimension of the regular orbits; i.e., we define:

$$K_\ell(\mathcal{G}) = \left\{G \in K(\mathcal{G}) : \dim[x]_G = \ell, G \text{ is connected}\right\}. \tag{4.21}$$

Since $K(\mathcal{G})$ is totally bounded under the metric $d_{\mathrm{Haus}(U)}$ each stratum $K_\ell(\mathcal{G})$ is totally bounded too (as every subset of a totally bounded set is totally bounded, see Lemma 72). So take any $\delta > 0$. For each ℓ, pick a finite set $\{G_i^\ell\}_{i=1}^{k_\ell}$ such that for all $G \in K_\ell(\mathcal{G})$, there exists a G_i^ℓ with $d_U(G, G_i) \leq \delta$. We then take our estimate, the **(Global) Error Minimising Symmetry** of $G_{\max}(f, \mathcal{G})$ as:

$$\hat{G}_\delta = \operatorname*{argmin}_{G \in \cup_\ell \{G_i^\ell\}} \hat{\mathcal{E}}(S_{\rho(G)} f_n) = \operatorname*{argmin}_{G \in \cup_\ell \{G_i^\ell\}} \sum_{i=1}^{n} (Y_i' - S_{\rho_m^{X_i'}(G)} f_n(X_i'))^2 \tag{4.22}$$

Where the empirical error is taken from an iid copy of the dataset $\mathcal{D}$ used in each f_n, given by $\mathcal{D}' = \{(X_i', Y_i')\}_{i=1}^n$. In the random design context this could be from a split of the data into two iid pieces.

Example 40 (Construction of δ-cover of $K(SO(3))$)**.** *Since $dim[x]_{SO(3)} = 2$, there are only three strata to cover. The easiest is $K_2(SO(3)) = \{SO(3)\}$, which is covered finitely by itself. The subgroups in $K_0(SO(3))$ are all finite, and the only connected finite group is the trivial group I. Thus we only need to really consider the stratum $K_1(SO(3))$ which contains only*

subgroups isomorphic to the circle S^1*, acting rotationally around the axes* u *in the unit sphere* S^2*, so* $K_1(SO(3)) = \{S^1_u : u \in S^2\}$*.*

Consider the distance between subgroups S^1_u *and* S^2_v*. For all* $g \in S^1_u$ *with angle* $\phi \in [0, \pi)$*, set* $h_g \in S^1_v$ *as the rotation around* v *with the same angle of rotation. Then consideration of the real part of the quaternions representing these rotations, we can find* $d_{SO(3)}(g, h_g) = 2^{-1/2} \| \log(g^{-1} h_g) \|_F = \theta$ *where* θ *is given by:*

$$\cos(\theta/2) = \cos(\phi/2)^2 + \langle u, v \rangle \sin(\phi/2)^2 = 1 - (1 - \langle u, v \rangle) \sin(\phi/2)^2 \geq \langle u, v \rangle \tag{4.23}$$

when $\langle u, v \rangle \geq 0$*. This implies that* $\theta \leq 2 \arccos(\langle u, v \rangle)$ *in this case, i.e., less than twice the angle between* u *and* v*. Thus we have:*

$$\sup_{g \in S^1_u} \inf_{h \in S^1_v} d_{SO(3)}(g, h) \leq \sup_{g \in S^1_u} d_{SO(3)}(g, h_g) \leq 2 \arccos(\langle u, v \rangle) \tag{4.24}$$

Hence the δ*-grid of* $K_1(SO(3))$ *is easily given by* $\{S^1_{u_{ij}}\}$ *where* $\{u_{ij}\}$ *forms a* $2\delta/\pi$ *grid of the unit sphere* S^2 *under the angular metric. These* u_{ij} *can in turn be generated from a* δ/π *grid of angles in spherical coordinates* $(\phi_i, \theta_i) \in [0, 2\pi) \times [0, \pi]$*, with* $u_{ij} = (\sin\theta \cos\phi, \sin\theta \sin\phi, \cos\phi)$*. This is depicted in Figure 4.3.*

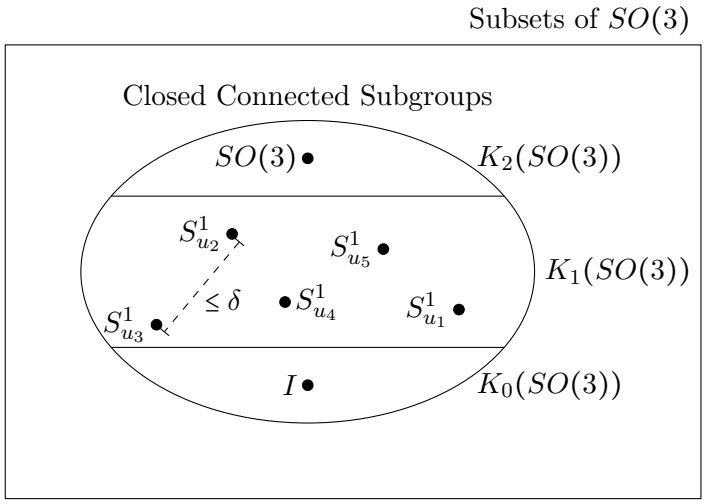


Fig. 4.3 Example of the search space of the error minimising symmetry when $\mathcal{G} = SO(3)$, as in Example 40. As δ decreases, the middle stratum gains more points in the δ-cover.

4.3.3 Global Rates for the Best Symmetric Estimator

We now can study the properties of the **Best Symmetric Estimator** of f:

$$\hat{f} = S_{\rho(\hat{G}_{\delta_n})} f_n \tag{4.25}$$

where f_n is any estimator satisfying assumption set (E), the partial symmetrisation operators are constructed as per section 4.2.1, δ_n is a deterministically chosen hyperparameter, and $\hat{G}_{\delta_n}$ is the global error minimising symmetry as per section 4.3.2. We first consider the case where $\mathcal{X}$ is compact, as with $[0,1]^d$, S^d, and $\mathbb{T}^d$. In this case the Best Symmetric Estimator achieves the desired rates of adaption to the maximal symmetry of f.

Theorem 41. *Let $\mathcal{X}$ be a compact space. Let $\delta_n = L_{\mathcal{G}}^{-1}\big(\frac{\phi_n^{\beta,d-d^{\mathcal{G}}}}{2L^2}\big)^{1/2\min(\beta,1)}$. Let $\{G_i^\ell\}_{i=1}^{k_\ell}$ be a finite δ-cover of the totally bounded sets $K_\ell(\mathcal{G}) \subseteq K(\mathcal{G})$ for $\ell = 0,\dots,\dim[x]_{\mathcal{G}}$. Let $\hat{G}_\delta$ be the minimiser of $\hat{\mathcal{E}}(f_{n,G})$ over $\cup_\ell\{G_i^D\}$. Then under assumptions (E) and (Q), the Best Symmetric Estimator $S_{\rho(\hat{G}_{\delta_n})} f_n$ achieves the rate:*

$$\sup_{f\in\mathcal{F}(L,\beta)} \frac{\mathbb{E}_f\big(\|S_{\rho(\hat{G}_{\delta_n})} f_n - f\|_2^2\big)}{\phi_n^{\beta,d-d^{G_{\max}(f,\mathcal{G})}}} \le 2(1 + \sup_{G\in K(\mathcal{G})} C_G) \tag{4.26}$$

Under many circumstances it is possible to bound C_G uniformly over the subgroups in $K(\mathcal{G})$ as in the example below. For this to be true, it is sufficient that $c_\star(x) = c$ for all $x \in \mathcal{X}$ and $\mathbb{E}((R_X^{[X]_U})^{-d^G}) < A$ for all $G \in K(\mathcal{G})$ for some constant $A > 0$.

Example 42. *Consider the case of $\mathcal{G} = SO(3)$ acting naturally on the unit sphere S^2 with a uniform $\mu = U(S^2)$. As shown in Example 32 we know that $\mathbb{E}((R_x^{[x]_{SO(3)}})^{-\dim[x]_{SO(3)}}) = 2$. The identity group has this expectation equal to 1 because R is constant. The closed connected subgroups with dimension 1 orbits are the groups S_u^1 each isomorphic to $SO(2)$. For regular points x we have $R_x^{S_u^1} = \sqrt{1-\langle x,u\rangle}$ because the orbits are circles of this radius around u. The inner product has the same distribution as the first coordinate of X by rotational symmetry, which is uniform on $[-1,1]$ (a special case for $U(S^2)$). Thus the distribution of $R_X^{S_u^1}$ is triangular, with $\mathbb{P}(\sqrt{1-\langle x,u\rangle} < a) = a^2/2$ for $a \in [0,\sqrt{2}]$ and thus has a linear density function over this region. The reciprocal expectation is thus simply*

$$\mathbb{E}\big((R_X^{S_u^1})^{-1}\big) = \int_0^{\sqrt{2}} a^{-1} f_{R_X^{S_u^1}}(a)\,\mathrm{d}a = \int_0^{\sqrt{2}} \mathrm{d}a = \sqrt{2} \tag{4.27}$$

This constant is greater than for the larger action because it has more singularities (i.e., the line through u rather than just the origin).

4.3.4 Local Rates for the Best Locally Symmetric Estimator

In the case that $\mathcal{X}$ is not compact, we can still establish a similar local integrated rates. Instead of minimising the empirical risk, we minimise over a local estimate of the integrated error of $S_{\rho_{\hat{G}_\delta}} f_n$ over a compact set $\Omega = \overline{\Lambda} \subseteq \mathcal{X}$ for a non-empty open set Λ. This is slightly more technical than the global case, but only because we have to account for the "no data" case. Specifically, we define the **Local Error Minimising Symmetry** as:

$$\hat{G}_\delta(\Omega) = \begin{cases} \operatorname*{argmin}_{G \in \cup_\ell \{G_i^\ell\}} \hat{\mathcal{E}}_\Omega(S_{\rho(G)} f_n) & \text{if } \sum_{i=1}^n \mathbf{1}_{X_i' \in \Omega} > 0 \\ I & \text{otherwise} \end{cases} \tag{4.28}$$

for every $x \in \mathcal{X}$, where:

$$\hat{\mathcal{E}}_\Omega(S_{\rho(G)} f_n) = \begin{cases} \frac{\sum_{i=1}^n \left(Y_i' - S_{\rho(G)} f_n(X_i')\right)^2 \mathbf{1}_{X_i' \in \Omega}}{\sum_{i=1}^n \mathbf{1}_{X_i' \in \Omega}} & \text{if } \sum_{i=1}^n \mathbf{1}_{X_i' \in \Omega} > 0 \\ L & \text{otherwise} \end{cases} \tag{4.29}$$

and all other terms are as for the global error minimising symmetry. The conditions on our covariate design and bandwidth ensure that the number of terms in Ω is expected to grow with $n\mu(\Omega)$, so with high probability (i.e., exponentially increasing in n) we will have an estimate in the first case of equation (4.28). When $\Omega = \overline{B_{\mathcal{X}}(x,h)}$ this is effectively a Nadaraya-Watson estimator of the pointwise risk. This then leads to the **Best Locally Symmetric Estimator**, $S_{\rho(\hat{G}_\delta(\Omega))} f_n$ which obtains rates in the following theorem.

Theorem 43. *Let Ω be a compact subset that is the closure of an open subset of $\mathcal{X}$. Let*

$$\delta = L_{\mathcal{G}}^{-1}\left(\frac{\phi_n^{\beta, d-d^{\mathcal{G}}}}{2L^2}\right)^{1/2\min(\beta,1)} \tag{4.30}$$

Let $\{G_i^\ell\}_{i=1}^{k_\ell}$ be a finite δ-cover of the totally bounded sets $K_\ell(\mathcal{G}) \subseteq K(\mathcal{G})$ for $\ell = 0, \ldots, \dim[x]_{\mathcal{G}}$. Let $\hat{G}_\delta$ be the local error minimising symmetry. Under assumptions (E) and (Q), we have that

$$\sup_{f \in \mathcal{F}(L,\beta)} \frac{\mathbb{E}\left((S_{\rho(\hat{G}_\delta(\Omega))} f_n(X) - f(X))^2 \mid X \in \Omega\right)}{\phi_n^{\beta, d-d^{G_{\max}(f,\mathcal{G})}}} \leq C_0 \tag{4.31}$$

where the constant C_0 is given by:

$$C_0 = \sup_{G \in K(\mathcal{G})} 2C_{G,\Omega} + 2 + (2C_{G,\Omega} + 4L)\tfrac{2\beta}{\mu(\Omega)(2\beta+d)} \tag{4.32}$$

and $C_{G,\Omega}$ is as defined in section 4.2.3.

4.4 Finite Sample Performance

We now examine the finite sample performance of best symmetric estimators using synthetic data in the contexts of the main examples of this chapter, namely 3D Rotations and translations. All code can be found in the supplementary material [Christie and Aston, 2024b].

4.4.1 3D Rotations

Suppose that $\mathcal{X} = \overline{B_{\mathbb{R}^3}(0,1)}$, the closed unit ball in $\mathbb{R}^3$. Let $\mathcal{G} = SO(3)$ (a well known compact Lie group) act naturally by matrix multiplication on $\mathcal{X}$ as in Example 30. Suppose that μ is the uniform measure on $\mathcal{X}$, so that c_x is the reciprocal of the Lebesgue volume of $\mathcal{X}$ for all x. To estimate the global error minimising symmetry we need:

1. to show that this action is $L_{SO(3)}$-Lipschitz; and
2. to construct a δ-cover of each $K_\ell(SO(3))$.

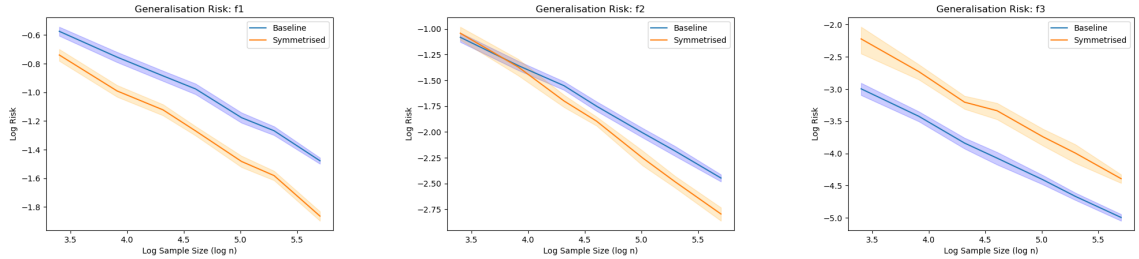
For this group and covariate space these are done in the examples 38 and 40 in the previous sections.

We have used a local constant estimator (LCE), defined explicitly in 2.1, originally proposed in Nadaraya [1964], Watson [1964]), with a rectangular kernel as our f_n. We consider the Hölder class $\mathcal{F}(L, \beta = 1)$ and simulated draws of datasets with $Y_i = f(X_i) + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $X_i \overset{iid}{\sim} U(\mathcal{X})$, for sample sizes $n \in \{30, 50, 75, 100, 150, 200, 300\}$. We sample two independent copies of the dataset for each simulation, one used to estimate f and the other used to estimate $G_{\max}(f, \mathcal{G})$ for the symmetrised estimator. The baseline LCE uses the union of these datasets to estimate f. We have used a bandwidth $h_n^G = n^{-1/(2\beta+d-d^G)}$ for each partially symmetrised estimator depending on $G \leq \mathcal{G} = SO(3)$, and the minimax optimal $h = h_n^I$ for the baseline LCE. When sampling from any subgroup of $\mathcal{G}$ we use the uniform distribution, as per the results in section 4.2.4. We have used 3 choices of f, with varying levels of symmetry:

Regression function	Maximal Symmetry
$f_1(x) = \cos(\|x\|_2)$	$SO(3)$
$f_2(x) = \cos(\sqrt{x_2^2 + x_2^3})$	$S^1_{e_1}$
$f_3(x) = x_1^2 + x_2 - 0.6x_3$	I

Note that f_1 does have other symmetries (for example it is invariant to the supergroup $O(3)$) though these are not contained in the search group $\mathcal{G} = SO(3)$. We compared an estimated risk of the baseline LCE, $\frac{1}{K}\sum_{i=1}^K (f_n(X_i') - f(X_i'))^2$ for $K = 200$, against the symmetrised LCE, $\frac{1}{K}\sum_{i=1}^K (S_{\hat{G}} f_n(X_i') - f(X_i'))^2$ for $X_i' \overset{iid}{\sim} \mu$ that are independent of both the data used to

the iid copy used estimate f_n and $\hat{G}_\delta$. The average results of 30 trials for each sample size are plotted on a log-log scale in Figure 4.4.



(a) f_1 (b) f_2 (c) f_3

Fig. 4.4 Estimated risk for a baseline Local Constant Estimator f_n in blue against the symmetrised $S_{\rho_{\hat{G}_\delta}} f_n$ in orange, as well as 95% Wald confidence intervals for the true risk. Plots are on a log-log scale.

We see that the true regression function has some level of symmetry, then our estimator does indeed obtain a faster rate than the baseline LCE. When there is no symmetry within the search group $\mathcal{G}$ our estimator performs slightly worse because of the variance in the estimated group and the half sized sample for estimation of f, but does obtain the same rate of risk decay. Note that the slopes slightly increase with n as the boundary effects diminish.

4.4.2 Translational Symmetry

Consider the situation of Example 29, where $\mathcal{X} = (\mathbb{R}/\mathbb{Z})^2$ is the (flat) 2-torus and $\mathcal{G} = (\mathbb{R}/\mathbb{Z})^2$ acts by translation. Recall that if f is invariant to subgroups of translation along an axis then f does not depend on this coordinate of the covariate vector X, so captures sparsity in the covariates. As in the previous example, we first establish the required Lipschitz bound and δ-grid:

1) The action is 1-Lipschitz because addition is Lipschitz; and

2) A δ-grid can be constructed by considering the angles of lines through the origin. All closed one dimensional subgroups have this form with a rational angle θ measured from the x-axis, and the distance between them (under the natural metric on $\mathcal{G}$), is bounded as:

$$d_{\mathcal{G}}(G, H) \leq \sqrt{2}\sin(\theta_G - \theta_H) \leq \sqrt{2}(\theta_G - \theta_H) \tag{4.33}$$

Thus a δ-grid is given by lines of angle θ_i where $\theta_i = 360°i/I$, $I = \lceil 360/(\sqrt{2}\delta) \rceil$ and $i \in \{0, 1, \ldots, I\}$. In practice we would like lines with low denominators so we take the

union of these grids over $J \leq I$, i.e.:

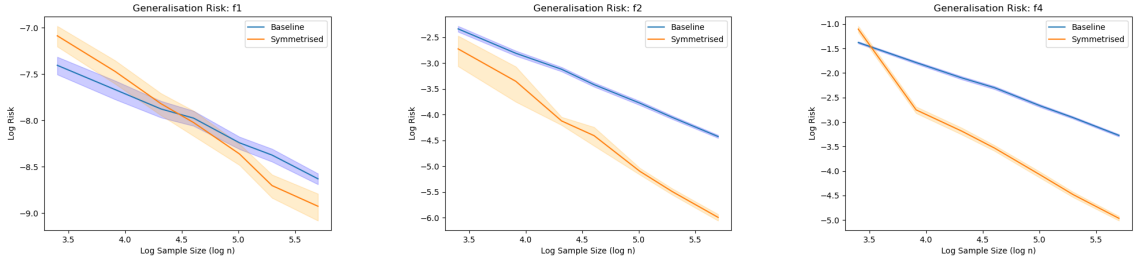
$$\{G_j^1\}_{j=1}^k = \bigcup_{J=1}^{I} \{S^1_{360°\, i/J} : i = 0, \ldots, J\} \tag{4.34}$$

where S^1_θ is the 1 dimensional group given by the line through the origin in $\mathbb{T}^2$ with angle θ.

We have used the same base estimator as in the previous example, and simulated draws in the same regression model for three choice of regression functions f:

Regression function	Maximal Symmetry
$f_1(x) = 1$	$\mathcal{G} = \mathbb{T}^2$
$f_2(x) = \sin(2\pi x_1)$	$G_2 = S^1_{\theta=0}$
$f_3(x) = \cos(2\pi(x_1 - x_2))$	$G_3 = S^1_{\theta=-\frac{\pi}{2}}$

The average estimated risks across 100 simulations are plotted in Figure 4.5, where we see that the symmetrised estimator outperforms the baseline LCE in terms of rate for all scenarios, and so has lower absolute performance for $n \geq 150$. At very low sample sizes the performance of the Best Symmetric Estimator is reduced by the variability in the estimated group.



(a) f_1 (b) f_2 (c) f_3

Fig. 4.5 Estimated generalisation risk for a baseline Local Constant Estimator f_n in blue against the symmetrised $S_{\rho_{\hat{G}_\delta}} f_n$ when the same data is used in both steps of the M-estimation, as well as 95% Wald confidence intervals for the true risk. Plots are on a log-log scale.

For low sample sizes we sometimes see that the symmetrised estimator performs worse than the baseline, which is likely due to a sub-optimal choice of bandwidth that creates instances where $D_X \cap B(g_i \cdot x, h)$ is empty.

4.5 Discussion

In this chapter we have given sufficient conditions for the adaption to the symmetries present in a non-parametric regression function f. This means that we can learn patterns present in

our non-parametric regression functions and use these patterns to better generalise. This answers the question posed in the previous chapter in the positive for the class of symmetric level sets. It has used the manifold averaged estimator but used the additional structure of this situation to prove stronger results: we can show it works in the random design context and remove some log factors. We now mention a few avenues for future work, and make a few remarks.

4.5.1 Low Smoothness Case

We have assumed (Q2), that $d^G < 2\beta$, so that the optimal bandwidth h decays slowly enough that each ball $B_{\mathcal{X}}(g_i \cdot x)$ will contain increasing numbers of X_i. It would be ideal to remove this assumption by somehow combining the data from all m balls into one, i.e., use the "quotient space local data" $\mathcal{D}^G = \{(g_j \cdot X_i, Y_i) : g_j \in \mathrm{Supp}(\rho_m^x(G))\}$ and taking the estimate of $f(x)$ as $f_n(x; \mathcal{D}^G)$. This would ensure that we can use approximately $O(mnh^d) = (n^{2\beta/(2\beta+d-d^G)})$ datapoints for the local estimation at x giving optimal variance. This is essentially what happens for a local constant estimator symmetrised by S_{ρ_G} if we only average over the points $f_n(g_i \cdot x)$ for which $B_{\mathcal{X}}(g_i \cdot x) \cap \mathcal{D}_X$ is non-empty. This would allow us to achieve mini-max optimal symmetry adaptive estimation even when $\beta = 1$ for any dimension of the orbit. For $\beta \in (1, d^G/2)$ one would hope that we could establish the same result, however the story becomes more complicated. If f is not G invariant and we use the dataset $\mathcal{D}^G$ then the introduced bias will change based on the method of estimation used; we might lose the fast decay of $\|f_n(x; \mathcal{D}^G) - S_{\rho_m^x(G)} f(x)\|_2^2$ that allows us to construct the M-estimator of unknown symmetries. We conjecture that there is a class of $\mathcal{G}$-stable base estimators f_n that satisfy both assumption set (E) and ensures this convergence property for these intermediate smoothness classes.

4.5.2 Local Symmetries

Whilst the requirement that f has a global compact symmetry may be common, it can be restrictive. A much more general condition is that f obeys a *local* symmetry, where the level set of $f(x)$ and the orbit $[x]_G$ intersect on a compact connected subset of the orbit containing x. This captures situations such as digit classification, where we cannot rotate a 6 arbitrarily before it becomes a 9 but we can make small rotational perturbations. The theorems in this chapter, even relating to the local error minimising symmetry, require that the local orbits are the same everywhere, which means that they must be able to be patched together into a global symmetry. It would be interesting to understand how the best locally symmetric estimator performs when the requirement of global symmetry is dropped.

4.5.3 Improvements on the Constants

One problem with the method presented is that it can be wasteful with the split between data used to estimate f and used to estimate $G_{\max}(f, \mathcal{G})$, which means we see worse finite sample performance when f has no symmetries. Alternatively, a leave one out cross-validation approach could also be examined.

Secondly we have ignored non connected subgroups of $\mathcal{G}$ in our search for $G_{\max}(f, \mathcal{G})$. Such groups allow for further reduction in variance and potentially increased performance, at the cost of a much larger search for the error minimising symmetry.

4.5.4 Lower Bounds on the Estimation of Maximal Symmetries

A natural question that arises in this chapter is "does $\mathbb{E}(d_{\text{Haus}(U)}(\hat{G}, G_{\max}(f, \mathcal{G})))$ converge uniformly to 0 for all $f \in \mathcal{F}(L, \beta)$, and if so, how quickly?". Unfortunately, this question has a negative answer because we can have a sequence of non-G-invariant functions that approximate a G-invariant f arbitrarily well in $L^2(\mathcal{X})$, and distinguishing the symmetries is at least as statistically challenging as distinguishing the functions. This is explicit in the following example.

Example 44. *Suppose that $\mathcal{X} = \overline{B_{\mathbb{R}^3}(0,1)}$ and $\mu = U(\mathcal{X})$. Consider the functions $f_3(x) = \exp(-(x_1^2 + x_2^2))$ and $f_2 = \exp(-(x_1^2 + x_2^3))$, which are invariant to rotations around the z-axis (S_z^1) and y-axis (S_y^1) respectively, as well as reflections through the x-y plane (R_1) and x-z plane (R_2) respectively, giving their maximal symmetries. Both are in the Hölder class $\mathcal{F}(L, 2)$ for some L. This is also true for any multiple of these functions by $c \in (0, 1]$.*

Suppose we have normally distributed noise of standard deviation 1, so $Y_i \mid X_i \sim N(f(X_i), 1)$. Consider that we can create hypotheses $P_0 : f = cf_3$ and $P_1 : f = cf_2$ with $KL(P_0, P_1)$ arbitrarily small Kullback-Leibler divergence $KL(P_0, P_1) = \frac{c^2}{2}\|f_x - f_y\|_2^2$. This is can be seen by using the fact that $KL(N(\mu_1, 1), N(\mu_2, 1)) = (\mu_1 - \mu_2)^2/2$ and writing ϕ for the density of a standard normal. So taking $c = \sqrt{n^{-1}\|f_x - f_y\|_2^{-2}\log(4)}$ ensures $KL(P_0, P_1) \leq n^{-1}\log 2$, and thus Le Cam's lemma [Tsybakov, 2008] implies

$$\inf_{\hat{G}} \sup_{f \in \mathcal{F}(L,2)} \mathbb{E}_f\left(d_{\text{Haus}(SO(3))}(\hat{G}, G_{\max}(f))\right) \geq \frac{1}{16} d_{\text{Haus}(SO(3))}(G_{\max}(f_0), G_{\max}(f_1)) \tag{4.35}$$

where the $G_{\max}$ terms have the second argument $SO(3)$ suppressed. Since this bound is a positive constant, we see this problem is essentially intractable.

This is not to say that the convergence is not possible for any particular regression function f, and it would be an interesting question of future study as to how this convergence behaves for different classes of symmetries.

4.5.5 A remark on equivariance

In this chapter we have considered only invariant functions, for which $f(g \cdot x) = f(x)$. In the machine learning literature there is often consideration of the class of **equivariant** functions, for which

$$f(g \cdot X) = g \star f(X) \tag{4.36}$$

for some action $\star$ on the domain of Y [Bronstein et al., 2021]. However, in terms of statistical theory, these functions behave very similarly to invariant functions. This is because the function f can be decomposed into a G-invariant function $\tilde{f} : \mathcal{X}/G \to \mathcal{Y}/G$ with $\tilde{f}(x) = [f(x)]_G$ (i.e., a function which picks the output's orbit) and another G-invariant function $r : \mathcal{X}/G \to G$ (that picks "starting" points) such that $f(x) = r(x) \star \iota_\star(\tilde{f}(x))$. Both pieces are easier to learn than the original function, both because the input space is reduced to $\mathcal{X}/G$ and the output space is split into $\mathcal{Y}/G$ and G. However, because we are still estimating both pieces the errors will still combine - obscuring any additional benefit from the quotient on $\mathcal{Y}$.

4.6 Proofs

4.6.1 Proofs in Section 4.2

Proof of Proposition 31. Each side length can afford at least $\lceil R_x(2h)^{-1} \rceil$ points with $2h$ spacing, and we can do this in each of the dimensions of $T_x\mathcal{M}$. This gives the first inequality in:

$$m = |\{a_i\}_{i=1}^N| \geq \lceil R_x^{[x]_U}(2h)^{-1} \rceil^{\dim[x]_U} \geq \lceil (R_x^{[x]_U}(2h)^{-1})^{\dim[x]_U} \rceil \geq \max(1, (R_x^{[x]_U}(2h)^{-1})^{\dim[x]_U}) \tag{4.37}$$

The second then follows as a basic fact about ceiling functions, as if $\{x\} = x - \lfloor x \rfloor = 0$ this is trivial and otherwise:

$$\lceil x^n \rceil = \left\lceil \sum_{k=0}^{n} \binom{n}{k} \times \lfloor x \rfloor^k \times \{x\}^{n-k} \right\rceil \leq \sum_{k=0}^{n} \binom{n}{k} \times \lfloor x \rfloor^k \times \lceil \{x\}^{n-k} \rceil \leq \sum_{k=0}^{n} \binom{n}{k} \times \lfloor x \rfloor^k \times 1 = \lceil x \rceil^n \tag{4.38}$$

Now the points $\{u_i\}$ have $d_{\mathcal{X}}(u_i, u_j) \geq 2h$. To see this, let n_i be the normal vector for which $u_i = a_i + n_i$ and $\langle n_i, a_i - x \rangle = 0$ for each i in the embedding space $\mathbb{R}^q$. Then we have (with d the metric on the embedding space $\mathbb{R}^q$),

$$d(u_i, u_j)^2 = d(a_i, a_j)^2 + d(n_i, n_j)^2 \geq d(a_i, a_j)^2 \geq (2h)^2 \tag{4.39}$$

by the orthogonality conditions. Then the isometry of the embedding of $\mathcal{X}$ into $\mathbb{R}^q$ ensures that the minimal path lengths in $\mathcal{X}$ connecting u_i and u_j are at least as long as the euclidean paths in the ambient space, giving the result. □

We now provide the proof regarding the pointwise risk of partially symmetrised estimators, which is best analysed by considering the squared bias and the variance of the estimator separately. We will use the shorthand $p_{x,h} = \mu(B_{\mathcal{X}}(x,h))$, and $g^i = g^i_{x,h}$.

Lemma 45 (Point-wise Bias of $S_{\rho^x_m} f_n$)**.** *Under the conditions of Proposition 33, we have:*

$$\left|\mathbb{E}(S_{\rho^x_m} f_n(x)) - S_{\rho^x_m} f(x)\right| \leq Bh^\beta + \sup_{i\in[m]} \exp(-np_{g^i\cdot x,h})\|f\|_\infty \tag{4.40}$$

when $h = an^{-1/(2\beta+d-d^G)}$, for all n in the random design regime. In the fixed deisgn regime these bounds hold for sufficiently large n.

Proof of Lemma 45. Using the definition of the symmetrisation operator, we see:

$$\begin{aligned}
\left|\mathbb{E}(S_{\rho^x_m} f_n(x)) - S_{\rho^x_m} f(x)\right| &= \left|\mathbb{E}\Big(\tfrac{1}{m}\sum_{i=1}^m f_n(g^i\cdot x) - f(g^i\cdot x)\Big)\right| && (4.41)\\
&= \left|\tfrac{1}{m}\sum_{i=1}^m \mathbb{E}\big(f_n(g^i\cdot x) - f(g^i\cdot x)\big)\right| && (4.42)\\
&\leq \tfrac{1}{m}\sum_{i=1}^m \left|\mathbb{E}\big(f_n(g^i\cdot x) - f(g^i\cdot x)\big)\right| && (4.43)\\
&= \tfrac{1}{m}\sum_{i=1}^m \left|\mathbb{E}\big(\mathbb{E}(f_n(g^i\cdot x)\mid \mathcal{D}_X) - f(g^i\cdot x)\big)\right| && (4.44)\\
&\leq \tfrac{1}{m}\sum_{i=1}^m \mathbb{E}\big(\left|\mathbb{E}(f_n(g^i\cdot x)\mid \mathcal{D}_X) - f(g^i\cdot x)\right|\big) && (4.45)\\
&= \tfrac{1}{m}\sum_{i=1}^m \mathbb{E}\big(\left|\mathbb{E}(f_n(g^i\cdot x)\mid \mathcal{D}_X) - f(g^i\cdot x)\right| \mid N_{g^i\cdot x,h} > 0\big)\mathbb{P}(N_{g^i\cdot x,h} > 0) + && (4.46)\\
&\qquad \mathbb{E}\big(\left|\mathbb{E}(f_n(g^i\cdot x)\mid \mathcal{D}_X) - f(g^i\cdot x)\right| \mid N_{g^i\cdot x,h} = 0\big)\mathbb{P}(N_{g^i\cdot x,h} = 0) && (4.47)\\
&\leq \tfrac{1}{m}\sum_{i=1}^m Bh^\beta + |f(g^i\cdot x)|\exp(-np_{g^i\cdot x,h}) && (4.48)
\end{aligned}$$

where $N_{x,h} = |B_{\mathcal{X}}(x,h)\cap \mathcal{D}_X|$. This uses assumption (E3c) and Lemma 71. In this fixed design regime these bounds hold for sufficiently large n. □

Lemma 46 (Point-wise Variance of $S_{\rho^x_m} f_n$)**.** *Under the conditions of Proposition 33, we have:*

$$\mathrm{Var}(S_{\rho^x_m} f_n(x)) \leq \tfrac{1}{m}\sup_{i\in[m]}\left(\tfrac{2V}{np_{g^i\cdot x,h}} + V\exp(-np_{g^i\cdot x,h}/8) + mB^2h^{2\beta}\right) \tag{4.49}$$

when $h = an^{-1/(2\beta+d-d^G)}$. This bound holds for all n in the random design regime, and for all sufficiently large n in the fixed design regime.

Proof of Lemma 46. First consider expanding the variance by the law of total variance:

$$\mathrm{Var}(S_{\rho_m^x} f_n(x)) = \mathbb{E}\big(\mathrm{Var}(S_{\rho_m^x} f_n(x) \mid \mathcal{D}_X)\big) + \mathrm{Var}\big(\mathbb{E}(S_{\rho_m^x} f_n(x) \mid \mathcal{D}_X)\big) \tag{4.50}$$

We now consider each term. Let $N_{x,h} = |B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X|$.

$$\mathbb{E}\big(\mathrm{Var}(S_{\rho_m^x} f_n(x) \mid \mathcal{D}_X)\big) = \mathbb{E}\big(\mathrm{Var}(\tfrac{1}{m}\sum_{i=1}^{m} f_n(g^i \cdot x) \mid \mathcal{D}_X)\big) \tag{4.51}$$

$$= \mathbb{E}\big(\tfrac{1}{m^2}\mathrm{Var}(\sum_{i=1}^{m} f_n(g^i \cdot x) \mid \mathcal{D}_X)\big) \tag{4.52}$$

$$= \tfrac{1}{m^2}\sum_{i=1}^{m} \mathbb{E}\big(\mathrm{Var}(f_n(g^i \cdot x) \mid \mathcal{D}_X)\big) \tag{4.53}$$

because of the strict locality of f_n, the construction of $\{g^i\}_{i=1}^m$ which ensures $2h$-spacing. Now we expand each expectation over the $\mathcal{D}_X$-events $\Xi_i = \{N_{g^i \cdot x,h} = 0\}$.

$$\mathbb{E}\big(\mathrm{Var}(f_n(g^i \cdot x) \mid \mathcal{D}_X)\big) = \mathbb{E}(\mathrm{Var}(f_n(g^i \cdot x) \mid \mathcal{D}_X) \mid \Xi_i)\mathbb{P}(\Xi_i) + \tag{4.54}$$

$$\mathbb{E}(\mathrm{Var}(f_n(g^i \cdot x) \mid \mathcal{D}_X) \mid \Xi_i^C)(1 - \mathbb{P}(\Xi_i)) \tag{4.55}$$

$$= 0 \times \mathbb{P}(\Xi_i) + \mathbb{E}\big(\tfrac{V}{N_{g^i \cdot x,h}} \mid N_{g^i \cdot x,h} > 0\big)(1 - \mathbb{P}(\Xi_i)) \tag{4.56}$$

$$\leq \mathbb{E}\big(\tfrac{V}{N_{g^i \cdot x,h}} \mid N_{g^i \cdot x,h} > 0\big) \tag{4.57}$$

$$= \mathbb{E}\big(\tfrac{V}{N_{g^i \cdot x,h}} \mid 0 < N_{g^i \cdot x,h} < np_{g^i \cdot x,h}/2)\big)\mathbb{P}(0 < N_{g^i \cdot x,h} < np_{g^i \cdot x,h}/2)) + \tag{4.58}$$

$$\mathbb{E}\big(\tfrac{V}{N_{g^i \cdot x,h}} \mid N_{g^i \cdot x,h} > np_{g^i \cdot x,h}/2)\big)\mathbb{P}(N_{g^i \cdot x,h} > np_{g^i \cdot x,h}/2)) \tag{4.59}$$

$$\leq V\mathbb{P}(N_{g^i \cdot x,h} < np_{g^i \cdot x,h}/2)) + \tfrac{2V}{np_{g^i \cdot x,h}} \tag{4.60}$$

$$\leq \tfrac{2V}{np_{g^i \cdot x,h}} + V\exp(-np_{g^i \cdot x,h}/8) \tag{4.61}$$

which uses the concentration inequality in Lemma 71. Thus we have:

$$\mathbb{E}\big(\mathrm{Var}(S_{\rho_m^x} f_n(x) \mid \mathcal{D}_X)\big) \leq \tfrac{1}{m}\sup_{i \in [m]}\big(\tfrac{2V}{np_{g^i \cdot x,h}} + V\exp(-np_{g^i \cdot x,h}/8)\big) \tag{4.62}$$

For the second term, consider:

$$\mathrm{Var}\big(\mathbb{E}(S_{\rho_m^x} f_n(x) \mid \mathcal{D}_X)\big) = \mathrm{Var}\big(\mathbb{E}(\tfrac{1}{m}\sum_{i=1}^{m} f_n(g^i \cdot x) \mid \mathcal{D}_X)\big) \tag{4.63}$$

$$= \mathrm{Var}\big(\tfrac{1}{m}\sum_{i=1}^{m} \mathbb{E}(f_n(g^i \cdot x) \mid \mathcal{D}_X)\big) \tag{4.64}$$

$$\leq \sup_{i \in [m]} \mathrm{Var}\big(\mathbb{E}(f_n(g^i \cdot x) \mid \mathcal{D}_X)\big) \tag{4.65}$$

For ease, let T_i be the random variable $\mathbb{E}(f_n(g^i \cdot x) \mid \mathcal{D}_X)$. Note that if $N_{g^i \cdot x,h} > 0$ then $T_i \in [f(x) - B(x)h^\beta, f(x) - B(x)h^\beta]$ by assumption (E3) and otherwise $T_i = 0$. Thus we can use assumption (E2) and Popoviciu's inequality for bounded random variables [Sharma et al., 2010] to find:

$$\mathrm{Var}(T_i) = \mathbb{E}(T_i^2 \mid \Xi)\mathbb{P}(\Xi_i) + \mathbb{E}(T_i^2 \mid \Xi_i^C)(1 - \mathbb{P}(\Xi_i)) - (\mathbb{E}(T_i \mid \Xi_i)\mathbb{P}(\Xi_i) + \mathbb{E}(T_i \mid \Xi_i^C)(1 - \mathbb{P}(\Xi_i)))^2 \tag{4.66}$$

$$= 0 \times \mathbb{P}(\Xi_i) + \mathbb{E}(T_i^2 \mid \Xi_i^C)(1 - \mathbb{P}(\Xi_i)) - (0 \times \mathbb{P}(\Xi_i) + \mathbb{E}(T_i \mid \Xi_i^C)(1 - \mathbb{P}(\Xi_i)))^2 \tag{4.67}$$

$$\leq \mathrm{Var}(T_i \mid \Xi_i^C) \tag{4.68}$$

$$\leq B^2 h^{2\beta} \tag{4.69}$$

Therefore we have the variance bound:

$$\mathrm{Var}(S_{\rho_m^x} f_n(x)) \leq \tfrac{1}{m} \sup_{i \in [m]} \left(\tfrac{2V}{np_{g^i \cdot x,h}} + V \exp(-np_{g^i \cdot x,h}/8) + mB^2 h^{2\beta} \right) \tag{4.70}$$

as required. □

Proof of Proposition 33. The pointwise error of the partially symmetrised estimator has:

$$\mathbb{E}\big((S_{\rho_m^x} f_n(x) - S_{\rho_m^x} f(x))^2\big) = \big(\mathbb{E}(S_{\rho_m^x} f_n(x)) - S_{\rho_m^x} f(x)\big)^2 + \mathrm{Var}\big(S_{\rho_m^x} f_n(x)\big) \tag{4.71}$$

$$\leq B^2 h^{2\beta} + \sup_{i \in [m]} \exp(-np_{g^i \cdot x,h}) \|f\|_\infty + \tag{4.72}$$

$$\tfrac{1}{m} \sup_{i \in [m]} \left(\tfrac{2V}{np_{g^i \cdot x,h}} + V \exp(-np_{g^i \cdot x,h}/8) + mB^2 h^{2\beta} \right) \tag{4.73}$$

Now with the choice of bandwidth $h = an^{-1/(2\beta + d - d^G)}$, the bound on m from Proposition 31, and the distribution property $\mu(B_{\mathcal{X}}(x,h)) \in [c_x h^d, C_x h^d]$ we have that:

$$h^{2\beta} \leq a^{2\beta} \phi_n^{\beta, d - d^G} \tag{4.74}$$

and

$$(mnp_{g^i \cdot x,h})^{-1} \leq \frac{(2h)^{d^G}}{(R_x^{[x]_U})^{d^G} n c_{g^i \cdot x} h^d} \tag{4.75}$$

$$= \frac{2^{d^G}}{c_{g^i \cdot x} (R_x^{[x]_U})^{d^G}} \times \phi_n^{\beta, d - d^G} \tag{4.76}$$

Thus we can substitute these back into equation (4.73) to find:

$$\mathbb{E}\big(S_{\rho_m^x} f_n(x) - S_{\rho_m^x} f(x))^2\big) \leq \sup_{i\in[m]} \big(2B^2 a^{2\beta} + + \frac{2^{1+d^G} V}{c_{g^i\cdot x}(R_x^{[x]_U})^{d^G}}\big)\phi_n^{\beta,\dim\mathcal{X}-d^G} + \tag{4.77}$$

$$(\|f\|_\infty + V)\exp(-np_{g^i\cdot x,h}/8) \tag{4.78}$$

$$\leq \sup_{g\in U} 2\big(B(g\cdot x)^2 a^{2\beta} + \frac{2^{d^G} V}{c_{g\cdot x}(R_x^{[x]_U})^{d^G}}\big)\phi_n^{\beta,\dim\mathcal{X}-d^G} + \tag{4.79}$$

$$(\|f\|_\infty + V)\exp\big(-\tfrac{c_{g\cdot x} a^d}{8} n^{\frac{2\beta-d^G}{2\beta+\dim\mathcal{X}-d^G}}\big) \tag{4.80}$$

$$\leq \sup_{g\in U}\Big(2\big(B^2 a^{2\beta} + \frac{2^{d^G} V}{c_{g\cdot x}(R_x^{[x]_U})^{d^G}}\big) + \tag{4.81}$$

$$(\|f\|_\infty + V)\big(\tfrac{16\beta}{c_{g\cdot x} a^d(2\beta-d^G)}\big)^{-\frac{2\beta}{d}}\Big)\phi_n^{\beta,\dim\mathcal{X}-d^G} \tag{4.82}$$

The result follows from noting that $\sup_g c_{g\cdot x}^{-1} = 1/\inf_g c_{g\cdot x}$ and that $\|f\|_\infty \leq L$ for all $f \in \mathcal{F}(L,\beta)$. □

Proof of Theorem 34. See that the pointwise bound of Proposition 33 can be used with the conditional expectation tower law to find:

$$\mathbb{E}((S_{\rho_m} f_n(X) - S_{\rho_m} f(X))^2 \mid X\in\Omega) = \mathbb{E}(\mathbb{E}((S_{\rho_m^X} f_n(X) - S_{\rho_m^X} f(X))^2 \mid X) \mid X\in\Omega) \tag{4.83}$$

$$\leq \mathbb{E}(C_{G,X} \mid X\in\Omega)\phi_n^{\beta,d-d^G} \tag{4.84}$$

Lastly we check the integrability of $C_{G,X}$. See that:

$$\mathbb{E}(C_{G,X} \mid X\in\Omega) = \mathbb{E}\Big(2\big(B^2 a^{2\beta} + \frac{2^{d^G} V}{c_\star(X)(R_X^{[x]_U})^{d^G}}\big) + (L+V)\big(\tfrac{c_\star(X) a^d(2\beta-d^G)}{16\beta}\big)^{\frac{2\beta}{d}} \;\Big|\; X\in\Omega\Big) \tag{4.85}$$

$$= 2B^2 a^{2\beta} + 2\mathbb{E}\big(\tfrac{2^{d^G} V}{c(R_X^{[x]_U})^{d^G}}\big) + (L+V)\mathbb{E}\Big(\big(\tfrac{c_\star(X) a^d(2\beta-d^G)}{16\beta}\big)^{\frac{2\beta}{d}} \;\Big|\; X\in\Omega\Big) \tag{4.86}$$

$$\leq 2B^2 a^{2\beta} + 2\big(\tfrac{2^{d^G} V}{c}\big)\mathbb{E}((R_X^{[x]_U})^{-d^G} \mid X\in\Omega) + (L+V)\Big(\big(\tfrac{c a^d(2\beta-d^G)}{16\beta}\big)^{\frac{2\beta}{d}}\Big) \tag{4.87}$$

Assumption (Q2) and its conditioned hypothesis in this theorem then control the expectation term, giving the integrability of $C_X \mid X\in\Omega$. □

Proof of Corollary 35. First, note that using the shift invariance of the Haar measure we have that if $g\sim U(G)$, then $gg^i \overset{D}{=} g\sim U(G)$ for all g^i, so

$$S_G S_{\rho_m^x}\phi(y) = m^{-1}\sum_{i=1}^m \mathbb{E}(\phi((gg^i)\cdot y) \mid \mathcal{D}) = m^{-1}\sum_{i=1}^m \mathbb{E}(\phi(g\cdot y) \mid \mathcal{D}) = S_G\phi(y) \tag{4.88}$$

for all $\phi \in L^2(\mathcal{X})$. Thus we can say that almost surely:

$$\|S_G f_n - S_G f\|_2 = \|S_G S_{\rho_m} f_n - S_G S_{\rho_m} f\|_2 \leq \|S_{\rho_m} f_n - S_{\rho_m} f\|_2 \tag{4.89}$$

using the fact that S_G is a projection with operator norm 1 in the last inequality. Lastly, we can then take expectations and use Proposition 34 to get the stated result. □

Proof of Corollary 36. The estimator $\hat{f}_{n,G}$ is a Monte-Carlo estimate of $S_G f_n$, so we have:

$$\mathbb{E}\|\hat{f}_{n,G} - S_G f\|_2^2 \leq 2\mathbb{E}\|\hat{f}_{n,G} - S_G f_n\|_2^2 + 2\mathbb{E}\|S_G f_n - S_G f\|_2^2 \tag{4.90}$$

using the sum of squares inequality $(a+b)^2 \leq 2a^2 + 2b^2$ (Lemma 69). The first term describes the Monte-Carlo error of the approximation $\hat{f}_{n,G} \approx S_G f_n$, which is bounded:

$$\begin{aligned}
\mathbb{E}\|\hat{f}_{n,G} - S_G f_n\|_2^2 &= \mathbb{E}\big((\hat{f}_{n,G}(X) - S_G f_n(X))^2\big) && (4.91)\\
&= \mathbb{E}\big(\mathbb{E}((\hat{f}_{n,G}(X) - S_G f_n(X))^2 \mid \mathcal{D}, X)\big) && (4.92)\\
&= \mathbb{E}\big(\mathrm{Var}(\hat{f}_{n,G}(X) \mid \mathcal{D}, X)\big) && (4.93)\\
&= \tfrac{1}{M}\mathbb{E}\big(\mathrm{Var}(f_n(g \cdot X) \mid \mathcal{D}, X)\big) && (4.94)\\
&= \tfrac{1}{M}\mathbb{E}\big((f_n(g \cdot X) - S_G f_n(X))^2\big) && (4.95)\\
&\leq \tfrac{3}{M}\mathbb{E}\big(\|f_n \circ (g\cdot) - f \circ (g\cdot)\|_2^2 + \|S_G f_n - S_G f\|_2^2 + \|(S_G f) \circ (g\cdot) - f \circ (g\cdot)\|_2^2\big) && (4.96)\\
&\leq \tfrac{6}{M}(C\phi_n^{\beta,d} + \|f\|_2^2) && (4.97)\\
&\leq \tfrac{6C+6L^2}{M} && (4.98)
\end{aligned}$$

using the fact that $S_G f_n = \mathbb{E}(\hat{f}_{n,G}(X) \mid \mathcal{D}, X)$, the independence of the g_i samples, the sum of squares inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, the G-invariance of μ, and lastly the bounds Assumption (E3) and the Hölder ball bound on $\|f\|_2^2$. Thus we have:

$$\begin{aligned}
\mathbb{E}(\|\hat{f}_{n,G} - S_G f\|_2^2) &\leq \tfrac{12}{M}(C + L^2) + 2C_G\phi_n^{\beta,d-d^G} && (4.99)\\
&\leq (12C + 12L^2 + 2C_G)\phi_n^{\beta,d-d^G} && (4.100)
\end{aligned}$$

when $M = n$, as required. □

4.6.2 Proof in Section 4.3

Proof of Proposition 39. Since f is continuous, we know that $f = f_0$ in the definition of invariance. Thus for any such $f \in \mathcal{F}(L, \beta)$, $x \in \mathcal{X}$, and $g \sim \rho_G^x$ we have:

$$\begin{aligned}
((S_{\rho_G^x} f)(x) - f(x))^2 &= \mathbb{E}(f(x) - f(g \cdot x))^2 && (4.101)\\
&= \mathbb{E}(\inf_{h \in H \cap U} f(h \cdot x) - f(g \cdot x))^2 && (4.102)\\
&\leq L^2 \mathbb{E}(\inf_{h \in H \cap U} d(h \cdot x, g \cdot x)^\alpha)^2 && (4.103)\\
&\leq L^2 (\sup_{g \in G \cap U} \inf_{h \in H \cap U} d(h \cdot x, g \cdot x)^\alpha)^2 && (4.104)\\
&\leq L^2 \sup_{g \in G \cap U} \inf_{h \in H \cap U} (L_{\mathcal{G}} d_{\mathcal{G}}(g, h))^{2\alpha} && (4.105)\\
&\leq L^2 L_{\mathcal{G}}^{2\alpha} d_{\mathrm{Haus}(U)}(G, H)^{2\alpha} && (4.106)
\end{aligned}$$

giving the required inequality. □

Proof of Theorem 41. Suppose that $\dim[x]_{G_{\max}(f,\mathcal{G})} = D$. Let G_i^D be any group in K^D with $d_{\mathrm{Haus}}(G_{\max}(f, \mathcal{G}), G_i^D) \leq \delta$ (of which there must be at least one). Then we have:

$$\begin{aligned}
\|S_{\rho_{\hat{G}_\delta}} f_n - f\|_2^2 &= \hat{\mathcal{E}}(S_{\rho_{\hat{G}_\delta}} f_n) + (\|S_{\rho_{\hat{G}_\delta}} f_n - f\|_2^2 - \hat{\mathcal{E}}(S_{\rho_{\hat{G}_\delta}} f_n)) && (4.107)\\
&\leq \hat{\mathcal{E}}(S_{\rho_{G_i^D}} f_n) + (\|S_{\rho_{\hat{G}_\delta}} f_n - f\|_2^2 - \hat{\mathcal{E}}(S_{\rho_{\hat{G}_\delta}} f_{n,h})) && (4.108)\\
&= \|S_{\rho_{G_i^D}} f_n - f\|_2^2 + (\hat{\mathcal{E}}(S_{\rho_{G_i^D}} f_n) - \|S_{\rho_{G_i^D}} f_n - f\|_2^2) + (\|S_{\rho_{\hat{G}_\delta}} f_n - f\|_2^2 - \hat{\mathcal{E}}(S_{\rho_{\hat{G}_\delta}} f_n)) && (4.109)
\end{aligned}$$

Taking expectations (over both the original data used for estimating f_n and the independent copy used for the calculating $\hat{\mathcal{E}}$) then gives:

$$R(S_{\hat{G}_\delta} f_n) \leq R(S_{G_i^D} f_n) + \sigma^2 - \sigma^2 \tag{4.110}$$

So we need only consider the properties of the group that best approximates $G_{\max}(f, \mathcal{G})$. We have:

$$\begin{aligned}
R(S_{G_i^D} f_n) &= \mathbb{E}(\|S_{\rho_{G_i^D}} f_n - f\|_2^2) && (4.111)\\
&\leq 2\mathbb{E}(\|S_{\rho_{G_i^D}} f_n - S_{\rho_{G_i^D}} f\|_2^2) + 2\|S_{\rho_{G_i^D}} f - f\|_2^2 && (4.112)\\
&\leq 2C_{G_i^D} \phi_n^{\beta, d - d^{G_i^D}} + 2 \int_{\mathcal{X}} (S_{\rho^x_{G_i^D}} f(x) - f(x))^2 \, \mathrm{d}\mu(x) && (4.113)\\
&\leq 2C_{G_i^D} \phi_n^{\beta, d - d^{G_i^D}} + 2 \int_{\mathcal{X}} L^2 L_{\mathcal{G}}^{2\alpha} d_U(G_i^D, G_{\max}(f, \mathcal{G})^{2\alpha}) \, \mathrm{d}\mu(x) && (4.114)\\
&\leq 2C_{G_i^D} \phi_n^{\beta, d - d^{G_i^D}} + 2L^2 L_{\mathcal{G}}^{2\alpha} \delta^{2\alpha} && (4.115)
\end{aligned}$$

where $\alpha = \min(\beta, 1)$. This comes from the sum of squares inequality $(a+b)^2 \leq 2a^2 + 2b^2$ (Lemma 69), Theorem 34, Proposition 39, and lastly the δ-closeness of G_i^D to $G_{\max}(f, \mathcal{G})$. Now we have:

$$2L^2 L_{\mathcal{G}}^{2\min(\beta,1)} \delta^{2\min(\beta,1)} \leq 2L^2 L_{\mathcal{G}}^{2\min(\beta,1)} (L_{\mathcal{G}}^{-1}\big(\frac{\phi_n^{\mathcal{G}}}{2L^2}\big)^{1/2\min(\beta,1)})^{2\min(\beta,1)} \tag{4.116}$$

$$\leq \phi_n^{\beta, d-d^{\mathcal{G}}} \tag{4.117}$$

$$\leq \phi_n^{\beta, d-d^{G\max(f,\mathcal{G})}} \tag{4.118}$$

Therefore we have:

$$R(S_{\rho_{\hat{G}_\delta}} f_n) \leq 2C_G \phi_n^{\beta, d-d^{G\max(f,\mathcal{G})}} + 2\phi_n^{\beta, d-d^{G\max(f,\mathcal{G})}} \tag{4.119}$$

$$\leq 2(\sup_{G \in K(\mathcal{G})} C_G + 1)\phi_n^{\beta, d-d^{G\max(f,\mathcal{G})}} \tag{4.120}$$

as required. □

Proof of Theorem 43. Let G_i^D be the minimiser of $d(G_i^D, G_{\max}(f, \mathcal{D}))$ over $\cup_\ell \{G_i^\ell\}$. First, we have the minimisation inequality:

$$\mathbb{E}\big((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big) \tag{4.121}$$

$$= \hat{\mathcal{E}}_\Omega(S_{\rho_{\hat{G}_\delta(\Omega)}} f_n) + \mathbb{E}\big((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big) - \hat{\mathcal{E}}_\Omega(S_{\rho_{\hat{G}_\delta(\Omega)}} f_n) \tag{4.122}$$

$$\leq \hat{\mathcal{E}}_\Omega(S_{\rho_{G_i^D}} f_n) + \mathbb{E}\big((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big) - \hat{\mathcal{E}}_\Omega(S_{\rho_{\hat{G}_\delta(\Omega)}} f_n) \tag{4.123}$$

$$= \mathbb{E}((S_{\rho_{G_i^D}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega) + \tag{4.124}$$

$$\hat{\mathcal{E}}_\Omega(S_{\rho_{G_i^D}} f_n) - \mathbb{E}\big((S_{\rho_{G_i^D}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big) + \tag{4.125}$$

$$\mathbb{E}\big((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big) - \hat{\mathcal{E}}_\Omega(S_{\rho_{\hat{G}_\delta(\Omega)}} f_n) \tag{4.126}$$

Now conditioned on the event Ξ' that Ω contains at least one X_i', we have:

$$\mathbb{E}(\hat{\mathcal{E}}_\Omega(S_{\rho_G} f_n) \mid f_n, \Xi') = \sigma^2 + \mathbb{E}\big((S_{\rho_G} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big) \tag{4.127}$$

for all subgroups $G \in \cup_\ell \{G_i^\ell\}$. In the complementary event Ξ'^C, we have $\hat{G}_\delta(\Omega) = I$ and so:

$$\mathbb{E}(\hat{\mathcal{E}}_\Omega(S_{\rho_G} f_n) \mid f_n, \Xi'^C) = L \tag{4.128}$$

Thus the integrated error splits over these events as:

$$\mathbb{E}\big((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid X \in \Omega\big) \tag{4.129}$$

$$= \mathbb{E}\Big(\mathbb{E}\big((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big)\Big) \tag{4.130}$$

$$= \mathbb{E}\Big(\mathbb{E}\big((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big) \mid \Xi_i\Big)\mathbb{P}(\Xi_i) + \tag{4.131}$$

$$\mathbb{E}\Big(\mathbb{E}\big((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega\big) \mid \Xi_i^C\Big)\mathbb{P}(\Xi_i^C) \tag{4.132}$$

$$\leq \Big(\mathbb{E}((S_{\rho_{G_i^D}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega) + \sigma^2 - \sigma^2\Big) \tag{4.133}$$

$$\Big(\mathbb{E}((S_{\rho_{\hat{G}_\delta(\Omega)}} f_n(X) - f(X))^2 \mid f_n, X \in \Omega) + L - L\Big)\mathbb{P}(\Xi_i'^C) \tag{4.134}$$

$$\leq 2C_{G_i^D,\Omega}\phi_n^{\beta,d-D} + 2\mathbb{E}((S_{\rho_{G_i^D}} f(X) - f(X))^2 \mid X \in \Omega) + \tag{4.135}$$

$$\sup_{G \in K(\mathcal{G})} (2C_{G,\Omega}\phi_n^{\beta,d} + 2\mathbb{E}((S_{\rho_G} f(x) - f(X))^2 \mid X \in \Omega))\mathbb{P}(\Xi'^C) \tag{4.136}$$

$$\leq 2C_{G_i^D,\Omega}\phi_n^{\beta,d-D} + 2L^2 L_{\mathcal{G}}^{2\alpha}\delta^{2\alpha} + \sup_{G \in K(\mathcal{G})} (2C_{G,\Omega}\phi_n^{\beta,d} + 4L)\mathbb{P}(\Xi'^C) \tag{4.137}$$

$$\leq 2C_{G_i^D,\Omega}\phi_n^{\beta,d-D} + 2\phi_n^{\beta,d-\dim[x]_{\mathcal{G}}} + \sup_{G \in K(\mathcal{G})} (2C_{G,\Omega} + 4L)\mathbb{P}(\Xi'^C) \tag{4.138}$$

using the sum of squares inequality (Lemma 69), the bounds from section 4.2, the fact that the rates $\phi_n^{\beta,d-\ell}$ are faster than $\phi_n^{\beta,d}$ for all ℓ, the bound of Proposition 39, and the choice of δ. Lastly note that:

$$\mathbb{P}(\Xi'^C) = (1 - \mu(\Omega))^n \leq \exp(-n\mu(\Omega)) \tag{4.139}$$

thus using Lemma 70 gives the result. □

Chapter 5

Estimation of Maximal Symmetries via Subgroup Lattices

In this chapter we provide a method for symmetry estimation that utilising the lattice structure of the closed subgroups of $\mathcal{G}$. This method relies on some hypothesis tests for symmetry we construct, and so leads to natural inferential properties of the estimate of $G_{\max}$. This is based on the work in the paper Christie and Aston [2023].

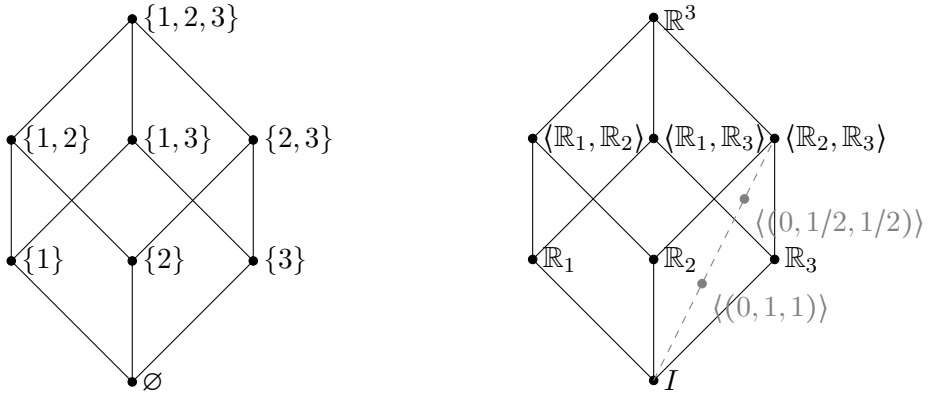
5.1 Introduction

In this chapter we present an alternative estimator for the maximal symmetry of f that avoids estimating f direction. This avoids much of the computational load of the M-estimator in the previous chapter. This estimator relies on a consideration of an inferential framework for the symmetries of f. Specifically, we will construct hypothesis tests for $H_0(G) : f$ *is* G*-invariant* and then consider how these tests can be used to find a confidence region in the lattice of subgroups of f.

This framework is an explicit generalisation of some techniques in variable subset selection [Mallows, 1973, Berk, 1978]. Consider a linear regression function $f : \mathbb{R}^p \to \mathbb{R}$ given by $f(X) = X^T\beta$ for which some features $i \in I$ have $\beta_i = 0$, where β_i are the regression coefficients in β. We could estimate I by considering subsets $A \subseteq \{1, \ldots, p\}$ and testing whether $\beta_A = 0$ (where the set subscript indicates the vector $(\beta_a : a \in A)$). These subsets can be structured in a *lattice* (see Section 5.2.2 and Figure 5.1), a directed graph with a node for each subset and an edge from $A \to B$ if $B \subseteq A$ (though we omit the directions from the figures as they always point down in this chapter). This is important because it shows how we can avoid running tests for some of the subsets A. If we test at the bottom of the lattice first, any rejection immediately tells us about the larger subsets containing the smaller subset so we can avoid testing for invariance to these subsets unnecessarily.

This linear example is in fact an explicit special case of the symmetries considered here. In this case the domain is $\mathcal{X} = \mathbb{R}^p$, and the symmetries are translations $\Phi_{\mathbb{T}} = \{x \mapsto x + u : u \in \mathbb{R}^p\}$. If f is invariant to translations along a particular axis (in the standard orthonormal basis $\{e_i\}$) then $\beta(X + ae_i) = \beta X$ for all $a \in \mathbb{R}$, which is equivalent to $\beta_i = 0$. This shows that variable selection in linear regression is precisely a symmetry estimation problem. There are of course many other sub-symmetries of this translation action, for example translations along the line from the origin through $(1, \ldots, 1)$. This is the idea of sufficient dimension reduction [Li, 1991], which is also a special case of this symmetry estimation problem.

As in the previous section, the question we seek to answer is "Given a large set of transformations Φ, what is the largest subset Ψ of transformations Ψ that satisfy $f = f \circ \psi$ for all $\psi \in \Psi$ for our particular regression f?". In the special cases above, Φ is either finite (as in variable selection) or linear (as in sufficient dimension reduction), but in our case it can be infinite and non-linear. To answer this, we will use the lattice structure of these subgroups to allow for the lattice testing regime as above. A lattice is a partially ordered set such that every subset has a unique supremum and infimum. The algebraic properties of the lattice of subgroups (where the ordering is given by subgrouping) allows us to combine the information in a collection of hypothesis tests for $H_0(\Phi_i) : f$ *is* Φ_i*-invariant* for some chosen subgroups $\Phi_i \leq \Phi$ into an confidence region of possible Φ. We can then take our estimate of the true maximal symmetry as the unique largest symmetry in this confidence region.



(a) Lattice of subsets $A \subseteq \{1, 2, 3\}$. (b) Lattice of (some) subgroups of $\mathbb{R}^p$.

Fig. 5.1 Variable subset selection lattice of subsets, and symmetry estimation lattice of subgroups. The symmetry here is the translation action of $\Phi_{\mathbb{T}}$ on $\mathbb{R}^3$, and we have taken a finite sub-lattice of the lattice of closed subgroups where $\mathbb{R}_i$ is the translation group along the i^{th} orthonormal basis vector of $\mathbb{R}^3$. We have also included grey and dashed edges to other symmetries such as $\{x \mapsto x + a(e_2 + e_3) : a \in \mathbb{Z}\}$ and the larger symmetry $\{x \mapsto x + (a/2)(e_2 + e_3) : a \in \mathbb{Z}\}$.

Thus in this chapter we construct tests for the hypotheses $H_0^{(i)} : f$ *is* Φ_i*-invariant* for certain subgroups $\Phi_i \leq \Phi$, and show how to borrow strength from these tests to make inferences about other subgroups of Φ. The tests we present require mild technical assumptions and trade theoretical guarantees, computational time, and statistical power. We show that these tests lead to a natural estimate of the maximal symmetry of f that can be used as in the previous chapter to adaptively smooth any regressor.

We describe in Section 5.2 how to structure the subgroups of the search group $\mathcal{G}$ into a lattice, describe confidence regions in this lattice, and give rules for how invariance behaves over this lattice. In Section 5.3 we then describe the estimation algorithm that searches for $G_{\max}(f, \mathcal{G})$ by testing the invariance of f to certain nodes of the lattice, and show how we can borrow information across tests to improve the estimation. Section 5.4 introduces two tests one can use in the estimation algorithm, each requiring different assumptions for the statistical problem. We then demonstrate the performance of the tests used, the symmetry estimator, and its use in nonparametric regression in simulation in Section 5.5. We then apply these methods to satellite measurements of the Earth's magnetic field in Section 5.6. All proofs can be found in the appendices, along with additional simulated and synthetic numerical results.

5.1.1 Main Contributions

We now summarise the technical contributions of this paper. All terms defined in later sections are italicized here. Let $\mathcal{G}$ be a group acting continuously on $\mathcal{X}$ and let f be any Hölder regression function. We show that:

1) The set of closed subgroups of $\mathcal{G}$ forms a *lattice* with *joins* given by the closure of the group generated by the two elements and *meets* given by intersections, which differs slightly from the usual subgroup lattice which also contains non-closed subgroups and has a different join operation;

2) We define *confidence regions of lattices* and give a loss function that measures the difference between the correct region of applicable symmetries and the inferred confidence region;

3) We define *lattice bases* and show that they allow us to reduce the number of tests conducted for specific invariances;

4) We develop a method for estimating the maximal invariant subgroup of f and for inferring the confidence region for the maximal invariant subgroups of the regression function within the chosen subgroup lattice;

5) We show that under suitable conditions on the hypothesis tests used in this method that the estimator is consistent; and

6) We present two hypothesis tests for the hypothesis $H_0(G) : f$ *is* G*-invariant*. The first, the *Asymmetric Variation Test*, is shown to be a consistent hypothesis test. The permutation variant of this test then allows us to make fewer assumptions on the noise distribution and the function class that f belongs to.

5.2 Inference for Subgroups

We now describe the structure used to do inference on the subgroups of some group $\mathcal{G}$. The first step is to structure the subgroups into a lattice. This is a well known mathematical result, see Schmidt [2011], Sankappanavar and Burris [1981]. We then show that the set of closed subgroups of $\mathcal{G}$ also form a lattice, which is not always a sub-lattice of the lattice of all subgroups. With this structure we are then able to ask the first statistical question: how can one do estimation and inference on a lattice of subgroups? We also show that we can reduce the lattice to the information stored in a particular set of subgroups, which is used to minimise the number of tests conducted for both statistical and computational benefit.

5.2.1 Lattices

A **lattice** L is a partially ordered set with unique supremums and infimums. Given two elements $A, B \in L$ the unique supremum is called their **join** and is written $A \vee B$, and their unique infimum, **meet**, is written $A \wedge B$. Joins and meets are associative, commutative, idempotent binary operations that satisfy the **absorption laws** $a \wedge (a \vee b) = a$ and $a \vee (a \wedge b) = a$. We say that a subset of a lattice L is a **sub-lattice** of L if it is closed under the joins and meets of the original lattice. Given some set $A \subseteq L$, we can take $\langle\langle A \rangle\rangle$ as the smallest sub-lattice of L that contains A and say that A **generates** $\langle\langle A \rangle\rangle$. We can arrange a lattice L into a **Hasse diagram**, a directed graph with nodes given by the elements $A, B \in L$ and edges from A to B if $B \leq A$, and there is no $C \in L$ with $B < C < A$. We typically omit the arrows on the edges in this chapter as they all point down.

Example 47. *In the context of variable selection we have a lattice of subsets of the variable indices, where the index subsets* $A, B \subseteq \{1, \ldots, d\}$ *have joins* $A \vee B = A \cup B$ *and meets* $A \wedge B = A \cap B$.

5.2.2 Subgroup Lattices

The subgroups of a group $\mathcal{G}$ form a lattice with the order given by subgrouping [Schmidt, 2011, Sankappanavar and Burris, 1981]. We write $L(\mathcal{G})$ for this lattice of subgroups. In this lattice the join is given by the groups generated by unions, i.e., $A \vee B = \langle A, B \rangle$ and meet by intersections, i.e., $A \wedge B = A \cap B$, for $A, B \in L$. Note that in the group context we can't simply take unions as these are not guaranteed to be groups, but the idea is essentially taking a union and adding in all the elements needed to form a group.

Example 48. *Consider the* ***dihedral group*** *of symmetries of the square D_4 (see section 2.3.1). This group has ten subgroups, including itself and the trivial group I. These are shown in the table of Figure 5.2 These can be ordered into the lattice in Figure 5.2, represented as a Hasse diagram.*

Subgroup	Elements
D_4	$\{I, R_{\pi/2}, R_\pi, R_{3\pi/2}, R_h, R_v, R_/, R_\backslash\}$
$\langle R_h, R_\pi\rangle$	$\{I, R_h, R_v, R_\pi\}$
$\langle R_{\pi/2}\rangle$	$\{I, R_{\pi/2}, R_\pi, R_{3\pi/2}\}$
$\langle R_/, R_\pi\rangle$	$\{I, R_\backslash, R_/, R_\pi\}$
$\langle R_h\rangle$	$\{I, R_h\}$
$\langle R_v\rangle$	$\{I, R_v\}$
$\langle R_\pi\rangle$	$\{I, R_\pi\}$
$\langle R_/\rangle$	$\{I, R_/\}$
$\langle R_\backslash\rangle$	$\{I, R_\backslash\}$
I	$\{I\}$

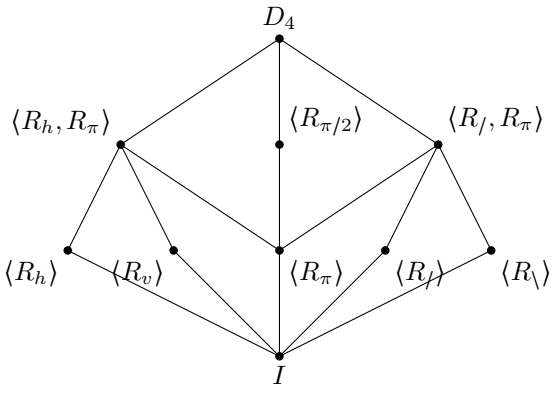


Fig. 5.2 The list of subgroups and their elements of the dihedral group D_4 (see section 2.3.1), and the Hasse diagram of the subgroup lattice. Each dot represents a subgroup, with edge from A down to B if $A \leq B$ and there is no other group that can be ordered between them.

One important subset of $L(\mathcal{G})$ is the collection of *closed* subgroups (closed in the topology of $\mathcal{G}$), which we write $K(\mathcal{G})$ here. These do not necessarily form a sub-lattice of $L(\mathcal{G})$ for arbitrary $\mathcal{G}$ as we can generate non closed subgroups with the original join operator $\langle G, H\rangle$ (consider two copies of high order finite rotations around distinct axes generating a proper dense subgroup of $SO(3)$). It does form a lattice as it still has unique supremums (just distinct ones from the supremum in L) and infimums, where the join is the topological closure of the group generated from G and H, i.e., $\overline{\langle G, H\rangle}$. This is established in the following proposition. In some cases (e.g. finite $\mathcal{G}$) this is equivalent to the original join, so the lattice in Example 48 is both $L(D_4)$ and $K(D_4)$.

Proposition 49. *Let $\mathcal{G}$ be a topological group. The closed subgroups $K(\mathcal{G})$ form a lattice with partial ordering given by subgrouping, meets given by intersection, and joins given by the topological closure of the group generated by the two subgroups.*

Example 50. *Consider the circle group $S^1 = \{z \in \mathbb{C} : |z| = 1\}$. The finite subgroups are those of rational rotations, and are of the form $C_a = \langle \exp(2\pi i/a)\rangle$ of order a. These have subgroups corresponding to the factors of a, and so the lowest parts of the subgroup lattice of S^1 are the groups of prime order. Above these you have groups with two (counting with repetition) prime factors, and so on. There are also subgroups of countable infinite order, for example $\langle \exp(i)\rangle$, and one subgroup of uncountable order - itself. Importantly, all of these infinite*

subgroups are all dense in S^1 (under the subset topology from $\mathbb{C}$). Thus the key difference between $L(S^1)$ and $K(S^1)$ is that these dense subgroups are missing.

5.2.3 Confidence Regions in Lattices

As statistical objects, groups typically have relationships characterised by inclusions - if a symmetry G applies in a particular situation then any subgroup $H \leq G$ surely must also apply. Therefore we are typically interested in confidence sets that contain all of their subgroups. We formalise this below, and give a measure of the optimality of these confidence sets.

Suppose we wish to estimate a group object G_0 within some lattice L, and that the lattice has a measure ν. We say that a region $\hat{R} \subseteq L$ is a $(1-\alpha)$-**confidence region** for G_0 if it is a measurable function of the data, contains all subgroups of all of its elements, and

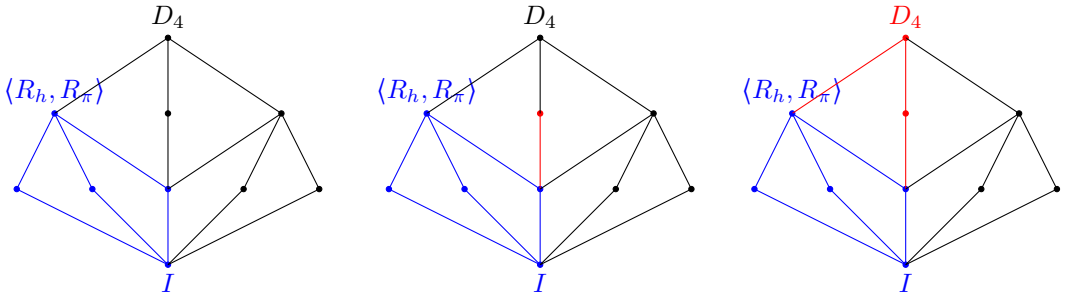
$$\mathbb{P}_f(G_0 \in \hat{R}) \geq 1 - \alpha. \tag{5.1}$$

The ideal confidence region would be the set of subgroups of G, which we denote A, so we judge the optimality of the region $\hat{R}$ by the measure of the groups that are different to A:

$$\mathcal{E}(\hat{R}) = \nu(\{G \in \hat{R} : G \notin A\} \cup \{G \in A : G \notin \hat{R}\}) = \nu\big((\hat{R} \cup A) \setminus (\hat{R} \cap A)\big) \tag{5.2}$$

where we assume that each of these is a measurable set. In our method later each lattice L will be finite and ν will be the discrete counting measure, so this will be true.

Example 51. *Consider the group D_4 and its subgroup $G_0 = \langle R_h, R_\pi \rangle$. In Figure 5.3, we show three examples and non-examples of possible samples of confidence regions for G_0, and show the excess over A.*



(a) $\mathcal{E}(\hat{R}_1) = 0$ (b) $\mathcal{E}(\hat{R}_2) = 1$ (c) Not a confidence region

Fig. 5.3 Two examples of confidence regions for the subgroup $G_0 = \langle R_h, R_\pi \rangle$ of D_4 shown in blue with excess in red. Also shown is an example of a subset that is not a confidence region, which fails because it does not include all of the subgroups of its elements. The errors when ν is the counting measure are given underneath each picture.

5.2.4 Lattice Bases

Lastly we require a structure theorem for lattices, giving a condition on a set of subgroups that simplifies the structure of the lattice we will work over. Importantly, it allows us to extrapolate from ℓ tests to a lattice of size up to 2^ℓ. This is a huge benefit of considering the group structure of the transformations.

We say that a set of closed subgroups $\{G_i\}_{i=1}^{\ell}$ of $\mathcal{G}$ satisfies the **intersection property** (I) if $G_i \cap G_j \in \{I, G_i, G_j\}$ whenever $i \neq j$. Any set of subgroups satisfying the intersection property will be called a **lattice base**.

Proposition 52. *Suppose the set of subgroups $\{G_i\}_{i=1}^{\ell}$ of $\mathcal{G}$ satisfies the intersection property (I). Then the smallest lattice (in the sub lattice ordering) containing $G_1, \ldots, G_\ell$ consists only of groups of the form $G_{i_1} \vee \cdots \vee G_{i_m}$ for some distinct $i_j \in [\ell]$, as well as the empty product I.*

This result allows us to identify the lattice $K = \langle\!\langle \{G_i\}_{i=1}^{\ell} \rangle\!\rangle$ with subsets of $\{1, \ldots \ell\}$, but where some subsets represent the same group in the lattice.

Example 53. *In the variable selection lattice of Figure 5.1a, a lattice base is the singletons set $\{\{i\} : i \leq p\}$. Equivalently for the translation group framing in Figure 5.1b, a lattice base is the collection $\{\mathbb{R}_i : i \leq p\}$.*

Example 54. *Consider the group D_4. One lattice base that generates the entire lattice $L(D_4)$ is given by:*

$$B = \left\{ \langle R_h \rangle, \langle R_v \rangle, \langle R_\pi \rangle, \langle R_/ \rangle, \langle R_\backslash \rangle, \langle R_{\pi/2} \rangle \right\} \tag{5.3}$$

It is easy to check that all but one on the pairwise intersections are trivial; the only non-trivial intersection being $\langle R_{\pi/2} \rangle \cap \langle R_\pi \rangle = \langle R_\pi \rangle$.

If we restrict our consideration to a finite sublattice of $K(\mathcal{G})$ we might lose the maximal symmetry of f from the original lattice. The following result ensures that there is a still a unique maximal object within this restricted lattice.

Proposition 55. *Suppose that $\mathcal{G}$ acts faithfully and continuously on $\mathcal{X}$, and that f is continuous. Let K be a finite sub-lattice of $K(\mathcal{G})$ that contains the trivial group I. Then there exists a unique node $G_{\max}(f, K)$ in K that f is invariant to, and for which $H \leq G_{\max}(f, K)$ for all $H \in K$ where f is H-invariant.*

5.3 Estimation of Maximal Invariant Subgroups

In the context of our nonparametric regression problem we seek to estimate the unique maximal invariant subgroup for f of some search group $\mathcal{G}$ acting continuously and faithfully on the domain $\mathcal{X}$. We assume that f is continuous, so that this maximal subgroup is

closed. We will estimate $G_{\max}(f, \mathcal{G})$ by considering a sequence of finite sub lattices K_ℓ that approximate $K(\mathcal{G})$. Each finite sub-lattice K_ℓ will be generated by lattice bases $\{G_i\}_{i=1}^{\ell}$ such that $\{G_i\}_{i=1}^{\ell} \subseteq \{G_i\}_{i=1}^{\ell+1}$. Choices for this base are discussed in section 5.3.4.

Suppose that we have a test for the hypothesis $H_0^{(i)}$: *f is G_i-invariant* against the alternative $H_1^{(i)}$: *f is not G-invariant* for each of the groups G_i in the lattice base. We will provide two possibilities for such a test in the next section. We will test these hypotheses at significance levels α_i, such that $\sum_{i\in\mathbb{N}} \alpha_i = \alpha \in [0, 1]$. The choice of α_i will depend on how many terms of the lattice base we wish to test, and could either have finitely many non-zero terms or infinitely many, as long as the sum converges. Choices for α_i are discussed in section 5.3.5.

We then use the information of these tests to infer whether f is G-invariant for each $G \in K_\ell$ using the rules for invariance of Proposition 12. We will first describe the process when finitely many tests have been conducted, and provide theoretical and computational results for the algorithm to that point.

5.3.1 Estimating Over a Finite Lattice

Let $\{G_i\}_{i=1}^{\ell}$ be a lattice base, and consider the lattice $K_\ell = \langle\!\langle \{G_i\}_{i=1}^{\ell} \rangle\!\rangle$, which by Proposition 52 contains only groups of the form $G_{i_1} \vee \cdots \vee G_{i_m}$ for some distinct $i_j \in [\ell]$. The rules for invariance (Proposition 12) guarantee that knowledge of invariance of the lattice generators G_i gives knowledge of invariance at all other groups G in the lattice K_ℓ. Therefore after we test each of the hypotheses $H_0^{(i)}$, we can then define:

$$\hat{K}_\ell = \left\{ G \in K_\ell : \mathrm{Test}_{\alpha_i}(\mathcal{D}, G_i) = 1 \text{ for all } G_i \leq G \right\}. \tag{5.4}$$

where $\mathrm{Test}_{\alpha_i}(\mathcal{D}, G_i)$ is the test for the null hypothesis $H_0^{(i)}$ at significance level α_i, taking the value 0 if this hypothesis is rejected and 1 otherwise. This object can be computed (after testing) as long as we can compute the maps $\mathbf{1}_{\{G_i \leq G\}}$ for each $G \in K_\ell$. We describe this computation in section 5.3.2.

Proposition 56. *The region $\hat{K}_\ell$ is a $(1-\alpha)$-confidence region for $G_{\max}(f, K_\ell)$. In particular, for all $f \in \mathcal{F}$ we have:*

$$\mathbb{P}_f(G_{\max}(f, K_\ell) \in \hat{K}_\ell) \geq 1 - \sum_{i=1}^{\ell} \alpha_i \geq 1 - \alpha. \tag{5.5}$$

From this confidence region $\hat{K}_\ell$, we can take a point estimate $\hat{G}_\ell$ as any maximal element. This will usually be unique, but if not then it can be chosen via any distribution on this set (usually the uniform distribution).

5.3.2 Computation of $\hat{K}_\ell$

In Algorithm 1 we provide one possible method for computation of $\hat{K}_\ell$. The main requirement for this is the ability to compute the inclusions $G \leq H$ for $G, H \in K_\ell$. This algorithm has the advantage of being iterable; if we wish to compute $\hat{K}_{\ell+1}$ then we simply store the TestResults vector as well as $\hat{K}_\ell$ and add another term to the for loop.

Algorithm 1 Computation of $\hat{K}_\ell$

procedure ComputeKHat($\mathcal{D}$, $\{G_i\}_{i=1}^{\ell}$)
  Initialise $\hat{K}_0 = \{I\}$
  Initialise TestResults = () as the empty tuple
  for $i \in \{1, \dots \ell\}$ **do**
    Compute $\text{Test}_{\alpha_i}(\mathcal{D}, G_i)$ and append to TestResults
    Set $\hat{K}_i = \{\}$.
    if $\text{Test}_{\alpha_i}(\mathcal{D}, G_i) = 1$ **then**
      Set $\hat{K}_i = \hat{K}_{i-1} \cup \{G \vee G_i : G \in \hat{K}_{i-1}\}$
      for $j \in \{1, \dots, i-1\}$ **do**
        if $\text{TestResults}_j = 0$ **then**
          Set $\hat{K}_i = \{G \in \hat{K}_i : G_j \not\leq G\}$
    else
      Set $\hat{K}_i = \{G \in \hat{K}_{i-1} : G_i \not\leq G\}$
  Return $\hat{K}_\ell$

The computational complexity of this algorithm will depend on the complexity of the ℓ tests conducted. For the tests we propose in this paper, see sections 5.4.1 and 5.4.2 for their complexities. A naive search for $G_{\max}$ would conduct up 2^ℓ tests for every symmetry in K_ℓ. Since the testing is usually the main computation, reducing to $O(\ell)$ tests saves significant computational time, and also allows for each test to be conducted at higher significance levels improving the statistical performance of the estimate.

Aside from this, we require (in the worst case) potentially $O(\ell^2 2^\ell)$ checks of inclusion $G \leq H$ (because $\hat{K}_\ell$ can be of size $O(2^\ell)$ and we might have to check inclusion ℓ^2 times for each). This is entirely determined by the choice of lattice base and is independent of n and the dimension of $\mathcal{X}$.

5.3.3 Computation of $\hat{G}$

Once $\hat{K}_\ell$ has been computed, we need to select a point estimate of $G_{\max}(f, K(\mathcal{G}))$. This amounts to filtering out the elements that are not maximal and then sampling uniformly from the remainder.

Algorithm 2 Computation of $\hat{G}_\ell$

```
1: procedure ComputeGHat(K̂_ℓ)
2:     Set MaxElems = {I}
3:     for G ∈ K̂_ℓ do
4:         for H ∈ MaxElems do
5:             if G ≥ H then
6:                 Remove H from MaxElems
7:             if G < H then
8:                 Continue the G for loop
9:         Add G to MaxElems
10:    Return Ĝ sampled from MaxElems uniformly
```

Note that it is not a problem to remove H from the maximal element set in line 6, because it is not maximal and this will not affect whether any maximal elements are added later. The worst case computational time can be bounded trivially by $|\hat{K}_\ell| \times |\max \hat{K}_\ell| \leq 2^{2\ell}$, though it is likely that the true worst case performance is faster because of the trade off in the sizes between $\hat{K}_\ell$ and $\max \hat{K}_\ell$.

5.3.4 Choosing the Lattice Base

The choice of the lattice base $\{G_i\}_{i=1}^{\ell}$, is fundamental to the method. We want to cover large portions of $K(\mathcal{G})$ with the lattices K_ℓ for minimal numbers of tests conducted. Recall that the Hausdorff distance $d_{K(\mathcal{G})}$ between compact subgroups $G, H \leq \mathcal{G}$ is given by:

$$d_{K(\mathcal{G})}(G,H) = \max\left(\sup_{g\in G}\inf_{h\in H} d(g,h), \sup_{h\in H}\inf_{g\in G} d(h,g)\right) \tag{5.6}$$

when $\mathcal{G}$ has a metric d, as with the matrix groups we have used as examples. The choice will depend on the goals of the lattice base selection. We have optimised for the following objectives:

(B1) K_ℓ contains symmetries of interest to the practitioner;

(B2) Maximise the number of groups, $|K_\ell|$;

(B3) Maximise the pairwise Hausdorff distances $d_{K(\mathcal{G})}(G_i, G_j)$ for the compact G_i;

(B4) Minimise $|G_i|$ or $\dim G_i$ for all G_i in the base; and

(B5) Maximise $|G|$ or $\dim G$ for all $G \in K_\ell$.

These goals are roughly in order of importance. If there are no symmetries of interest then the method described can still proceed, but of course any prior interest should be the top priority. The second goal speaks to the efficiency of this method: as the ratio $|K|/2^\ell$ approaches 1 we

get maximal information about the whole lattice $K(\mathcal{G})$ out of the ℓ tests conducted. The third spaces the groups around K, which ensures that you don't waste tests on groups that the data cannot distinguish. The fourth and fifth mean that we should run the tests on smaller groups where we can cover with samples more easily, but that they give information about larger groups.

The following recursive method for selecting the lattice base provides a general solution that achieves a balance of the goals above. Suppose we are given a metric group $\mathcal{G}$, a desired number of groups ℓ, and some starting symmetries of interest $B_0 = \{G_1, \ldots, G_k\} \subseteq K(\mathcal{G})$ that form a lattice base. Then for $i = 1, 2, \ldots$, we pick an element $g \notin \bigvee_{G \in B_i} G$ such that:

1) $\overline{\langle g \rangle} \cap G \in \{I, G, \overline{\langle g \rangle}\}$ for all $G \in B_i$;

2) If $\mathcal{G}$ is finite then $|g| = \min_{h \notin \bigvee_{G \in B_i}} |h|$;

3) $\dim \overline{\langle g \rangle} \leq 1$; and

4) $\prod_{G \in B_i} d_{\mathcal{G}}(g, G)$ is (roughly) maximised.

Then set $B_{i+1} = B_i \cup \{\overline{\langle g \rangle}\}$. The product in (iv) encourages equal spacing between the other groups for a given sum of the distances. This of course requires the practitioner to know about various aspects of the group theory of $\mathcal{G}$, which are well known for all of the groups in Section 2.3.1 among many others.

5.3.5 Choice of α_i

There are a few simple choices for the significance levels α_i for a finite sample n. The first is to set $\alpha_i = \alpha/m$ for $i \leq m$, which is a Bonferroni correction for multiple testing. This is attractive in its simplicity but allows for only a finite portion of $K(\mathcal{G})$ to be searched.

Another choice is to pick a sequence with positive values for all i, for example $\alpha = \frac{1}{2^i}$. This would allow for infinitely many groups to be tested, though with likely very low powers for the later groups.

Lastly, we can pick a fixed level α and simply test *all* groups in the lattice base at this level, as is done in many practical variable selection problems [Heinze et al., 2018] (although not necessarily recommended given issues of post model selection inference). This of course can affect the probability $\mathbb{P}_f(G_{\max}(f, K_\ell) \in \hat{K}_\ell)$ but in practice often seems not to increase it beyond α because the tests run on the same dataset are not independent. See Section 5.9.2 as a continuation of Example 61 for an example of this.

5.3.6 Theoretical Properties of $\hat{G}_\ell$ and $\hat{K}_\ell$

We now turn to the understanding how the properties of the hypothesis test used determines the properties of the estimate $\hat{G}_\ell$ and the confidence region $\hat{K}_\ell$ for a fixed lattice base of size ℓ.

Theorem 57. *Suppose that the hypothesis test used is consistent at every significance level $\alpha \in (0,1)$. For any fixed significance levels α_i (that do not depend on n), we have:*

(1) $\mathbb{P}_f\big(\hat{G}_\ell \leq G_{\max}(f, K_\ell)\big) \to 1$ *when $\hat{G}_\ell$ is sampled uniformly from* $\max \hat{K}_\ell$.

Moreover, there exist deterministic sequences of significance levels $\alpha_i(n)$ for which:

(2) $(\mathrm{Test}_{\alpha_i(n)}(\mathcal{D}, G_i) : i = 1, \ldots, \ell) \xrightarrow{\mathbb{P}_f} 1 - 2(\mathbf{1}_{G_i \leq G_{\max}(f,K_\ell)} : i = 1, \ldots, \ell)$;

(3) $\hat{G}_\ell$ is a consistent estimator of $G_{\max}(f, K_\ell)$, when $\hat{G}_\ell$ is sampled uniformly from $\max \hat{K}_\ell$;

(4) $\hat{K}_\ell \xrightarrow{\mathbb{P}_f} \{G \in K_\ell : G \leq G_{\max}(f, K_\ell)\}$; *and*

(5) $\mathcal{E}(\hat{K}_\ell) \xrightarrow{\mathbb{P}_f} 0$,

where $\xrightarrow{\mathbb{P}_f}$ denotes convergence in probability under the measure $\mathbb{P}_f$. The sequence of significance values can be calculated if the power of Test_α *has a known lower bound.*

The first statement says that we are probabilistically guaranteed eventually not to *over*-estimate the symmetry. This means that the method presented is robust to type I errors of the tests used, as errors on the true symmetries of f result in missed symmetries but do not introduce errant ones. The statements (2)-(5) are equivalent formulations of the stronger statement of consistency.

5.4 Hypothesis Testing for a Particular Symmetry

We now turn to constructing tests for the hypothesis $H_0 : f$ *is G-invariant* against the alternative $H_1 : f$ *is not G-invariant* for use in the subgroup estimation, where G is any group acting measurably on $\mathcal{X}$. We present two options that make trade-offs in terms of generality and guarantees. The first requires stronger assumptions but has a theoretical guarantee of convergence, whereas the second makes weaker assumptions but has fewer guarantees and requires more computation.

5.4.1 Asymmetric Variation Test

This test relies on the idea that if we know how quickly f can vary locally (i.e., we have a relationship between $|f(x) - f(y)|$ and $d_{\mathcal{X}}(x, y)$) then we can construct a test statistic that measures the effect of permuting the covariates in X_i under random transformations from

the symmetry to test G.

If we can assume that the regression function f is in a class $\mathcal{F}$ of *bounded variation*, then we can define $V(x,y) = \sup_{f\in\mathcal{F}} |f(x) - f(y)|$. This bound V is used to identify if f varies more quickly than $V(g\cdot x, y)$. An example of such a class are β-Hölder continuous functions $\mathcal{F}(L,\beta)$ (defined in Section 2.1) for which $V(x,y) = Ld_{\mathcal{X}}(x,y)^{\min(\beta,1)}$. The only real choice here is L, as the continuous f can be uniformly well approximated by Lipschitz functions of some steepness, so in practice one would use the bound $V(x,y) = Ld_{\mathcal{X}}(x,y)$ using an assumed L. If this bound is unknown then it could be estimated using techniques in Section 5.9.1.

Suppose that we also have a bound p_t such that $\mathbb{P}(|\epsilon_i - \epsilon_j| > t) \le p_t$ for the independent mean zero additive noise $\epsilon_i = Y_i - f(X_i)$, for all positive t. One example of this is $p_t = \frac{2\sigma}{t}\frac{\exp(-t^2/4\sigma^2)}{\sqrt{2\pi}}$ for independent and identically distributed Gaussian ϵ_i with variance σ^2 (Proposition 2.1 of Adams [2020]), or in general one can use a Chebyshev inequality. Let μ_g be any distribution on the group G, and let $g \sim \mu_g$.

We can now define the test statistic based on a resampling procedure. Let m be the number of resamples. First, for $j \in \{1,\dots,m\}$ set $D^{g_j}_{I(j)J(j)} = |Y_{I(j)} - Y_{J(j)}| - V(g_j\cdot X_{I(j)}, X_{J(j)})$, where g_j is sampled independently from μ_g, $I(j)$ is sampled uniformly from $[n] = \{1,\dots,n\}$, and $J(j)$ is the index of the nearest neighbour to $g_j\cdot X_{I(j)}$ in the original data $\mathcal{D}_X = \{X_i\}_{i=1}^n$. Our test statistic is then $N_t^g = |\{D^{g_j}_{I(j)J(j)} \ge t\}|$, the distribution of which is controlled by the following lemma.

Lemma 58. *Under the null hypothesis of G-invariance, and for any fixed choice of resamples m, N_t^g is stochastically bounded from above by a binomial random variable Z with m trials and probability of success p_t for sufficiently large sample sizes n. I.e.,*

$$\mathbb{P}(N_t^g > c) \to p \le \mathbb{P}(Z > c) \tag{5.7}$$

for all $c \in \mathbb{R}$.

Using this lemma it is straightforward to calculate p-values for this hypothesis via:

$$p_{val} = \sum_{k=N_t^g}^{m} \binom{m}{k} p_t^k (1-p_t)^{m-k} \tag{5.8}$$

which are guaranteed to have asymptotic type I error control at any significance level $\alpha \in [0,1]$, i.e.,

$$\mathbb{P}_f(p_{val} \le \alpha) \to a \le \alpha \tag{5.9}$$

for all f that satisfy the null hypothesis of G-invariance.

Choices of t and of μ_g

The test presented here works for any choice of t and variable g, though particular choices of these will affect the power of the test. For example, if we choose t such that $p_t = 1$ then our p-value will always be 1 (as all the mass of Z will be on $Z = m$). Similarly if $g = e \in G$ almost surely then we will also only reject with probability at most α.

For the choice of t, we suggest calculating N_t^g from the sample of $D^{g_j}_{I(j)J(j)}$ at some grid of t values $t_0 < t_1 < \cdots < t_k$ with the values of p_{t_i} spread over the interval $(0, 1)$, and then taking the p-value of H_0 as the minimum p-value of each $N_{t_i}^g$. This is justified as the information of the test is entirely contained in the set $\{D^{g_j}_{I(j)J(j)}\}_{j=1}^m$, i.e., since the p-value is at most $\mathbb{P}(Z \geq N_t^g \mid N_t^g)$ for all t, we can take an infimum over t.

For the choice of μ_g, we suggest using a uniform distribution only on some set of topological generators, i.e., a set A with $\overline{\langle A \rangle} = G$. This means that we don't sample the identity or other elements that generate only subgroups of G to which f may be invariant, but if there is an element that breaks the invariance then one of the generators will too (by Lemma 65 in Section 5.8.3) so we should capture that chance.

Computational complexity

The main computational work is finding the nearest neighbour to each resampled point $g_j \cdot X_{I(j)}$. A naive algorithm would search across all n samples to find $J(j)$, which means that the algorithm would cost $O(mn)$ distance calculations. One improvement would be to store the feature vectors X_i in a k-d tree [Bentley, 1975]. Building such a tree takes $O(n \log n)$ operations, but can find nearest neighbours to a new point in $O(\log n)$ time. This speeds up the asymmetric variation to $O(m \log n)$.

Consistency of this test

Under the conditions above and some mild conditions on the noise distribution, and when the number of resamples m increases slower than the true sample size n, we can prove that the asymmetric variation test is consistent. The condition on the support of μ_X amounts to restricting $\mathcal{X}$ to the closure of the support as a practitioner would usually do. The condition that the noise admits a density is satisfied in many usual cases in regression (e.g. Gaussian noise). The condition that we can bound the concentration of the noise tightly is somewhat restrictive, but reflects the difficulty of the problem - if the asymmetry is obscured by more noise then it is much more difficult to identify it. The condition on μ_g ensures that we sample from enough of G to spot points where f is not G-invariant. A sufficient condition for this is the sample uniformly from some set of topological generators as suggested above, or from the uniform distribution on G when this exists.

Proposition 59 (Consistency of the Asymmetric Variation Test)**.** *Set m such that $m \to \infty$ and $m/n \to 0$ and fix $t > 0$. Suppose that the law of X has a dense support on $\mathcal{X}$. Suppose that ϵ admits a density f_ϵ with respect to Lebesgue measure on $\mathbb{R}$ that is decreasing in $|\epsilon|$. Suppose that μ_g is chosen such that $\mathbb{P}_f(f(g \cdot X) \neq f(X)) > 0$ for any f that is not G-invariant. Then the asymmetric variation test is consistent, i.e. $p_{val} \overset{\mathbb{P}_f}{\to} 0$ for all f that are not G-invariant.*

5.4.2 Permutation Variant of the Asymmetric Variation Test

The asymmetric variation test (in the previous subsection) relies on both the existence of, and the knowledge of, the bounds $V(x, y)$ and p_t. In this section we show that we can relax these assumptions and sometimes gain statistical power, but with increased computational cost and without the theoretical guarantee of consistency.

Suppose for now that μ_g is a distribution on G such that $d_{\mathcal{X}}(g_1 \cdot X_1, g_2 \cdot X_2) \overset{D}{=} d_{\mathcal{X}}(X_1, X_2)$ when $g_i \overset{iid}{\sim} \mu_g$ and $X_i \overset{iid}{\sim} \mu_X$. If μ_X is G-invariant, then any distribution on G suffices. Then under the null hypothesis $H_0 : f$ *is G-invariant* we have the equality:

$$\frac{|Y_i - Y_j|}{d_{\mathcal{X}}(X_i, X_j)} \overset{D}{=} \frac{|f(g_i \cdot X_i) + \epsilon_i - f(g_j \cdot X_j) - \epsilon_j|}{d_{\mathcal{X}}(g_i \cdot X_i, g_j \cdot X_j)} = \frac{|Y_i - Y_j|}{d_{\mathcal{X}}(g_i \cdot X_i, g_j \cdot X_j)} \tag{5.10}$$

for any variation bound V for f. Thus instead of comparing outputs at transformed *points* $g_{I(i)} \cdot X_{I(i)}$ to non transformed points $X_{J(i)}$ as with the asymmetric variation test we now compare the slopes above for transformed *datasets*:

$$\mathcal{D}^{(k)} = \{(g_i^{(k)} \cdot X_i, Y_i)\}_{i=1}^n \tag{5.11}$$

where $g_i^{(k)} \overset{iid}{\sim} \mu_g$ for all i and $k \in \{1, \ldots, B\}$. To ensure independence we split in half by uniform sampling; half gets used for a baseline computation and half gets used for the transformed datasets. For any non G-invariant function f we break the second equality of 5.10, and when there is a bound V as with the previous test then the right hand side is frequently larger than the left. This leads to a resampling test with minimal assumptions defined in algorithm 3.

Under the null hypothesis and the assumption that μ_g preserves the distances, we know equation 5.10 ensures interchangeability of the variables $A(k)$ and A_0, so the type 1 error of the hypothesis test $\text{Test}_\alpha(\mathcal{D}, G) = 1 - 2\mathbf{1}_{p_{val} \leq \alpha}$ is controlled exactly at level α in finite samples. If we drop the assumption that μ_g preserves pairwise distances then we can in fact show that this test still has a valid asymptotic size under a modified condition that holds much more generally.

Algorithm 3 Permutation Variant of Asymmetric Variation Test

procedure PERMVARTEST($\mathcal{D}$, μ_g, B)
  for $k \in \{1, \dots, B\}$ **do**
    Sample $\mathcal{I} \subseteq [n]$ of size $\lfloor n/2 \rfloor$
    Compute $S^0 = \{|Y_i - Y_j|/d_{\mathcal{X}}(X_i, X_j) : i \neq j \in [n] \setminus \mathcal{I}\}$
    $A_0(k) \leftarrow \max S_0$
    Sample $g_i \overset{iid}{\sim} \mu_g$ for $i \in [n]$
    Compute $S^k = \{|Y_i - Y_j|/d_{\mathcal{X}}(g_i^{(k)} \cdot X_i, g_j^{(k)} \cdot X_j) : i \neq j \in \mathcal{I}\}$
    $A_G(k) \leftarrow \max S^k$
  $p_{val} \leftarrow |\{k : A_G(k) \leq A_0(k)\}|/B$
  return p_{val}

Proposition 60. *Suppose that μ_g is such that there exists a positive constant $T > 0$ such that:*

$$\mathbb{P}\big(d(g_i \cdot X_i, g_j \cdot X_j) \leq t\big) \leq \mathbb{P}\big(d(X_i, X_j) \leq t\big) \tag{5.12}$$

for all $t \leq T$, when $i \neq j$ and $g_i, g_j \overset{iid}{\sim} \mu_g$ and $X_i, X_j \overset{iid}{\sim} \mu$. Suppose that $f \in \mathcal{F}(L, 1)$ for some $L \in \mathbb{R}_{>0}$. Then the permutation variant is asymptotically valid at any size $\alpha \in (0, 1/2)$.

Proof of Proposition 60. We rely on some preliminary facts:

$$\mathbb{P}\Big(\max S^k = \max_{i,j} \tfrac{|f(g_i^{(k)} \cdot X_i) - f(g_j^{(k)} \cdot X_j)| + |\epsilon_i - \epsilon_j|}{d(g_i^{(k)} \cdot X_i, g_j^{(k)} \cdot X_j)}\Big) \to 1 \tag{5.13}$$

$$\mathbb{P}\Big(\max S^0 = \max_{i,j} \tfrac{|f(X_i) - f(X_j)| + |\epsilon_i - \epsilon_j|}{d(X_i, X_j)}\Big) \to 1 \tag{5.14}$$

$$\mathbb{P}\big(\text{argmax}_{i,j \in \mathcal{I}} S^k \text{ has } d(g_i^{(k)} \cdot X_i, g_j^{(k)} \cdot X_j) \leq T\big) \to 1 \tag{5.15}$$

$$\mathbb{P}\big(\text{argmax}_{i,j \in [n] \setminus \mathcal{I}} S^0 \text{ has } d(X_i, X_j) \leq T\big) \to 1 \tag{5.16}$$

The first two statements are due to the fact that $|Y_i - Y_j| = |f(X_i) - f(X_j)| + |\epsilon_i - \epsilon_j|$ with probability 1/2 (and the null hypothesis allows for any transformation of the X_i inside of f) and otherwise $|Y_i - Y_j|$ is the smaller quantity $\big||f(X_i) - f(X_j)| - |\epsilon_i - \epsilon_j|\big|$ with probability 1/2. Thus as $n \to \infty$ the maximum will occur for the larger possibility with probability converging to 1. The second two statements are due to the fact that

$$\frac{|f(g_i \cdot X_i) - f(g_j \cdot X_j)| + |\epsilon_i - \epsilon_j|}{d(g_i \cdot X_i, g_j \cdot X_j)} \leq L + \frac{|\epsilon_i - \epsilon_j|}{d(g_i \cdot X_i, g_j \cdot X_j)} \tag{5.17}$$

and so the maximum of S^k will eventually occur for small $d(g_i \cdot X_i, g_j \cdot X_j)$ (rather than for large $|\epsilon_i - \epsilon_j|$, which is has the same distribution independently of the distance term) which takes the second term to ∞, and of course the same is true for S^0.

Now note that each of these four statements must also hold conditional on $\mathcal{D}$ with conver-

gence replaced by convergence in probability, because the statements above guarantee the L^1 convergence of these $\mathcal{D}$-random variables.

Lastly note that the condition on the distribution of each g_i ensures that

$$\mathbb{P}\Big(\tfrac{|\epsilon_i-\epsilon_j|}{d(X_i,X_j)} \le \theta \mid d(X_i,X_j) \le T\Big) \le \mathbb{P}\Big(\tfrac{|\epsilon_i-\epsilon_j|}{d(g_i\cdot X_i,g_j\cdot X_j)} \le \theta \mid d(g_i\cdot X_i, g_j\cdot X_j) \le T\Big) \tag{5.18}$$

for all $\theta \ge 1/T$ (noting that the noise terms are independent of the X_i and g_i terms).

Let $\mathcal{A}$ denote the intersection of the events considered in equations 5.13. We know that for sufficiently large n:

$$\mathbb{P}\Big(\max_{i,j\in[n]\setminus\mathcal{I}} \tfrac{|\epsilon_i-\epsilon_j|}{d(X_i,X_j)} \le \max_{i,j\in\mathcal{I}} \tfrac{|\epsilon_i-\epsilon_j|}{d(g_i^{(k)}\cdot X_i, g_j^{(k)}\cdot X_j)} \,\Big|\, \mathcal{A},\mathcal{D}\Big) \le 1/2 \tag{5.19}$$

as the maximum will occur for large arguments with high probability which then use the above condition, and using the independence of the samples from the data splitting.

Thus conditioned on $\mathcal{D}$ the permutation statistics are independent trials with probability of success $q > 1/2$ and so Algorithm 3 gives valid (though perhaps conservative) p-values under the null hypothesis (as $F_{pval\,B}(\alpha B) \le F_K(\alpha B) \le \alpha$ for $K \sim \text{Binom}(B, 1/2)$, see Lemma 73). □

Note that the proof here relies on the Lipschitz bound of f through equation (5.17), but the algorithm does not require knowledge of the particular Lipschitz constant L. This means that we can avoid the problems of potentially having to estimate it from the previous test.

Choice of sampling distribution on the group

The practitioner has a full choice of the sampling distribution μ_g from the group G; any choice will give a valid test as long as it satisfies the condition of Proposition 60. This condition can be examined empirically with a Q-Q plot of the empirical quantiles of $\{d(X_i,X_j) : i \ne q \in [n]\}$ and $\{d(g_i\cdot X_i, g_j\cdot X_j) : i \ne j \in [n], g_i \overset{iid}{\sim} \mu_g\}$. If the quantiles of the original data have more mass for lower values then we assume that the sampling distribution is appropriate. If we know the sampling distribution of the covariates X_i exactly one can instead verify the condition theoretically.

This condition holds very broadly, but can be broken for some actions of non-compact groups. For example, consider the action of $R_{>0}$ by scaling on $\mathbb{R}^d$ (so $g\cdot X = gX$): if μ_g only places mass on the interval $(0,1)$ then it will contract the distances between transformed data points. This can of course be avoided by picking a distribution with mass on

the values above one, which is why this testing approach can work even for non-proper actions.

Regarding the optimal choice: we want to sample from as much of the group as possible but still want to maximise the mass of $d(g_i \cdot X_i, g_j \cdot X_j)$ near 0. As with the previous test we can achieve both goals by constructing a set of topological generators of G (where possible) and sampling uniformly from this set. Other good choices are simply to pick the uniform distribution on G (if it is compact) or on a symmetric[1] neighbourhood of the identity (if it is locally compact).

Computational complexity

This permutation method requires $O(nB)$ samples and transformations of the inputs X_i and then the computation of $O(n^2B)$ distance matrices. The remaining computations (the computing the maxima and the comparisons) are negligible in comparison. Thus the overall complexity is $O(n^2B)$ which is at least B times more intensive than the Asymmetric Variation Test (with the choices suggested in its section).

5.5 Numerical Experiments

We now demonstrate the performance of the point estimate $\hat{G}_\ell$ and of the confidence region $\hat{K}_\ell$ on synthetic datasets. The first simulation is a low dimensional regression estimation where we can directly quantify the error $\mathcal{E}(\hat{K}_\ell)$ directly. The second simulation then considers the effect of using the estimated group $\hat{G}_\ell$ to symmetrise an estimator $\hat{f}$ of f and consider the mean squared prediction error.

In Table 5.1 we summarise the simulations with the proportions of correct point estimates of the maximal symmetry (i.e., where $\hat{G}_\ell = G_{\max}(f, K_\ell)$) when $\hat{G}_\ell$ is sampled uniformly from $\max \hat{K}_\ell$. Note that almost all of the failures are overestimation of the confidence region, so these do not indicate type 1 failures. Also note that the permutation variant appears to reach a plateau in power, particularly in the dimension 2 example of Example 61. This is due to the density of samples with sample pairwise distances, which makes the test more susceptible to noise. Where simulations were not run these are marked with a dash.

All tests were conducted locally on a 2020 13" MacBook Pro laptop with a 2 GHz Quad-Core Intel Core i5 processor and 16GB of RAM. All code is available on GitHub at: https://github.com/lchristie/estimating_symmetries.

[1] symmetric here meaning that $A \subseteq G$ has $\{a^{-1} : a \in A\} = A$.

Table 5.1 Proportions of correct symmetry estimation.

Example	Scenario	Test	n = **20**	**50**	**100**	**200**	**500**
61	Dimension 2	AVT	0.18	0.67	0.98	1	1
		PV	0.010	0.295	0.545	0.650	0.590
	Dimension 4	AVT	0	0.015	0.010	0.180	0.910
		PV	0	0.015	0.150	0.755	0.930
	Dimension 6	AVT	0	0	0	0	0
		PV	0	0	0.005	0.110	0.715
62	Spherically Symmetric	PV	1	1	1	1	-
	Rotationally Symmetric	PV	0	0.6	1	1	-
	Totally Asymmetric	PV	0	0.02	0.95	1	-
	Misspecified Lattice	PV	0	0	0.05	0.87	-

Example 61 (Finite Group in Low Dimensions)**.** *Consider the regression functions:*

$$f_d : (x_1, \ldots, x_d) \mapsto \exp(-|x_1|) \tag{5.20}$$

for each dimension $d = 2, 3, 4, \ldots$ *and suppose that* $X_i \overset{iid}{\sim} N(0, 4I_d)$ *and* $\epsilon_i \overset{iid}{\sim} N(0, 0.05^2)$*. Consider the group* D_4 *acting by rotations and reflections of the* x_1*-*x_2 *plane. We consider the lattice base:*

$$\left\{G_i\right\}_{i=1}^{\ell} = \left\{\langle R_h\rangle, \langle R_v\rangle, \langle R_\pi\rangle, \langle R_/\rangle, \langle R_\backslash\rangle, \langle R_{\pi/2}\rangle\right\} \tag{5.21}$$

which has $K_\ell = L(D_4)$*. Note that* R_h *reflects through the* x_2 *line in this* x_1*-*x_2 *plane, and* R_v *through the* x_1 *line etc. Then the symmetries* R_h*,* R_v*, and* R_π *all apply to* f_d *as they only change signs of* x_1 *and* x_2*. However* $R_/$*,* $R_\backslash$*, and* $R_{\pi/2}$ *do not because they exchange* x_1 *and* x_2 *(as well as change signs). Therefore the correct object we wish to estimate is:*

$$G_{\max}(f_d, K_\ell) = \langle R_h, R_\pi\rangle \tag{5.22}$$

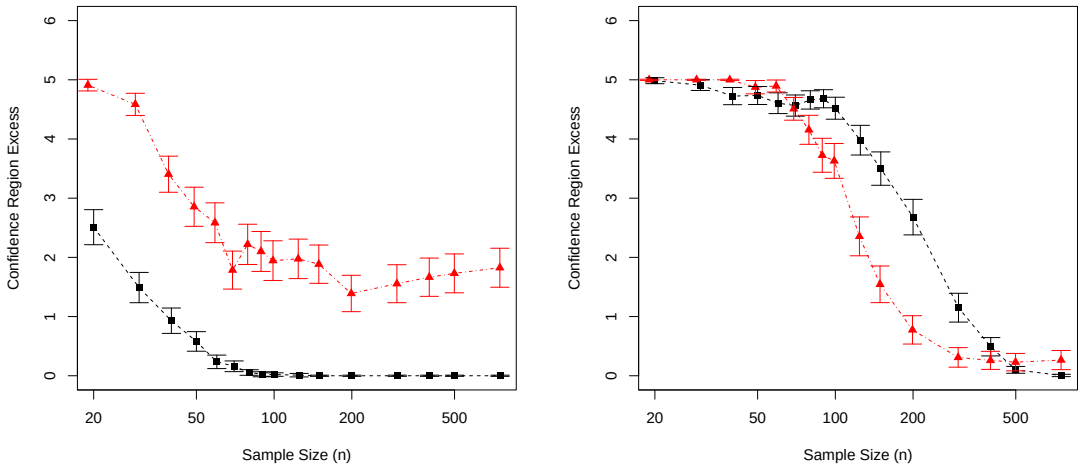
for every dimension d*, and the correct test result vector would be* $(1, 1, 1, 0, 0, 0)$*.*

We tested each null hypothesis $H_0^{(i)}$ *:* f *is* G_i*-invariant at the* $\alpha_i = 0.05$ *significance level. This means that* $\hat{K}_\ell$ *will be a* 0.7 *confidence region for* $G_{\max}(f, K_\ell)$ *per Theorem 57, but we see that it contains this object much more frequently in section 5.9.2. Each sampling distribution from the groups are uniform on the group. The asymmetric variation tests had* $t = 2$*,* $\sigma = 0.1$*,* $p_t = \frac{2\sigma}{t}\frac{\exp(-t^2/4\sigma^2)}{\sqrt{2\pi}}$*, and the number of resamples equal to the number of samples. The permutation variants used* $B = 100$*. Both tests assume that* $f \in \mathcal{F}(1, 1)$*, the class of* 1*-Lipschitz functions.*

The results of these simulations for dimensions 2 and 4 are plotted in Figure 5.4 for the performance of the confidence region $\hat{K}_\ell$*. Note that if* $\mathcal{E}(\hat{K}_\ell) = 0$ *then* $\hat{G}_\ell = G_{\max}(f, K_\ell)$*.*

These plots show that the Asymmetric Variation Test can detect this asymmetry reliably, though this performance decays with higher dimension as the X_i become spaced with further distances. The permutation variant interestingly struggles more in low dimensions, exhibiting moderate errors even when $n = 300$, though outperforms the AVT in dimension 4 and above. Results for the higher dimensional simulations are available in section 5.9.2.

In this example computations of the asymmetric variation test when $n = 100$ and $d = 4$ each took 0.004s and computations of the permutation variant each took 0.118s.



(a) $\mathcal{E}(\hat{K}_\ell)$ when $d = 2$ (b) $\mathcal{E}(\hat{K}_\ell)$ when $d = 4$

Fig. 5.4 This figure describes the results of the simulations in Example 61. Subfigures (a) and (b) track the confidence region excess $\mathcal{E}(\hat{K}_\ell)$ for each test used (black for the asymmetric variation test and red for the permutation variant), in dimension 2 for (a) and dimension 4 for (b). The error bars are Wald 95% Confidence Intervals for the true expected excess.

Example 62 (Simulation of Symmetrised Estimators)**.** *We now consider an example of estimating continuous symmetries. Suppose that the domain $\mathcal{X}$ is $\mathbb{R}^3$ and let $\mathcal{G} = SO(3, \mathbb{R})$ act on $\mathcal{X}$ by matrix multiplication. Let K_ℓ be the sub-lattice of $K(SO(3, \mathbb{R}))$ with lattice base:*

$$\{G_i\}_{i=1}^{\ell} = \{S^1_{e_1}, S^1_{e_2}, S^1_{e_3}, S^1_{u_1}, \ldots, S^1_{u_6}, SO(3)\}. \tag{5.23}$$

Here the axes $u_1, \ldots u_6$ (given explicitly in appendix 5.9.2) correspond to non-parallel unit vectors through the vertices of a regular icosahedron, rotated such that the distances to the standard basis vectors are roughly maximised (as with section 5.3.4). We consider three possible regression functions with differing levels of symmetry:

Regression Function	***Maximal Symmetry***	
$f_1(x) = \sin(\|x\|_2)$	$G_{\max}(f_1, K_\ell) = SO(3)$	*Spherically Symmetric*
$f_2(x) = \sin(\|x_1\|) + \cos(\sqrt{x_2^2 + x_3^2})$	$G_{\max}(f_2, K_\ell) = S^1_{e_1}$	*Rotationally Symmetric*
$f_3(x) = \sin(\|x_1\|) + 0.2x_2 - \cos(x_3)$	$G_{\max}(f_3, K_\ell) = I$	*Totally Asymmetric*
$f_4(x) = f_2(R_{0.1,e_3}x)$	$G_{\max}(f_3, K_\ell) = I$	*Symmetry not in* K_ℓ

where $R_{0.1,e_3}$ *is the rotation matrix of 0.1 radians around the* z *axis. This means that* f_4 *has symmetries that are isomorphic to those of* f_2*, but are not included in the lattice* K_ℓ*. This demonstrates the usefulness of these methods even in the case when the lattice* K_ℓ *is misspecified.*

We simulate regression data $\mathcal{D} = \{(X_i, f(X_i) + \epsilon_i\}_{i=1}^n$ *and* $\mathcal{D}' = \{(X_i', f(X_i') + \epsilon_i'\}_{i=1}^n$ *where* $X_i \overset{iid}{\sim} N(0, 2I_3)$*,* $X_i' \overset{iid}{\sim} N(0, 2I_3)$*, and* $\epsilon_i, \epsilon_i' \overset{iid}{\sim} N(0, 0.01^2)$*. We again assume that* $f \in \mathcal{F}(1,1)$*, the class of* 1*-Lipschitz functions.*

We estimate $\hat{G}_\ell$ *using the permutation variant of the asymmetric variation test with* $B = 100$ *and* μ_g *sampled uniformly from each group. We then consider the problem of estimation of* f*, using an estimate that has been symmetrised by* $\hat{G}_\ell$*. In particular, we consider three regressors:*

(A) A standard local constant estimator (LCE) $\hat{f}(x, \mathcal{D})$ *with bandwidths selected by cross-validation (the default in the R-package* NP *[Hayfield and Racine, 2008]);*

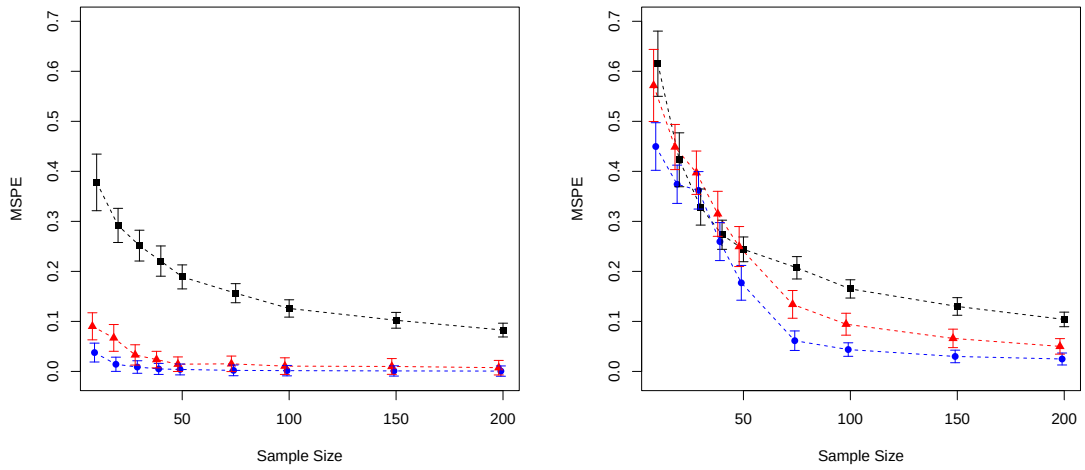
(B) A symmetrised LCE $\hat{f}^A_{\hat{G}}\big(\pi_{\hat{G}}(x), \{(\pi_{\hat{G}}(X_i), Y_i)\}_{i=1}^n\big)$ *(utilising data projection); and*

(C) A symmetrised LCE that splits $\mathcal{D}$ *into independent pieces half the size of* $\mathcal{D}$*, then using the first half to estimate* $\hat{G}_\ell$ *only and the second half to estimate* $\hat{f}$ *afterwards as with (B).*

Here $\pi_G : \mathcal{X} \to \mathcal{X}/G$ *is the natural projection given by* $\pi_I(x) = x$*,* $\pi_{SO(3)}(x) = \|x\|_2$*, and* $\pi_{SL(3)}(x) = \mathbf{1}_{x \neq 0}$*. The natural projection for the circle groups are given by* $\pi_{S^1_u}(x) = (x \cdot u, \sqrt{\|x\|^2 - (x \cdot u)^2})$*. We have chosen local constant, rather than local linear for example, estimators because of computational time and because we do not have boundary effects with the domain.*

The mean squared prediction errors are plotted in Figures 5.5 and 5.6. We see that for scenarios 1 and 2 the MSPE decays rapidly and is able to correctly generalise once $\hat{G}$ *is more likely than not to estimate the correct symmetry. When there is no symmetry present, as with scenario 3, the we quickly see convergence of* $\hat{f}^A_{\hat{G}} - \hat{f} \overset{\mathbb{P}_f}{\to} 0$*. Interestingly, we see no empirical benefit from splitting the data into independent pieces, suggesting that the errors of the group estimation and the functional estimation are uncorrelated. In the fourth scenario at*

low sample sizes, using the errant symmetry to symmetrise the estimator reduces the variance more than it increases bias resulting in smaller MSPE. At higher sample sizes it correctly rejects the $S^1_{e_1}$ *symmetry and the estimator converges to the baseline.*



(a) Scenario 1 (b) Scenario 2

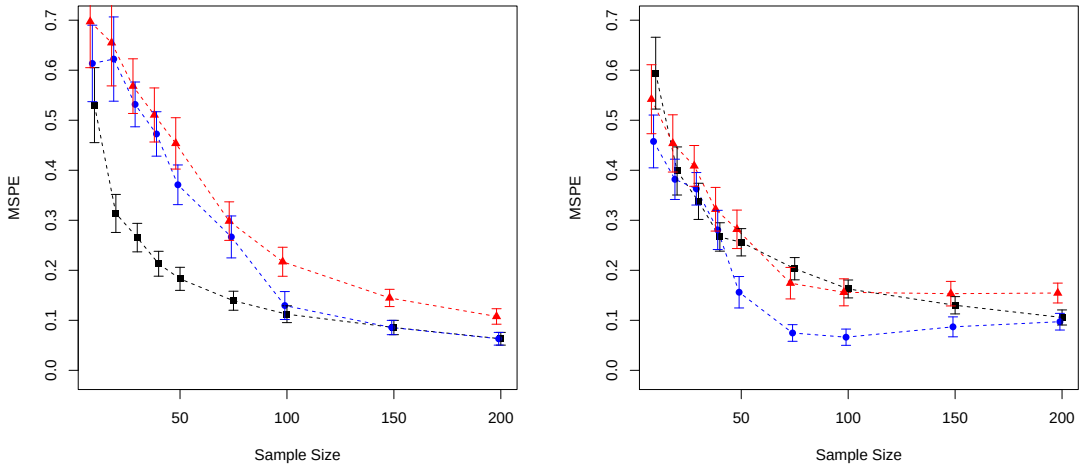
Fig. 5.5 Mean squared prediction errors for the three estimators, under scenarios 1 and 2. The averages over 100 simulations are shown in black for $\hat{f}$, in blue for $\hat{f}^A_{\hat{G}}$, and in red for $\hat{f}^S_{\hat{G}}$. The error bars are Wald 95% Confidence Intervals for the true mean squared predictive error.

5.6 Applications to Real Data

We now demonstrate these methods on two real world datasets. The first is from geophysics, where we seek to estimate the Earth's magnetic field. In this example, we show that our confidence region built from satellite readings contains the identity only. This supports the practice of modelling an asymmetric magnetic field, as in Thébault et al. [2015]. The second example explores the spatio-temoral relationship of sunspots. We show that in this example our estimated maximal symmetry of this relationship is rotational, and consistent with the phenomenon known as Spörer's law [Babcock, 1961] for most of the 12 available solar cycles. The other four cycles have $\hat{G}_\ell = I$, which indicates that there are longitudinal effects along with the latitudinal effects in Spörer's law.

5.6.1 Satellite Based Magnetic Field Data

The European Space Agency (ESA) launched the SWARM mission in 2013 to measure the Earth's magnetic field and other related properties. This consists of three satellites orbiting



(a) Scenario 3 (b) Scenario 4

Fig. 5.6 Mean squared prediction errors for the three estimators, under scenarios 3 and 4. The averages over 100 simulations are shown in black for $\hat{f}$, in blue for $\hat{f}^A_{\hat{G}}$, and in red for $\hat{f}^S_{\hat{G}}$. The error bars are Wald 95% Confidence Intervals for the true mean squared predicted error.

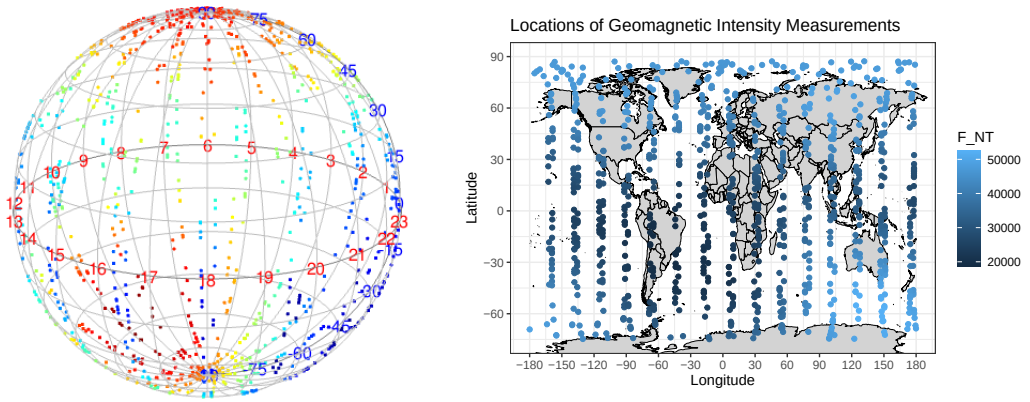
between 450km and 530km above the Earth's surface. These measurements complement the observations at ground based observatories to better model the field for use in navigation and geospatial modelling. This vector field $B : \mathbb{R}^3 \to \mathbb{R}^3$ is generated primarily by the rotation of the inner core, but is also affected by: magnetised material in the Earth's crust; electrical currents in the upper atmosphere caused by solar winds; and the interaction of the sun's magnetic field. Without these additional effects it is reasonable to model the field as one would a bar magnet, with a rotational symmetry through the magnetic poles.

Given these effects can break this symmetry, the vector field B on and above the Earth's surface at time t is typically modelled as the gradient of a potential field $B = \nabla V$ where:

$$V(r,\theta,\phi,t) = a \sum_{n=1}^{13} \sum_{m=0}^{n} \left(\frac{a}{r}\right)^{n+1} \Big(g_n^m(t)\cos(m\phi) + h_n^m(t)\sin(m\phi)\Big) P_n^m(\cos(\theta)) \tag{5.24}$$

in the International Geomagnetic Reference Field (IGRF) [Thébault et al., 2015]. Here $a = 6,371.2$km is the Earth's radius, θ is the geocentric co-latitude, ϕ is the east longitude. The functions P_n^m as the Schmidt quasi-normalised associated Legendre functions of degree n and order m. The Gauss coefficients g_n^m and h_n^m are then fitted using observational data from ground based stations.

Using the data of the ESA SWARM satellites, accessed via the VirES client (https://vires.services/), we can estimate the symmetries of the field. A full description of the data source is available in appendix 5.9.2. We use 1000 measurements of the field intensity $\|B\|_2$ as the response variable chosen uniformly at random from the observations on the 25th of February 2023. These measurements are plotted in Figure 5.7.



(a) Geo-spatial plot (b) Mercator Projection

Fig. 5.7 A visualisation of the measurements of the Earth's magnetic field, using the 1000 samples on February 25, 2023. In subfigure (a), red dots represent the high intensity measurements that occur towards the poles, and blue dots the regions of lowest intensity towards the equator. Red numbers indicate longitude and blue numbers represent latitude. In subfigure (b) intensity high magnetic field intensity is displayed in light blue and low intensity in dark blue.

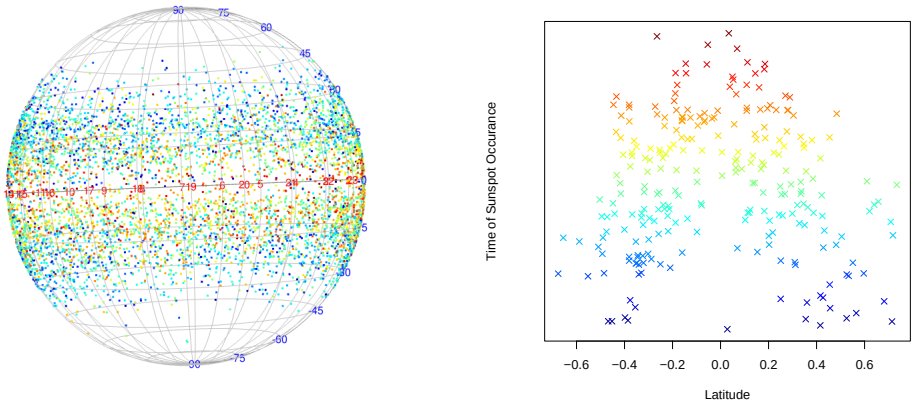
We have used the same symmetry estimation procedure of Example 62, with six additional symmetries in the lattice base (corresponding to vertices of a further offset icosahedron). These axes of these S_u^1 rotational symmetries are given explicitly in Section 5.9.2. We have used the permutation variant because we do not know the noise distribution explicitly. Unlike Example 62, the data is not distributed symmetrically around the sphere, but examinations of the Q-Q plots in Section 5.9.2 show that the condition of Proposition 60 holds empirically for the uniform distribution on all tested subgroups.

All symmetries were rejected at the $\alpha = 0.05$ significance level, so the estimated maximal symmetry is just trivial group I. This supports the modelling approach of the IGRF and rejects the simpler bar magnet model of the Earth's magnetic field.

5.6.2 Sunspots and Spörer's Law

The Sun's photosphere has small regions of low intensity, known as *sunspots*. These sunspots occur due to complex interactions between aspects of the inner and outer solar magnetic fields, as described in Babcock [1961]. As noted in García-Portugués et al. [2020], the distribution of the locations of these spots over each orbit is approximately symmetric around the rotational axis of the sun. However, this distribution depends on time within each solar cycle (the period between reversals of the polarity of the magnetic field) - the latitudes decrease over the course of the period. This phenomenon, known as *Spörer's Law*, is explained by an intensification of the twisting that causes the sunspots that culminates in a polarity reversal.

Records of the sunspots over the last 13 solar cycles (cycles 12 - 24) are readily available in the R package rotasym [García-Portugués et al., 2020]. This was originally collected in Debrecen Photoheliographic Data (DPD) sunspot catalogue [Baranyi et al., 2016]. We have taken a uniform sample of 250 sun spots from the 21st solar cycle and taken a local average of time of occurrence of sunspots from the entire catalogue of that cycle. This data is represented as the vector (x_i, y_i, z_i, t_i) with $x_i^2 + y_i^2 + z_i^2 = 1$. The locations of the spots in this orbit are shown in Figure 5.8. The plot of time vs. latitude shows the Spörer's law seems valid—the sunspots closest to the equator do appear later in the cycle than the ones away from it.



(a) Helio-spatial Plot (b) Plot of time vs. latitude

Fig. 5.8 Locations of all sunspots during the 21st recorded solar cycle, from the rotasym R package. The colours represent the time with the cycle, with blue towards the start and red towards the end. The blue numbers in subfigure (a) indicate latitude, the red longitude in degrees. Also shown is a marginal of time against latitude in subfigure (b) for 250 points uniformly sampled from this cycle.

We can examine the symmetries of this relationship between the locations $X = (x, y, z)$ of the observed spots and the mean time of occurrence conditioned on locations $f(x, y, z) = \mathbb{E}(t \mid x, y, z)$. Whilst this is not the usual regression situation, where we would predict location given the time, it is mathematically equivalent. Specifically, if the spatial distribution conditioned on time is invariant to G, then the conditional mean $\mathbb{E}(t \mid x, y, z)$ must be too[2], and so we can use the symmetry testing and estimation techniques to give the largest possible symmetry of this conditional distribution.

We have conducted tests on the same lattice base as in the previous application, so the lattice base is:

$$\{G_i\}_{i=1}^{15} = \{S_{u_1}, S_{u_2}, \ldots, S_x, S_y, S_z, SO(3)\} \tag{5.25}$$

The vectors u_i are given in Section 5.9.2. We used the permutation variant with $B = 1000$ because we do not know the noise distribution. Sampling from each group was done by sampling an angle uniformly on $[0, 2\pi)$ and constructing the rotation around the axis by this angle. As suggested in Section 5.4.2, we checked Q-Q plots between the original and transformed pairwise distance distributions, which showed that these sampling distributions are suitable in light of Proposition 60. These are available for cycle 21 in Section 5.9.2.

For this cycle and all others with data available in **rotasym**, the confidence regions for the maximal symmetry of $\mathbb{E}(t \mid X)$ are shown in table 5.2. Four of the cycles are consistent with rotational symmetry of the conditional mean around the z-axis. This is in agreement with the symmetries of the distributions of García-Portugués et al. [2020], where all cycles except for the 22 have a symmetric marginal distribution of locations around the z-axis. Interestingly the rejection of symmetry of the conditional distribution suggests that perhaps there is a unique conditional mean location that follows a spiral pattern in towards the equator.

5.7 Discussion

We have given methods for finding and using symmetry in regression function estimation. We have shown that we can generalise the idea of variable subset selection into selection of non-linear combinations of the variables by placing them within the framework of subgroup lattices. We have developed methods for testing and used these to estimate the maximal invariant symmetry of a regression function. We have demonstrated the power and applicability of these tests across several numerical experiments, showing that when the symmetry exists our methods do detect this consistently and that this can significantly improve estimates of

[2]This is seen by considering first the marginal of the spatial distribution (which must be G-invariant), and then applying Bayes Law and the invariant of the conditional and the marginal as required.

Table 5.2 Confidence regions for each solar cycle

Solar Cycle	Confidence Region
12	$\{I\}$
13	$\{I, G_{11}, G_{15}\}$
14	$\{I\}$
15	$\{I\}$
16	$\{I, G_{15}\}$
17	$\{I\}$
18	$\{I\}$
19	$\{I\}$
20	$\{I, G_{15}\}$
21	$\{I\}$
22	$\{I\}$
23	$\{I\}$
24	$\{I, G_{15}\}$

the regression function. The estimation of the maximal symmetry using this method has two main advantages over the M-estimation in the previous chapter. It has built in inferential properties and it is often much faster computationally because it avoids estimating f directly.

A limitation to this method is that by restricting to a finite sub lattice of $K(\mathcal{G})$ we might miss some true symmetries of f that are not within this lattice. If the sub lattice chosen is also an δ-cover of $K(\mathcal{G})$ (in the metric of the previous chapter), then we would hope that the symmetries close to the true maximal symmetry would not be rejected because the power of the tests would be lower. It would be interesting to bound the power of this test in terms of the symmetric gap $\|f - S_G f\|_2^2$ (under the conditions that G is compact and μ is G-invariant) and then investigate the behaviour of the lattice estimator when the base is chosen as a δ-cover using the δ from Theorem 41. Increasing the size of the lattice base in this case worsens the multiple testing problem but could still lead to consistent estimation if $\mathbb{P}(p_{val} > \alpha) \to 1$ under the null hypothesis for appropriate choices of α.

The main limitation of this method is that $\hat{G}_\ell$ is biased towards more symmetry because it relies on a hypothesis test that does not always have power. This can help alleviate the sub-lattice problem because we can still end up with an estimate ideally close to the true symmetry rather than simply rejecting everything because the only true symmetry in the lattice is the identity. This is observed in the misspecified lattice in Example 62.

This method could be improved by a more thorough treatment of the multiple testing effects. We have taken perhaps the most conservative approach of just using the Bonferonni correction, but there are many other methods to control the family wise error rate such as

Holm's step-down procedure [Holm, 1979]. We could also look to adapt the DAG testing approach used in Müller et al. [2023] as mentioned in Section 2.8, noting that the lattice we use also forms a directed acyclic graph (DAG).

It is also worth considering methods akin to forward selection and backwards elimination in the variable selection literature. One could, for example, look to minimise residual sum of squares of the smoothed models $\hat{f}_G$ over the lattice K at each level as with forward selection. This would extend the M-estimation strategy of the previous chapter to make use of the lattice structure of the subgroups.

One issue with the hypothesis tests used is that the power of the tests depends highly on sampling data with small distances $d(X_i, X_j)$ for $X_i, X_j \overset{iid}{\sim} \mu$. As the dimension of the covariate space increases this distribution can shift right—if $X_i \overset{iid}{\sim} U([0,1]^a)$ pairwise distances will be closer with greater probability than if $X_i \overset{iid}{\sim} U([0,1]^b)$ whenever $a < b$. This is more a problem with an "effective dimension" of the distribution μ rather than the space we observe the data in: if μ is concentrated on a low dimensional submanifold then the pairwise distances can remain small and the tests retain power.

5.8 Proofs and Further Mathematical Details

5.8.1 Proofs in Section 5.2

Proof of Proposition 49. We need only show that $K(\mathcal{G})$ contains supremums and intersections (with respect to the subgroup ordering) of arbitrary pairs $G, H \in K(\mathcal{G})$, and that these are given by the join and meet operations in the statement. If $G, H \in K(\mathcal{G})$, then consider that $G \cap H$ is still the infimum as with the subgroup lattice. For the supremum, consider that $G, H \leq \overline{\langle G, H \rangle}$ by definition, so this is an upper bound. If $G, H \leq A \in K(\mathcal{G})$ then $\langle G, H \rangle \leq A$ and so $\overline{\langle G, H \rangle} \leq \overline{A} = A$, and so $\overline{\langle G, H \rangle}$ is in fact the least upper bound and we are done. □

Proof of Proposition 52. First, note that $\langle G_1, \dots G_\ell \rangle_{K(\mathcal{G})}$ must contain all groups of the specified form by the definition of a lattice. Thus we need to show that the set of such groups forms a lattice. Let $G = G_{i_1} \vee \cdots \vee G_{i_m}$ and $H = G_{j_1} \vee \cdots \vee G_{j_b}$ be two groups in the set considered. The supremum $G \vee H$ is also of the form specified because of the commutativity, associativity, and idempotency of the operation $\vee$ [Schmidt, 2011].

For the infimum, first define $\mathcal{I} = \{i \in [\ell] : G_i \leq G \cap H\}$, and let $A = \bigvee_{i \in \mathcal{I}} G_i$. See that $A \leq G \cap H$ by definition. Take any element $g \in G \cap H$, and take sequences $\{g_n^G\}_{n \in \mathbb{N}} \subseteq \langle G_{i_a} : a \in [m] \rangle$ and $\{g_n^H\}_{n \in \mathbb{N}} \subseteq \langle G_{j_a} : a \in [a] \rangle$ both convergent to g. If there is a subsequence of either in $\langle G_i : i \in \mathcal{I} \rangle$ then $g \in A$. If not, then by the pigeon hole principle there must exist a subsequence of $\{g_n^G\}$ contained in some G_{i_a} for $i_a \notin \mathcal{I}$, and similarly for $\{g_a^H\}$ in some $G_{j_{a'}}$ for $j_{a'} \notin \mathcal{I}$.

But then $g \in G_{i_a} \cap G_{j_{a'}}$ as both are closed groups. However, since $i_a, j_{a'} \notin \mathcal{I}$ we know that $G_{i_a} \cap G_{j_{a'}} = I$ (as if one were contained in the other then the index of the smaller group would be in $\mathcal{I}$), and so $g = e \in A$. Therefore $A = G \cap H$ and so the set described contains all infimums and so is a lattice. □

Recall from section 2 that Proposition 13 required proof, which we now do here all with a few supporting lemmas.

Lemma 63. *Let $\mathcal{G}$ be a group with subgroups $G, H \leq \mathcal{G}$. Suppose that $f : \mathcal{X} \to \mathbb{R}$ is G-invariant. If $H \leq G$ then f is H-invariant. If f is H invariant then f is $\langle G, H \rangle$-invariant.*

Proof of Lemma 63. The first claim is clear: if $f(g \cdot X) = f(X)$ for all $g \in G$ almost surely then it is true for all $g \in H \subseteq G$. For the second, take $x \in \mathcal{X}$ and $a \in \langle G, H \rangle$. Express $a = a_1^{i_1} \cdots a_k^{i_k}$, where $a_j \in G$ or $a_j \in H$ and $i_j \in \mathbb{N}$ for all $j \in [k]$. Then see that

$$f(a \cdot x) = f(a_1^{i_1} \cdot (a_2^{i_2} \cdots a_k^{i_k} \cdot X)) = f(a_2^{i_2} \cdots a_k^{i_k} \cdot X) \tag{5.26}$$

almost surely so the second claim follows by an induction on k. □

Lemma 64. *Suppose that f is not G-invariant for some subgroup $G \leq \mathcal{G}$. Then f is not H-invariant for any H with $G \leq H \leq \mathcal{G}$. If f is H invariant then f is not G' invariant for any G' with $\langle G', H \rangle = \langle G, H \rangle$.*

Proof of Lemma 64. The first result is just the contraposition of Lemma 63. The second part also follows as if f were G' invariant, then it would be $G \leq \langle G', H \rangle = \langle G, H \rangle$ invariant. □

We can now prove Proposition 13.

Proof of Proposition 13. Let $G_{\max} = \langle H \leq G : f \text{ is } H\text{-invariant} \rangle$, noting that f is I-invariant and so there is at least one such H. Then the previous lemma gives the first condition of maximal invariance, and the second is trivial and implies also uniqueness. □

Proof of Proposition 12. These rules are simply a restating of the Lemmas 63 and 64. □

Proof of Proposition 55. This follows from exactly the same reasoning as Proposition 13, except here we need only consider finitely many groups H and have to take a closure operator:

$$A = \overline{\langle H \in K : f \text{ is } H\text{-invariant} \rangle} \tag{5.27}$$

Since $I \in K$, we know $A \geq I$. Since K is a finite lattice, A must also be in K. Lastly note that Lemma 63 implies that f is A-invariant and so $A \leq G_{\max}(f, \mathcal{G})$. □

5.8.2 Proofs in Section 5.3

Proof of Proposition 56. We check the conditions (R1) - (R3) in order. The first is trivial because each test is a measurable function.

For the second condition, suppose that $G \in \hat{K}_\ell$ has subgroup H. Then if $G_i \leq H$, we must have $G_i \leq G$. Thus we must have $\text{Test}_{\alpha_i}(\mathcal{D}, G_i) = 1$ for all $G_i \leq H \leq G$, so $H \in \hat{K}_\ell$.

Since $\hat{K}_\ell$ always contains the trivial group I we need only consider the case where $G_{\max}(f, K) > I$. Let $\mathcal{I} = \{i \in [\ell] : G_i \leq G_{\max}(f, K)\}$. If $G_{\max}(f, K)$ is not contained in $\hat{K}_\ell$, then at least one of the hypotheses $H_0^{(i)}$ is rejected for some $i \in \mathcal{I}$. Thus:

$$\mathbb{P}\big(G_{\max}(f, K) \notin \hat{K}_\ell\big) = \mathbb{P}\big(\text{Test}_{\alpha_i}(\mathcal{D}, G_i) = -1 \text{ for some } i \in \mathcal{I}\big) \tag{5.28}$$

$$\leq \sum_{i \in \mathcal{I}} \mathbb{P}\big(\text{Test}_{\alpha_i}(\mathcal{D}, G_i) = -1\big) \tag{5.29}$$

$$\leq \sum_{i \in \mathcal{I}} \alpha a_i \leq \sum_{i=1}^{\ell} \alpha_i \leq \sum_{i \in \mathbb{N}} \alpha_i = \alpha \tag{5.30}$$

which gives condition (R3). □

Proof of Theorem 57. First note that $\mathbb{P}_f(\text{Test}_{\alpha_i}(\mathcal{D}, G_i) = 0) \to 1$ for all $G_i \notin G_{\max}(f, K_\ell)$ by the assumption of consistency of the test. Thus:

$$\mathbb{P}_f(\hat{G}_\ell \notin G_{\max}(f, K_\ell)) \leq \mathbb{P}_f(\text{Test}_{\alpha_i}(\mathcal{D}, G_i) = 1 \text{ for some } G_i \notin G_{\max}(f, K_\ell)) \tag{5.31}$$

$$\leq \sum_{G_i \notin G_{\max}(f, K_\ell)} 1 - \mathbb{P}_f(\text{Test}_{\alpha_i}(\mathcal{D}, G_i) = 0) \to 0 \tag{5.32}$$

giving the first statement. Now let P_α be a lower bound on the power of the consistent test used, which must be an increasing sequence (with respect to n) for each fixed α converging to 1. We first construct the sequences of significance levels. Simply take some fixed starting points $\alpha_i(1)$ for $i \in \{1, \ldots, \ell\}$ and set $\alpha_i(n) = \alpha_i(n-1)$ until $P_{\alpha_i(n)} \geq 1 - \alpha_i(n)/2$, then set $\alpha_i(n) = \alpha_i(n-1)/2$. With this sequence of significance levels, we have the convergences:

$$\mathbb{P}_f(\text{Test}_{\alpha_i(n)}(\mathcal{D}, G_i) = 1) \to 1 \qquad \forall G_i \leq G_{\max}(f, K_\ell) \tag{5.33}$$

$$\mathbb{P}_f(\text{Test}_{\alpha_i(n)}(\mathcal{D}, G_i) = 0) \to 1 \qquad \forall G_i \notin G_{\max}(f, K_\ell) \tag{5.34}$$

From these it is clear that TestResults (as defined in algorithm 1) converges in probability to the vector of subgroups of $G_{\max}(f, K_\ell)$.

This then implies that $\mathbb{P}_f(G \in \hat{K}_\ell) \to 0$ for each $G \notin \{G \in K_\ell : G \leq G_{\max}(f, K_\ell)\}$, so we have the convergence in probability of $\hat{K}_\ell$ to $\{G \in K_\ell : G \leq G_{\max}(f, K_\ell)\}$. Since this set

has only one maximal element, $G_{\max}(f, K_\ell)$, we know that $\hat{G}_\ell$ converges in probability to this. □

5.8.3 Proofs in Section 5.4

Proof of Lemma 58. Under the null hypothesis of G-invariance, we know that

$$\begin{aligned} |Y_i - Y_j| &= |f(X_i) - f(X_j) + \epsilon_i - \epsilon_j| && (5.35)\\ &= |f(g \cdot X_i) - f(X_j) + \epsilon_i - \epsilon_j| && (5.36)\\ &\leq |f(g \cdot X_i) - f(X_j)| + |\epsilon_i - \epsilon_j| && (5.37)\\ &\leq V(g \cdot X_i, X_j) + |\epsilon_i - \epsilon_j|, && (5.38) \end{aligned}$$

so we know that $D_{ij}^g = |Y_i - Y_j| - V(g \cdot X_i, X_j) \leq |\epsilon_i - \epsilon_j|$ for all $i, j \in \{1, \ldots, n\}$ and $g \in G$. In particular, this is true for j chosen such that $d(g \cdot X_i, X_j)$ is minimised. If we then count how often D_{ij}^g is larger than some threshold $t \in \mathbb{R}_{\geq 0}$, we can bound the probability using the concentration of $|\epsilon_i - \epsilon_j|$.

For fixed m, the $D_{I(j)J(j)}^{g_j}$ are asymptotically independent (because it is vanishingly unlikely that we sample the same $g_j \cdot X_{I(j)}$ or that two $g_j \cdot X_{I(j)}$ share a nearest neighbour). Thus under the null hypothesis, N_t^g is eventually stochastically bounded by a $\text{Binom}(m, p_t)$ variable. □

We require the following result to simplify the sampling distributions when testing for symmetry in section 5.4.

Lemma 65. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is continuous and G acts continuously on $\mathcal{X}$. If G is topologically generated by a finite set $\{g_1, \ldots, g_k\}$ and f is not G-invariant, then f is not invariant to at least one group generated by a topological generator $\langle g_i \rangle$.*

Proof of Lemma 65. Let $g \in G$ be such that $\mathbb{P}(f(g \cdot X) = f(X)) < 1$, and let $\{h_i\}_{i=1}^\infty$ be a sequence in $\langle g_1, \ldots, g_k \rangle$ that approximates g. Suppose for the sake of contradiction that f is invariant to every $\langle g_i \rangle$. Then $f(h_i \cdot X) = f(X)$ almost surely (as h_i is a finite product of only g_i terms) for every $i \in \mathbb{N}$. But then the continuity of f (noting that $\mathcal{X}$ is a first countable metric space) implies

$$f(X) = \lim_{i \to \infty} f(h_i \cdot X) = f(\lim_{i \to \infty} h_i \cdot X) = f(g \cdot X) \tag{5.39}$$

almost surely - a contradiction. □

The proof of proposition 4.1 relies on several supporting Lemmas. In the following we use the notation $X \xrightarrow{\mathbb{P}} x$ to indicate that the random variable X converges in probability to x.

Lemma 66. *In the context of Proposition 59, if the support of μ_X is dense in $\mathcal{X}$ then*

$$d_{\mathcal{X}}(g \cdot X_{I(1)}, X_{J(1)}) \xrightarrow{\mathbb{P}} 0 \tag{5.40}$$

Proof. For any $\eta \in \mathbb{R}_{\geq 0}$, the probability that $d_{\mathcal{X}}(g \cdot X_{I(1)}, X_{J(1)}) \geq \eta$ given I, g is equivalent to a binomial variable K with n trials and probability of success $p = \mu_X(\mathcal{X} \setminus B(g \cdot X_{I(1)}, \eta))$ being 0. Since the support of μ_X is dense in $\mathcal{X}$, $\mu_X(B(g \cdot X_{I(1)}, \eta)) > 0$, and so

$$\mathbb{P}(d_{\mathcal{X}}(g \cdot X_{I(1)}, X_{J(1)}) \geq \eta) \leq \mathbb{P}(K = 0) = (1-p)^n \tag{5.41}$$

As $n \to \infty$, this clearly goes to 0 for all η, and for all I, g. □

In the following let $F_X(x)$ be the distribution function of the real valued random variable X.

Lemma 67. *If X and Y are independent real random variables, and Y is symmetrically distributed around 0 (so $F_Y(t) = 1 - F_Y(-t)$ for all $t \in \mathbb{R}$) and admits a density f_Y with respect to Lebesgue measure on $\mathbb{R}$ that is decreasing in $|y|$. Then $\mathbb{P}(|X+Y| \geq t) > \mathbb{P}(|Y| \geq t)$.*

Proof. First note $\mathbb{P}(X + Y \geq t) = \mathbb{E}_X(\mathbb{P}(Y \geq t - X \mid X)) = 1 - \mathbb{E}_X(F_Y(t - X))$. Thus

$$\mathbb{P}(|X+Y| \geq t) = \mathbb{P}(X+Y \geq t) + \mathbb{P}(X+Y \leq -t) \tag{5.42}$$

$$= 2 - \mathbb{E}_X(F_Y(t-X) + F_Y(t+X)) \tag{5.43}$$

Let $\phi_t(x) = F_Y(t-x) + F_Y(t+x)$. Since Y is absolutely continuous (w.r.t. Lebesgue measure) with density function f_Y, we can see that ϕ_t has a critical point at $x = 0$. Moreover, if $x > 0$ then

$$\phi_t'(x) = f_Y(t+x) - f_Y(t-x) = f_Y(t+x) - f_Y(-t+x) < f_Y(t+x) - f_Y(-t-x) = 0 \tag{5.44}$$

as f_Y is decreasing in $|y|$. Similarly $\phi_t'(x) > 0$ if $x < 0$. Thus $x = 0$ is a maximum and so we have

$$\mathbb{P}(|X+Y| \geq t) = 2 - \mathbb{E}_X(F_Y(t-X) + F_Y(t+X)) \geq 2 - \mathbb{E}_X(F_Y(t) + F_Y(t)) = \mathbb{P}(|Y| \geq t). \tag{5.45}$$

as required. □

Lemma 68. *If $X \sim Binom(n,p)$ and $Y \sim Binom(n,q)$ with $p < q$, then $F_X(Y) \xrightarrow{p} 1$*

Proof. First note that $F_X(Y) = F_{\tilde{X}}(\frac{Y-np}{\sqrt{np(1-p)}})$ where $\tilde{X} = \frac{X-np}{\sqrt{np(1-p)}}$, which has $\tilde{X} \xrightarrow{D} N(0,1)$ by the de Moivere - Laplace theorem (DLT). Then we also set:

$$\tilde{Y} = \frac{Y-np}{\sqrt{np(1-p)}} = \left(\frac{Y-nq}{\sqrt{nq(1-q)}} + \sqrt{n}\frac{q-p}{\sqrt{q(1-q)}}\right)\frac{\sqrt{q(1-q)}}{\sqrt{p(1-p)}}. \tag{5.46}$$

Again by DLT, the first term converges to $N(0, \frac{q(1-q)}{p(1-p}))$, but the second term diverges to $+\infty$ (as $q > p$). Thus $\tilde{Y} \xrightarrow{p} \infty$ and so clearly $F_{\tilde{X}}(\tilde{Y}) \xrightarrow{p} 1$, as required. □

We can now return to the proof of Proposition 59.

Proof of Proposition 59. Since the support of μ_X is dense in $\mathcal{X}$, Lemma 4 gives $d_{\mathcal{X}}(g \cdot X_{I(j)}, X_{J(j)}) \xrightarrow{p} 0$. Now consider

$$D^{g_j}_{I(j)J(j)} = |Y_{I(j)} - Y_{J(j)}| - V(g \cdot X_{I(j)}, X_{J(j)}) \tag{5.47}$$

$$= |f(X_{I(j)}) + \epsilon_{I(j)} - f(X_{J(j)}) - \epsilon_{J(j)}| - V(g \cdot X_{I(j)}, X_{J(j)}) \tag{5.48}$$

We have that $V(g \cdot X_{I(j)}, X_{J(j)}) \to 0$ as the distance goes to 0, and also $f(X_{J(j)}) - f(g \cdot X_{I(j)}) \xrightarrow{p} 0$. Thus with X as a variable with the same law as each X_i and g with the same law as each g_j,

$$D^{g_j}_{I(j)J(j)} \xrightarrow{D} |(f(X) + \epsilon_\ell) - f(g \cdot X) - \epsilon_k| = |\phi(g, X) + \eta| \tag{5.49}$$

where $\phi(g, X) = f(X) - f(g \cdot X)$ and $\eta = \epsilon_l - \epsilon_k$. Definitionally, this means

$$\mathbb{P}(D^{g_j}_{I(j)J(j)} \geq t) \to \mathbb{P}(|\phi(g, X) + \eta| \geq t) \tag{5.50}$$

Under H_0, $\phi(g, X) = 0$ almost surely, so this probability is given by $P^0_t = \mathbb{P}(|\eta| \geq t)$. Under H_1, ϕ must take non zero values with some positive probability (using the condition on the distribution μ_G). Thus we can use Lemma 67 to say that

$$P^1_t = \mathbb{P}(|\phi(g, X) + \eta| \geq t \mid H_1) > P^0_t \tag{5.51}$$

Let N be large enough that $\mathbb{P}(D^{g_j}_{I(j)J(j)} \geq t) > (P^1_t - P^0_t)/2 > P^0_t$ for all $n > N$.

Now consider that under the alternative hypothesis, $N^g_t \xrightarrow{D} \text{Binom}(m, \mathbb{P}(D^{g_j}_{I(j)J(j)} \geq t))$, which is stochastically bounded from below by $A \sim \text{Binom}(m, (P^1_t - P^0_t)/2)$ for $n > N$. This gives, using Lemma 68, with $A' \sim \text{Binom}(m, P^0_t)$, that

$$p_{val} = 1 - F_{A'}(N^g_t) \leq 1 - F_A(N^g_t) \xrightarrow{\mathbb{P}} 0 \tag{5.52}$$

as required. □

5.9 Related Topics

5.9.1 Estimating Lipschitz Constant of f

As noted in section 5.4.1, if the bound V is not assumed then we may need to estimate it from the data. In the same vein as this test a natural estimate for V is $\hat{L}d(x, y)$, where $\hat{L}$ is

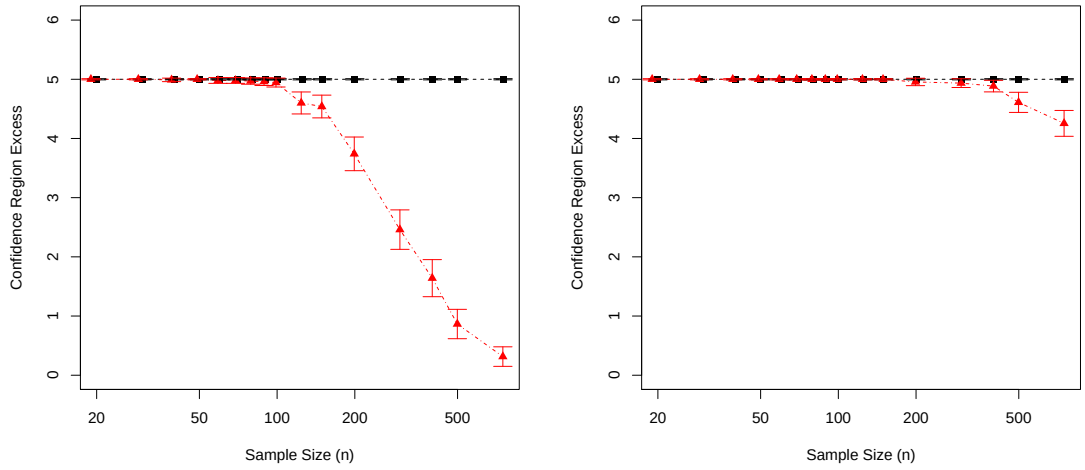
the smallest positive L such that $N_t^e \leq Q_{m,p_t}(\alpha)$ where N_t^e is the same tests statistic before but where μ_g only has support on the identity e and $Q_{m,p_t}(\alpha)$ is the α-quantile of a binomial variable with m trials and probability of success p_t.

Of course, such an estimate is very much not independent of the asymmetric variation test so this could be estimated using a data-splitting approach if needed.

5.9.2 Further Details on the Experiments and Applications

High dimensions for Example 61

As in Example 61, we plot the confidence region excess $\hat{\mathcal{E}}(\hat{K}_\ell)$ for dimensions 6 and 10 in Figure 5.9. We see that the asymmetric variation test is unable to reject any symmetries in these higher dimensional settings, but that the permutation variant can. This is mostly due to the large spacing between the datapoints in these high dimensions. Note that $X_i \overset{iid}{\sim} N(0, 2I_d)$ so the data points are spaced further in higher dimensions.



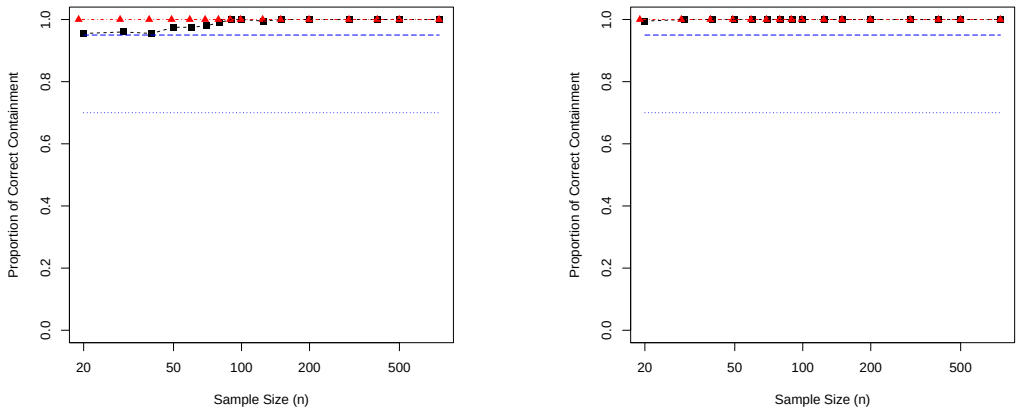
(a) $\mathcal{E}(\hat{K}_\ell)$ when $d = 6$ (b) $\mathcal{E}(\hat{K}_\ell)$ when $d = 10$

Fig. 5.9 A replica of Figure 5.4 for dimensions $d = 6$ and $d = 10$.

Effect of using a fixed α_i in Example 61

As mentioned in Section 5.3.5 and in the variable selection literature [Heinze et al., 2018], one might choose to conduct all tests at the same significance level, i.e., $\alpha_i = a$. We can use the experiments in Example 62 to examine how this choice affects the confidence region. In Figure 5.10 we plot the proportion of the confidence regions that contain the true maximal symmetry $G_{\max}(f, K_\ell)$. Each test was conducted at the significance level $\alpha_i = 0.05$, which

guarantees that $\hat{K}_\ell$ is a 0.7-confidence region for $G_{\max}$ (using the Bonferroni correction). However, we see in all trials the tests were conservative enough to keep the containment proportions within $1 - \alpha_i$ for all simulations.



(a) Dimension 2 (b) Dimension 4

Fig. 5.10 Plot of the frequency with which $G_{\max}(f, K_\ell) \in \hat{K}_\ell$ in the 200 simulations of Example 61. The black squares represent the results when using the asymmetric variation test and the red triangles for the permutation variant. The lower and upper blue lines represent the 0.7 and 0.95 confidence region error thresholds respectively.

Power and size of the hypothesis tests

We now provide the results of each test conducted in Example 61 in order to gauge the empirical size and power of the asymmetric variation test and the permutation variant. These are summarised in tables 5.3 and 5.4. Recall that for each dimension d, the function f_d is invariant to the groups $G_1 = \langle R_h \rangle$, $G_2 = \langle R_v \rangle$, and $G_3 = \langle R_\pi \rangle$ and is not invariant to $G_4 = \langle R_/ \rangle$, $G_5 = \langle R_\backslash \rangle$, or $G_6 = \langle R_{\pi/2} \rangle$.

The results show that both tests have the correct size empirically, and are conservative. The asymmetric variation test is shows its consistency in dimensions 2 and 4, but lacks power at these sample sizes in dimension 6. The permutation variant never rejects a null hypothesis, but has interesting behaviour under the alternatives in dimension 2. We see that the power reaches a plateau at around 0.6 indicating that this approach is potentially not consistent.

Icosahedral vertices

In Example 62 our lattice base consisted of circle groups $S^1_{u_i}$ around axes corresponding to the vertices of a regular icosahedron. To be precise, these axes are given by the normalisation

Table 5.3 Proportions of rejections for the asymmetric variation test in Example 6.1.

Dimension	**Hypothesis**	**Group**	n = **20**	**50**	**100**	**200**	**500**
2	H_0	$\langle R_h \rangle$	0.005	0	0	0	0
		$\langle R_v \rangle$	0.02	0.01	0	0	0
		$\langle R_\pi \rangle$	0.025	0.02	0	0	0
	H_1	$\langle R_/ \rangle$	0.345	0.775	1	1	1
		$\langle R_\backslash \rangle$	0.375	0.785	0.985	1	1
		$\langle R_{\pi/2} \rangle$	0.295	0.735	0.980	1	1
4	H_0	$\langle R_h \rangle$	0.005	0	0	0	0
		$\langle R_v \rangle$	0	0	0	0	0
		$\langle R_\pi \rangle$	0	0	0	0	0
	H_1	$\langle R_/ \rangle$	0.005	0.04	0.050	0.33	0.930
		$\langle R_\backslash \rangle$	0	0.02	0.050	0.29	0.940
		$\langle R_{\pi/2} \rangle$	0	0.02	0.045	0.29	0.915
6	H_0	$\langle R_h \rangle$	0	0	0	0	0
		$\langle R_v \rangle$	0	0	0	0	0
		$\langle R_\pi \rangle$	0	0	0	0	0
	H_1	$\langle R_/ \rangle$	0	0	0	0	0
		$\langle R_\backslash \rangle$	0	0	0	0	0
		$\langle R_{\pi/2} \rangle$	0	0	0	0	0

Table 5.4 Proportions of rejections for the permutation variant in Example 6.1.

Dimension	**Hypothesis**	**Group**	n = **20**	**50**	**100**	**200**	**500**
2	H_0	$\langle R_h \rangle$	0	0	0	0	0
		$\langle R_v \rangle$	0	0	0	0	0
		$\langle R_\pi \rangle$	0	0	0	0	0
	H_1	$\langle R_/ \rangle$	0.00	0.300	0.525	0.64	0.590
		$\langle R_\backslash \rangle$	0.01	0.295	0.475	0.64	0.580
		$\langle R_{\pi/2} \rangle$	0.00	0.285	0.550	0.63	0.565
4	H_0	$\langle R_h \rangle$	0	0	0	0	0
		$\langle R_v \rangle$	0	0	0	0	0
		$\langle R_\pi \rangle$	0	0	0	0	0
	H_1	$\langle R_/ \rangle$	0	0.015	0.155	0.730	0.915
		$\langle R_\backslash \rangle$	0	0	0.185	0.675	0.935
		$\langle R_{\pi/2} \rangle$	0	0.020	0.155	0.720	0.920
6	H_0	$\langle R_h \rangle$	0	0	0	0	0
		$\langle R_v \rangle$	0	0	0	0	0
		$\langle R_\pi \rangle$	0	0	0	0	0
	H_1	$\langle R_/ \rangle$	0	0	0.010	0.120	0.685
		$\langle R_\backslash \rangle$	0	0	0.000	0.155	0.725
		$\langle R_{\pi/2} \rangle$	0	0	0.005	0.095	0.675

of:

$$u_1 = (0, 1, \varphi) \qquad u_2 = (0, 1, -\varphi) \qquad u_3 = (1, \varphi, 0) \tag{5.53}$$

$$u_4 = (1, -\varphi, 0) \qquad u_5 = (\varphi, 0, 1) \qquad u_6 = (\varphi, 0, -1) \tag{5.54}$$

where $\varphi = \frac{1}{2}(1 + \sqrt{5})$ is the golden ratio. In the real world data applications (Section 5.6, we also tested for the circle groups around the additional six axes:

$$u_7 = (1, 0, \varphi) \qquad u_8 = (1, 0, -\varphi) \qquad u_9 = (\varphi, 1, 0) \tag{5.55}$$

$$u_{10} = -\varphi, 1, 0) \qquad u_{11} = (0, \varphi, 1) \qquad u_{12} = (0, \varphi, -1) \tag{5.56}$$

Downloading magnetic field data from VirES

The data in the file "SWARM_DATA.CSV" used in section 5.6.1 was downloaded from the European Space Agency's VirES client (https://vires.services/). The variable names as in our csv file correspond to the variable names that can be selected using the custom download parameters, i.e., F_{NT} is the magnetic field intensity and F_{error} is their reported standard error on this measurement. We have downloaded only the data from satellite "A" on the 25th of February 2023. This data is contained in a csv file in the GitHub reposository linked above.

Q-Q plots in section 5.6.1

Here we check the appropriateness of the sampling distributions for the real data example in Section 5.6.1. As suggested in Section 5.4.2, we examine the Q-Q plots of the marginal distributions of $d(g_i \cdot X_i, g_j \cdot X_j)$ against the original $d(X_i, X_j)$ where the angle of rotation is sampled uniformly on $[0, 2\pi]$. These plots are supplemented by the p-values of Kolmogorov-Smirnoff tests for the equality of these distributions. We see that the uniform sampling angles in Figure 5.11 shift the pairwise distances right (i.e. the black line is above over a neighbourhood of the origin) and so are suitable for use in this case. In these figures the "Group Number" corresponds to the index of G_i in the lattice base.

Q-Q plots in section 5.6.2

Here we check the appropriateness of the sampling distributions for the real data example in Section 5.6.2. As suggested in Section 5.4.2, we examine the Q-Q plots of the marginal distributions of $d(g_i \cdot X_i, g_j \cdot X_j)$ against the original $d(X_i, X_j)$ where the angle of rotation is sampled uniformly on $[0, 2\pi]$. These plots are supplemented by the p-values of Kolmogorov-Smirnoff tests for the equality of these distributions. We see that the uniform sampling angles in Figure 5.11 shift the pairwise distances right (i.e. the black line is above over a

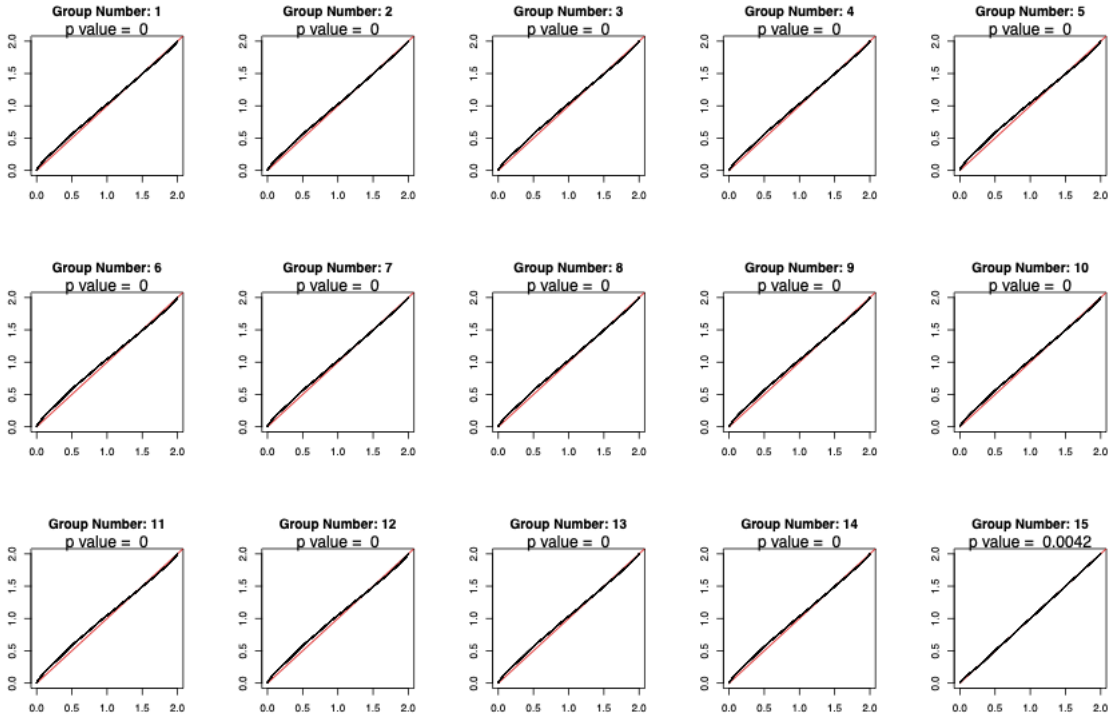


Fig. 5.11 Quantile-Quantile plots of $d_{\mathcal{X}}(X_i, X_j)$ vs $d_{\mathcal{X}}(g_i \cdot X_i, g_j \cdot X_j)$ for each of the groups in Section 5.6.2. Here the sampling of the group elements is from the uniform angle regime, where g_i has an rotational angle sampled from $U([0, 2\pi])$ in radians. Also presented are p-values for Kolmogorov-Smirnov tests for equality of the distributions.

neighbourhood of the origin) and so are suitable for use in this case. In these figures the "Group Number" corresponds to the index of G_i in the lattice base.

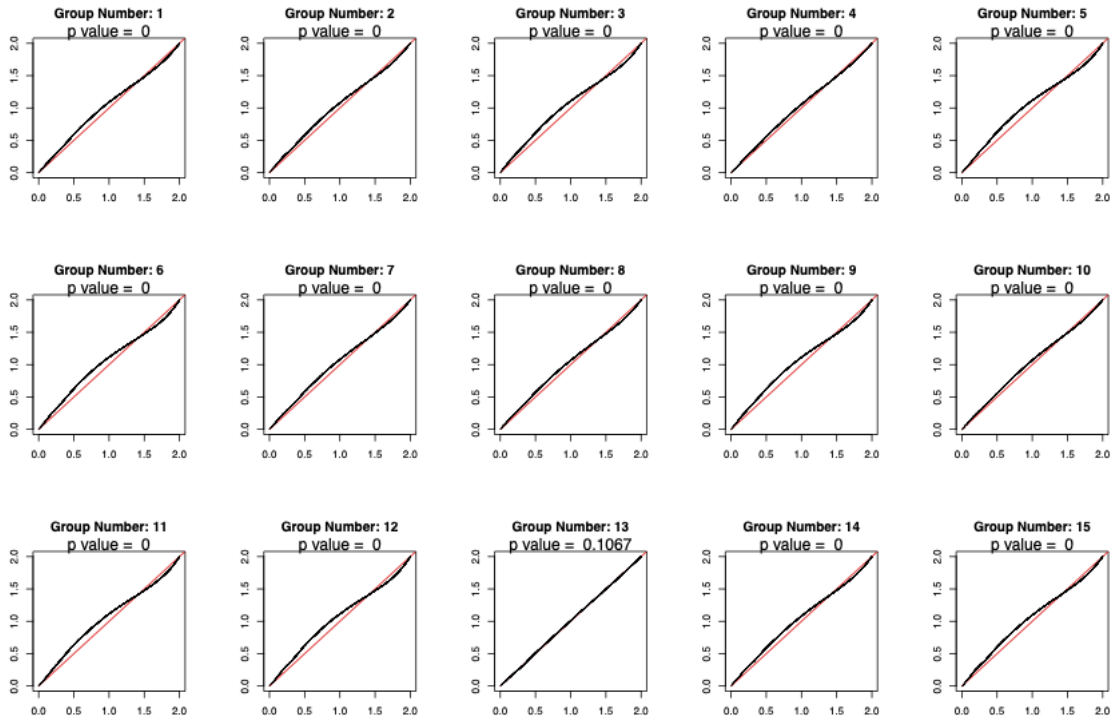


Fig. 5.12 Quantile-Quantile plots of $d_{\mathcal{X}}(X_i, X_j)$ vs $d_{\mathcal{X}}(g_i \cdot X_i, g_j \cdot X_j)$ for each of the groups in Section 5.6.2. Here the sampling of the group elements is from the uniform angle regime, where g_i has an rotational angle sampled from $U([0, 2\pi])$ in radians. Also presented are p-values for Kolmogorov-Smirnov tests for equality of the distributions.

Chapter 6

Conclusion

In this thesis we have examined the benefits of estimating the level sets of regression functions, which can allow us to find adaptive dimension reduction and achieve faster convergence of our estimators. This allows us to generalise the benefits of searching for sparsity to a natural non-linear context.

In Chapter 2 we have summarised the relevant background material on nonparametric regression, differential geometry, and group theory. We also introduced some mathematical results about G-invariant functions and their maximal invariant subgroups. We explored how symmetries have been used in statistics and machine learning, both explicitly and implicitly in many cases. Lastly we briefly covered the previous tools used to estimate sub-level sets of densities and regression functions.

In Chapter 3 we presented the manifold averaged estimator (MAVE) and showed that it reduces dimensions from the regression problem when the level sets are known. This works for every sparse function and can often bring pointwise convergence rates up to the speed of additive regression but without introducing asymptotic bias. When the level sets are unknown we explored a simple method for estimating them that can lead to adaptive dimension reduction in some circumstances. We demonstrated this behaviour in some numerical experiments where we see an order of magnitude improvement in pointwise MSE relative to a local polynomial baseline even when estimating the level set.

In Chapter 4 we restricted ourselves to searching for particular structure in the level sets: orbits of group actions. This is a still a generalisation of covariate sparsity, which is equivalent to a translation invariance of the regression function. We showed that using the MAVE for level sets that contain these orbits can be strengthened to give integrated risk with rates corresponding to (open manifold part of) the lower dimensional quotient space. We developed a metric structure to the space of closed connected subgroups and use this to show

that we can construct an M-estimator of the maximal symmetry of the regression function. This estimator works quickly enough to achieve adaptive dimension reduction, removing dimensions of the orbits even if we do not know them ahead of time. This is demonstrated on some numerical simulations where we see these faster rates occur.

Lastly in Chapter 5 we considered an alternative estimator of the maximal symmetry of a regression function. This approach is based on the lattice structure of the closed subgroups of $\mathcal{G}$. We show that we can construct a lattice base and conduct hypothesis tests for G-invariance along this base in order to build a confidence region of applicable symmetries of the regression function. We present two options for these tests that trade off theoretical guarantees for general applicability. These methods are demonstrated on synthetic data and are applied to two real world data sets: to the strength of the earth's magnetic field and to the locations of sunspot occurrences within solar cycles.

This means that symmetry provides a natural non-linear extension of sparsity into a non-linear framework where we can achieve the same adaptive dimension reduction. However it is still a relatively rigid framework with restricted class of structures within level sets. We now give some extensions to this work that: extend this framework by considering coordinate transformations; consider alternative methods of learning the level set; and use alternative tests for symmetry.

6.1 Future Work

6.1.1 Coordinate Transformation Invariance

Many symmetries are not invariant to changes in the representation of the data. As an example, consider the usual action of $SO(3)$ on $\mathbb{R}^3$ and the regression function $f(x) = x_1^2 + x_2^2 + 0.25x_3^2$. This function is not $SO(3)$ invariant under its usual action, but if we scale the third coordinate (so we have $\tilde{x} = (x_1, x_2, 2x_3)$) then the function in this rescaled coordinate system is the $SO(3)$ invariant $\tilde{f}(\tilde{x}) = \tilde{x}_1^2 + \tilde{x}_2^2 + \tilde{x}_3^2$. This can be viewed as a problem of the choice of action: we could simply have said that $SO(3)$ acts by first applying this coordinate transformation, rotating in this new basis, and inverting the coordinate transform.

To avoid having to make a "good choice" of action we can look to a special case of this problem in covariate sparsity: if f is sparse in some other basis for the covariate space then this might be missed if we only consider the basis we receive the data in. This problem is often solved by considering multi-index models, as considered in Example 28 earlier. In the same way, we might ask whether $f \circ R$ is G-invariant for some coordinate transformation R. Methodologically, this could mean minimising the loss function in equation (4.22) over R

in the M estimation. This would add a large computational load (if using the M-estimator direction) but this could be reduced by only considering symmetries up to equivalence under the transformation. For example, rotations around differing axes would be equivalent because we could choose R such that the rotation axes become aligned.

Estimating the level set directly avoids this problem: if the coordinates are transformed then the level set changes appropriately. This may affect the average performance of the estimator of the level sets but won't affect the worst case (for the class of functions with smooth level sets - any smooth coordinate transform will preserve these).

6.1.2 G-Stable Estimators

In chapter 4 we discussed that the smoothness assumption (Q2), that $2\beta < d^G$ limits the ability of symmetrised estimators to achieve optimal dimension reduction in all cases. The main work that needs to be done to overcome this is to find base estimators that satisfy both assumption set (E) and have the property that symmetrising the data (rather than the estimator) leads to the same asymptotic behaviour. We know that local constant estimators satisfy this for the case that $\beta = 1$, but not for higher order β. For local polynomial estimators it may be the case that we have to change the basis of the polynomial space from the usual monomial basis to another with G-invariant polynomials and orthogonal ones. This is in fact what happens with local constant estimators, where the polynomial basis is just the always invariant $e_1(x) = 1$. However for higher degree polynomial spaces (which are required for assumption set (E) for $\beta > 1$) this basis will be different depending on the symmetry used G.

6.1.3 Other Methods for Estimating the Level Set

As $h_O \to 0$, the level-set near x will converge to its tangent space in the metric d_{x,h_O}. Thus really we could consider only an estimate of the tangent space by looking at the (estimated) derivatives of f on the ball $B(x, h_O)$. The true level set $[x]_f$ will have minimal angle between $\nabla f|_u$ and the normal to this tangent hyperplane, which leads to the estimator for this normal vector:

$$\hat{n} = \operatorname*{argmin}_{n \in S^d} \int_{u \in B(x,h_O)} \frac{\langle \widehat{\nabla f|_u}, n \rangle}{\|\widehat{\nabla f|_u}\|_2} \, \mathrm{d}u \tag{6.1}$$

and then taking $\widehat{[x]_f} = \{x + v : \langle v, n \rangle = 0\}$, where $\widehat{\nabla f|_u}$ is some appropriate estimator of the gradient vector of f at the point u. If the direction of this gradient can be estimated sufficiently quickly we might hope to be able to prove that this estimator achieves dimension adaptive rates. This is essentially a local single index model, that uses global information about the derivatives of f to estimate the projection matrix. This estimator of $[x]_f$ could be extended to higher order approximations if needed.

6.1.4 Other Tests of Invariance

The asymmetric variation test and the permutation variant have two main limitations: they require many samples with small pairwise distances which is sometimes unlikely in high dimensional situations, and they can lack power when the noise level is much greater than the signal. There are two potential solutions to these problems that we discuss here.

Covariate Projections

In order to shorten pairwise distances we could project the data to a chosen submanifold of $\mathcal{X}$. For example, if we wish to test for invariance to rotations about the x-axis (S^1_x) when $\mathcal{X} = \mathbb{R}^3$, we could project the covariates to a hyperplane orthogonal to this axis and test for invariance to the conditional mean of the response given these projections (i.e., we forget the x-coordinate). If the original regression function were S^1_x-invariant then this conditional mean would also be invariant to the rotations of this hyperplane. In general this would work for any μ-measurable idempotent map $\psi : \mathcal{X} \to \mathcal{X}$ with $\psi(g \cdot x) = g \cdot \psi(x)$, as the conditional mean $\mathbb{E}(Y_i \mid \psi(X_i) = x)$ will be G-invariant whenever $\mathbb{E}(Y_i \mid X_i = x)$ is. This means that we would retain a valid test in these cases, as rejection of the G-invariant of the the projected conditional mean is sufficient for rejection of the underlying regression function. However there are some non G-invariant function with G-invariant projected conditional means (consider $f(x, y, z) = (y^2 + z^2) + x \sin(y + z)$ which is not S^1_x-invariant but the projection above creates an S^1_x-invariant conditional mean). This means that this idea would remove power against some alternatives as a price for gaining power in others.

Convex Hull Area

The tests in Chapter 5, particularly the permutation variant, can be heavily affected by pairs of samples with large noise $|\epsilon_i - \epsilon_j|$ and small pairwise distances $d(X_i, X_j)$. This is because the maximum essentially looks only at such points and ignores all of the rest. Instead of computing the maximal slopes $A_0(k)$ and $A_G(k)$ we could instead look to a convex hull of the points $(d(X_i, X_j), |Y_i - Y_j|) \in \mathbb{R}^2$ (perhaps with a fake point $(\infty, \max_{ij} |Y_i - Y_j|)$ added). The area of this hull (integrated with respect to some probability measure on the x axis that might depend on n and the empirical measure of the $d(X_i, X_j)$) would make more use of points with larger pairwise distances and so would measure more of the data. This could help overcome the power plateau observed in Figure 5.4.

6.2 Open Questions

6.2.1 Which Estimators are Sufficiently Local?

Assumption (E1) that $f_n(x)$ and $f_n(y)$ are independent conditioned on $\{X_i\}_{i=1}^n$ for sufficiently large n, is used heavily in the analysis of the MAVE to reduce the variance. This assumption is trivially satisfied for local polynomial estimators using a kernel of compact support, but it, or the variant (E1)' is also likely to hold for other estimation procedures. It would be an interesting question to see whether wavelet basis projection estimators, smoothing splines, or neural networks with appropriate designs have this property as well.

6.2.2 Rates of Estimation of Maximal Symmetries

In the discussion of Chapter 4 we noted that symmetry converge is not bounded uniformly because of the near invariance of some asymmetric functions. If we instead consider the non-uniform rate for the specific regression function f we may find that maximal symmetries can be estimated at a parametric rate (in the local Hausdorff loss function we have considered). This would mirror the results of Ichimura [1987] showing the parametric rate consistency of single index models.

6.2.3 Which Symmetries and Distributions Satisfy Proposition 60?

The permutation variant of the permutation test requires the condition that the pairwise distances "spread out" when the covariate distribution is convolved with the group action, i.e.,

$$\mathbb{P}(g_i \cdot X_i, g_j \cdot X_j) \leq t) \leq \mathbb{P}(d(X_i, X_j) \leq t) \tag{6.2}$$

for $g_i, g_j \overset{iid}{\sim} \mu_g$ and $X_i, X_j \overset{iid}{\sim} \mu$ for all $t \leq T$ for some positive constant T. This condition is likely to hold very broadly because the symmetrisation of the distribution will lessen regions of high density where these small pairwise distances are most likely to occur. It would be interesting to have a complete characterisation of the distributions and groups this works for so that we would not have to judge the suitability from the empirical distributions.

References

Hervé Abdi. Partial least square regression (pls regression). Encyclopedia for research methods for the social sciences, 6(4):792–795, 2003.

Stefan Adams. High-dimensional probability lecture notes. https://warwick.ac.uk/fac/sci/maths/people/staff/stefan_adams/high-dimensional_probability_ma3k0-notes.pdf, May 2020. Accessed: 2023-04-02.

Krizhevsky Alex. Learning multiple layers of features from tiny images. https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf, 2009.

Horace Welcome Babcock. The topology of the sun's magnetic field and the 22-year cycle. Astrophysical Journal, vol. 133, p. 572, 133:572, 1961.

Tünde Baranyi, Lajos Győri, and András Ludmány. On-line tools for solar data compiled at the debrecen observatory and their extensions with the greenwich sunspot data. Solar Physics, 291:3081–3102, 2016.

Ludwig Baringhaus. Testing for spherical symmetry of a multivariate distribution. The Annals of Statistics, pages 899–917, 1991.

Jon L Bentley. Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9):509–517, 1975.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. Advances in Neural Information Processing Systems, 33:17605–17616, 2020.

Kenneth N Berk. Comparing subset regression procedures. Technometrics, 20(1):1–6, 1978.

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. Lecture Notes-Monograph Series, pages 177–186, 2007.

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. Advances in neural information processing systems, 34:18673–18684, 2021.

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. Journal of the American statistical Association, 80(391):580–598, 1985.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478, 2021.

References

Laurent Cavalier. Nonparametric estimation of regression level sets. Statistics: A Journal of Theoretical and Applied Statistics, 29(2):131–160, 1997.

Evangelos Chatzipantazis, Stefanos Pertigkiozoglou, Kostas Daniilidis, and Edgar Dobriban. Learning augmentation distributions using transformed risk minimization. Transactions on Machine Learning Research, 2023.

Kenny Chiu and Benjamin Bloem-Reddy. Hypothesis tests for distributional group symmetry with applications to particle physics. In NeurIPS 2023 AI for Science Workshop, 2023.

Louis G Christie and John AD Aston. Estimating maximal symmetries of regression functions via subgroup lattices. arXiv preprint arXiv:2303.13616, 2023.

Louis G Christie and John AD Aston. Symmetry: a general structure in nonparametric regression. arXiv preprint arXiv:2404.12943, 2024a.

Louis G Christie and John AD Aston. Suplement to "symmetry: a general framework in nonparametric regression". 2024b.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 113–123, 2019.

Antonio Cuevas, Wenceslao González-Manteiga, and Alberto Rodríguez-Casal. Plug-in estimation of general level sets. Australian & New Zealand Journal of Statistics, 48(1): 7–19, 2006.

Tammo T Dieck. Transformation groups. de Gruyter, 1987.

Ian L Dryden and Kanti V Mardia. Statistical shape analysis: with applications in R, volume 995. John Wiley & Sons, 2016.

Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976, pages 85–100. Springer, 1977.

Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. Advances in Neural Information Processing Systems, 34:17273–17283, 2021.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In International Conference on Machine Learning, pages 2959–2969. PMLR, 2021.

Tom Fletcher. Terse notes on riemannian geometry. http://www.sci.utah.edu/~fletcher/RiemannianGeometryNotes.pdf, 2010. Accessed: 2024-02-05.

Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. Journal of the American statistical Association, 76(376):817–823, 1981.

Eduardo García-Portugués, Davy Paindaveine, and Thomas Verdebout. On optimal tests for rotational symmetry against new classes of hyperspherical distributions. Journal of the American Statistical Association, 115(532):1873–1887, 2020.

Thomas Gaudelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy BR Hayter, Richard Vickers, Charles Roberts, Jian Tang, et al. Utilizing graph machine learning within drug discovery and development. Briefings in bioinformatics, 22 (6):bbab159, 2021.

Phillip I Good. Permutation, parametric and bootstrap tests of hypotheses: a practical guide to resampling methods for testing hypotheses. 2005.

Alfred Haar. Der massbegriff in der theorie der kontinuierlichen gruppen. Annals of Mathematics, pages 147–169, 1933.

Thomas Hawkins. Emergence of the theory of Lie groups: An essay in the history of mathematics 1869–1926. Springer Science & Business Media, 2012.

Tristen Hayfield and Jeffrey S Racine. Nonparametric econometrics: The np package. Journal of Statistical Software, 27(5):1–32, 2008. doi: 10.18637/jss.v027.i05. URL https://www.jstatsoft.org/index.php/jss/article/view/v027i05.

Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection – a review and recommendations for the practicing statistician. Biometrical Journal, 60(3):431–449, 2018. doi: https://doi.org/10.1002/bimj.201700067.

Jeff Henrikson. Completeness and total boundedness of the hausdorff metric. MIT Undergraduate Journal of Mathematics, 1(69-80):10, 1999.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.

Sture Holm. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, pages 65–70, 1979.

Marian Hristache, Anatoli Juditsky, Jorg Polzehl, and Vladimir Spokoiny. Structure adaptive approach for dimension reduction. Annals of Statistics, pages 1537–1566, 2001.

Kevin H Huang, Peter Orbanz, and Morgane Austern. Quantifying the effects of data augmentation. arXiv preprint arXiv:2202.09134, 2022.

Zhen Huang and Bodhisattva Sen. Multivariate symmetry: Distribution-free testing via optimal transport. arXiv preprint arXiv:2305.01839, 2023.

Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic pca for riemannian manifolds modulo isometric lie group actions. Statistica Sinica, pages 1–58, 2010.

Hidehiko Ichimura. Estimation of single index models. PhD thesis, Massachusetts Institute of Technology, 1987.

Wen Jiang and Liang Tang. Atomic cryo-em structures of viruses. Current Opinion in Structural Biology, 46:122–129, 2017.

David G Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London mathematical society, 16(2):81–121, 1984.

Arlene KH Kim and Richard J Samworth. Global rates of convergence in log-concave density estimation. The Annals of Statistics, 44(6):2756–2779, 2016.

Israel Kleiner. The evolution of group theory: A brief survey. Mathematics Magazine, 59(4): 195–215, 1986.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In International Conference on Machine Learning, pages 2747–2755. PMLR, 2018.

John Lafferty and Larry Wasserman. Rodeo: Sparse, greedy nonparametric regression. Annals of Statistics, 36(1):28–63, 2008.

Serge Lang. Algebra, volume 211. Springer Science & Business Media, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

John M Lee. Smooth manifolds. In Introduction to smooth manifolds, pages 1–31. Springer, 2013.

Ker-Chau Li. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414):316–327, 1991.

Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. Advances in neural information processing systems, 32, 2019.

Zhenhua Lin, Hans-Georg Müller, and Byeong U Park. Additive models for symmetric positive-definite matrices and lie groups. Biometrika, 110(2):361–379, 2023.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. arXiv preprint arXiv:2005.00178, 2020.

Colin L Mallows. Some comments on c_p. Technometrics, 42(1):87–94, 1973.

Kanti V Mardia and Ian L Dryden. The statistical analysis of shape data. Biometrika, 76(2): 271–281, 1989.

Kanti V Mardia and Peter E Jupp. Directional statistics, volume 494. John Wiley & Sons, 2009.

James Stephen Marron and Ian L Dryden. Object Oriented Data Analysis. Chapman and Hall/CRC, 2021.

John Milnor. Curvatures of left invariant metrics on lie groups. Advances in Mathematics, 21(3):293–329, 1976.

Michael Mitzenmacher and Eli Upfal. Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis. Cambridge university press, 2017.

Anthony P Morse. The behavior of a function on its critical set. Annals of Mathematics, 40 (1):62–70, 1939.

Manuel M Müller, Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Isotonic subgroup selection. arXiv preprint arXiv:2305.04852, 2023.

Elizbar A Nadaraya. On estimating regression. Theory of Probability & Its Applications, 9 (1):141–142, 1964.

John Nash. The imbedding problem for riemannian manifolds. Annals of Mathematics, 63 (1):20–63, 1956.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In NIPS workshop on deep learning and unsupervised feature learning, volume 2011, page 4. Granada, 2011.

Evelien Otte and Ronald Rousseau. Social network analysis: a powerful strategy, also for the information sciences. Journal of information Science, 28(6):441–453, 2002.

Judea Pearl. Causal inference in statistics: An overview. Statistics Surveys, 3:96—-146, 2009.

James Ramsay and Bernard W Silverman. Functional Data Analysis. Springer New York, NY, 2005.

Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Optimal subgroup selection. The Annals of Statistics, 51(6):2342–2365, 2023.

Olinde Rodrigues. Des lois géométriques qui régissent les déplacements d'un système solide dans l'espace, et de la variation des coordonnées provenant de ces déplacements considérés indépendamment des causes qui peuvent les produire. Journal de mathématiques pures et appliquées, 5:380–440, 1840.

Hanamantagouda P Sankappanavar and Stanley Burris. A course in universal algebra. Graduate Texts Math, 78:56, 1981.

Arthur Sard. The measure of the critical values of differentiable maps. Bulletin of the American Mathematical Society, 48(1):883–890, 1942.

Ichirô Satake. On a generalization of the notion of manifold. Proceedings of the National Academy of Sciences, 42(6):359–363, 1956.

Roland Schmidt. Subgroup lattices of groups. In Subgroup Lattices of Groups. de Gruyter, 2011.

Samuel S Shapiro, Martin B Wilk, and Hwei J Chen. A comparative study of various tests for normality. Journal of the American Statistical Association, 63(324):1343–1372, 1968.

Rajesh Sharma, Madhu Gupta, and Girish Kapoor. Some better bounds on the variance with applications. Journal of Mathematical Inequalities, 4(3):355–363, 2010.

Charles J Stone. Consistent nonparametric regression. The Annals of Statistics, pages 595–620, 1977.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. Annals of Statistics, pages 1040–1053, 1982.

Charles J Stone. Additive regression and other nonparametric models. The Annals of Statistics, 13(2):689–705, 1985.

Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. InterStat, 5(16.10):1249–1272, 2004.

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. Advances in Neural Information Processing Systems, 36, 2023.

Erwan Thébault, Christopher C Finlay, Ciarán D Beggan, Patrick Alken, Julien Aubert, Olivier Barrois, Francois Bertrand, Tatiana Bondar, Axel Boness, Laura Brocco, et al. International geomagnetic reference field: the 12th generation. Earth, Planets and Space, 67:1–19, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1):267–288, 1996.

References

Alexandre B Tsybakov. On nonparametric estimation of density level sets. The Annals of Statistics, 25(3):948–969, 1997.

Alexandre B Tsybakov. Introduction to Nonparametric Estimation. Springer Science & Business Media, 2008.

Aad W Van der Vaart. Asymptotic Statistics, volume 3. Cambridge University Press, 2000.

Geoffrey S Watson. Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A, pages 359–372, 1964.

Herman Wold. Estimation of principal components and related models by iterative least squares. Multivariate Analysis, pages 391–420, 1966.

Min Xu and Richard J Samworth. High-dimensional nonparametric density estimation via symmetry and shape constraints. Annals of Statistics, 49(2):650–672, 2021.

Appendix A

Other Useful Lemmas

Lemma 69 (Sum of Squares Inequality)**.** *For any $a \in \mathbb{R}^m$, we have:*

$$\Big(\sum_{i=1}^{m} a_i\Big)^2 \leq m \sum_{i=1}^{m} a_i^2 \tag{A.1}$$

Proof. This is a basic application of the Cauchy-Schwarz inequality: consider

$$m^{-1}\sum a_i = \sum a_i b_i \leq \big(\sum a_i^2\big)^{1/2}\big(\sum b_i^2\big)^{1/2} = \big(\sum a_i^2\big)^{1/2}\big(m^{-1}\big)^{1//2} \tag{A.2}$$

where $b_i = m^{-1}$, then simply square and rearrange. □

Lemma 70. *For any positive constants A, B, a, α, β we have:*

$$An^{-\beta} + B\exp(-an^{\alpha}) \leq (A + B\big(\tfrac{\beta}{a\alpha}\big)^{\frac{\beta}{\alpha-1}})n^{-\beta} \tag{A.3}$$

if $\alpha \neq 1$, and otherwise:

$$An^{-\beta} + B\exp(-an^{\alpha}) \leq (A + B\tfrac{\beta}{a})n^{-\beta} \tag{A.4}$$

Proof. First, note that $\exp(-an^{\alpha})n^{\beta}$ is maximised when $n = n^{\star} = \big(\frac{\beta}{a\alpha}\big)^{\frac{1}{\alpha-1}}$ if $\alpha \neq 1$, so

$$\begin{aligned} An^{-\beta} + B\exp(-an^{\alpha}) &= An^{-\beta} + B\exp(-an^{\alpha})n^{\beta}n^{-\beta} && \text{(A.5)}\\ &\leq An^{-\beta} + B\exp(-a(n^{\star})^{\alpha})(n^{\star})^{\beta}n^{-\beta} && \text{(A.6)}\\ &\leq An^{-\beta} + B(n^{\star})^{\beta}n^{-\beta} && \text{(A.7)}\\ &\leq (A + B\big(\tfrac{\beta}{a\alpha}\big)^{\frac{\beta}{\alpha-1}})n^{-\beta} && \text{(A.8)} \end{aligned}$$

as required. Similarly if $\alpha = 1$ then the maximum of $\exp(-an^{\alpha})n^{\beta}$ occurs at $n = \beta/a$ giving the second bound. □

Lemma 71. *In the random design context, where* $X_i \overset{iid}{\sim} \mu$*, we have:*

$$\mathbb{P}(|B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X| = 0) \leq \exp(-np_{x,h}) \tag{A.9}$$

$$\mathbb{P}(|B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X| \leq \tfrac{np_{x,h}}{2}) \leq \exp(-np_{x,h}/8) \tag{A.10}$$

where $p_{x,h} = \mu(B_{\mathcal{X}}(x,h))$*. In the fixed design context these bounds also hold under the conditions in Chapter 4 for sufficiently large* n*.*

Proof. We can use the Taylor expansion of $\ln(1-x)$ around $x = 0$ to bound the first term:

$$\mathbb{P}(|B_{\mathcal{X}}(x,h) \cap D_X| = 0) = \mathbb{P}(X_i \notin B_{\mathcal{X}}(x,h) \forall X_i) \tag{A.11}$$

$$= (1-p_{x,h})^n \tag{A.12}$$

$$= \exp(n\ln(1-p_{x,h})) \tag{A.13}$$

$$= \exp\big(n\big(-\sum_{k=1}^{\infty} p_{x,h}^k k^{-1}\big)\big) \tag{A.14}$$

$$\leq \exp(-np_{x,h}) \tag{A.15}$$

For the second, we note that $|B_{\mathcal{X}}(x,h) \cap \mathcal{D}_X| \sim \text{Binom}(n, p_{x,h})$ and take the Chernoff bound:

$$\mathbb{P}(N \leq np/2) \leq \exp(-np_{x,h}/8) \tag{A.16}$$

for $N \sim \text{Binom}(n,p)$, a proof of which is in Mitzenmacher and Upfal [2017]. □

Lemma 72. *Let* (X, d_X) *be a totally bounded metric space. Any sub metric space* (Y, d_Y) *of* (X, d_X)*, where* $Y \subseteq X$ *and* $d_Y(y_1, y_2) = d_X(y_1, y_2)$ *for all* $y_1, y_2 \in Y$*, is also totally bounded.*

Proof. Let $\epsilon > 0$ and let $\{x_i\}_{i=1}^m$ be such that $X \subseteq \cup_{i=1}^m B_X(x_i, \epsilon/2)$, which exists by the definition of total boundedness. Let Y be any subset of X and pick $y_i \in B_X(x_i, \epsilon/2) \cap Y$ if this set is non-empty and pick any other $y_i \in Y$ otherwise. Then for all $y \in Y$ we have:

$$d(y, y_i) \leq d(y, x_i) + d(x_i, y_i) \leq \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} \tag{A.17}$$

where x_i is any such that $y \in B_X(x_i, \epsilon/2)$. Thus $\{y_i\}_{i=1}^m$ forms an ϵ-cover of Y as required. □

Lemma 73. *Let* $n \geq 9$ *be an integer and* $p \in (0, 0.5)$*, and suppose that* $A \sim Binom(n,p)$*. The cumulative distribution function of* A *satisfies:*

$$F_A(k) \leq \frac{k}{n} \tag{A.18}$$

for all $1 \leq k < \lfloor n/2 \rfloor$*.*

Proof. Consider first the alternative variable $Z \sim \text{Binom}(n, 1/2)$. We know that

$$F_Z(k) = \frac{1}{2^n} \sum_{m=0}^{k} \binom{n}{m} \tag{A.19}$$

so the result reduces to showing that $\sum_{m=0}^{k} \binom{n}{m} \leq \frac{2^n}{n}(k)$ for all $1 \leq k < \lfloor n/2 \rfloor$. Now note that the left hand side is a convex function in k so it suffices to show that this is true at the endpoints. See that when $n > 5$

$$\sum_{m=0}^{1} \binom{n}{m} = \binom{n}{0} + \binom{n}{1} = 1 + n \leq 2^n/n \tag{A.20}$$

for the first endpoint. For the second, see that if n is even then:

$$\sum_{m=0}^{\frac{n}{2}-1} \binom{n}{m} = \frac{1}{2}\Big(\big(\sum_{m=0}^{n} \binom{n}{m}\big) - \binom{n}{n/2}\Big) \tag{A.21}$$

$$= 2^{n-1} - \frac{1}{2}\binom{n}{n/2} \tag{A.22}$$

$$= \frac{2^n}{n} \times \Big(\frac{n}{2} - \frac{n}{2^{n+1}}\binom{n}{n/2}\Big) \tag{A.23}$$

So if $\frac{n}{2^{n+1}}\binom{n}{n/2} \geq 1$ we have the other end point. This is equivalent to

$$\log_2 \binom{n}{n/2} \geq \log_2 \frac{2^{n+1}}{n} \tag{A.24}$$

which is true for $n \geq 8$. If n is odd then $\lfloor n/2 \rfloor - 1$.

If n is odd then we consider the end point $k = \lfloor n/2 \rfloor - 1$:

$$\sum_{m=0}^{k} \binom{n}{m} = \frac{1}{2}\big(\sum_{m=0}^{n} \binom{n}{m}\big) - \binom{n}{k} \tag{A.25}$$

$$= 2^{n-1} - \binom{n}{k} \tag{A.26}$$

$$= \frac{2^n}{n} \times \Big(\frac{n}{2} - \frac{n}{2^n}\binom{n}{k}\Big) \tag{A.27}$$

See that the bracketed term is less that k for $n \geq 10$ giving the result. □

Lemma 74. *Let $n \geq 8$ be an even integer and let $k = \lfloor n/2 \rfloor - 1$. Then*

$$\frac{n}{2} - \frac{n}{2^{n+1}}\binom{n}{n/2} \leq k \tag{A.28}$$

Proof. This only requires that $\frac{n}{2^{n+1}}\binom{n}{n/2} \geq 1$, which follows from the fact that it is true for $n = 8$ and that it is a increasing function of n. Indeed consider that:

$$\frac{\frac{n}{2^{n+1}}\binom{n}{n/2}}{\frac{n+2}{2^{n+3}}\binom{n+2}{n/2+1}} = \frac{n}{n+2} \times \frac{2^{n+3}}{2^{n+1}} \times \frac{\binom{n}{n/2}}{\binom{n+2}{n/2+1}} \tag{A.29}$$

$$= \frac{4n}{n+2} \frac{n!(n/2+1)!(n/2+1)!}{(n/2)!(n/2)!(n+2)!} \tag{A.30}$$

$$= \frac{4n}{n+2} \frac{(n/2+1)^2}{(n+1)(n+2)} \tag{A.31}$$

$$= \frac{n}{n+2} \frac{(n+2)^2}{(n+1)(n+2)} \tag{A.32}$$

$$= \frac{n}{n+1} \leq 1 \tag{A.33}$$

□

Lemma 75. *Let $n \geq 9$ be an odd integer and let $k = \lfloor n/2 \rfloor - 1$. Then*

$$\frac{n}{2} - \frac{n}{2^n}\binom{n}{k} \leq k \tag{A.34}$$

Proof. First note that $k = (n-3)/2$. Now the statement is equivalent to

$$\frac{3}{2} \leq \frac{n}{2^n}\binom{n}{(n-3)/2} \tag{A.35}$$

Again this is easily verified for $n = 9$ and the right hand side is an increasing function as the ration of sequential terms decays:

$$\frac{\frac{n}{2^n}\binom{n}{(n-3)/2}}{\frac{n+2}{2^{n+2}}\binom{n+2}{(n-1)/2}} = \frac{n}{n+2}\frac{2^{n+2}}{2^n}\frac{\binom{n}{(n-3)/2}}{\binom{n+2}{(n-1)/2}} \tag{A.36}$$

$$= \frac{4n}{(n+2)} \frac{n!((n-1)/2)!(n-(n-1)/2)!}{(n+2)!((n-3)/2)!(n-(n-3)/2)!} \tag{A.37}$$

$$= \frac{4n}{(n+2)} \frac{((n-3)/2+1)!(n-(n-1)/2)!}{(n+1)(n+2)((n-3)/2)!(n-(n-1)/2+1)!} \tag{A.38}$$

$$= \frac{4n}{(n+2)} \frac{((n-3)/2+1)}{(n+1)(n+2)(n-(n-1)/2+1)} \tag{A.39}$$

$$= \frac{n}{n+1} \frac{2n-2}{(n+2)^2(n/2+3/2)} \tag{A.40}$$

$$\leq \frac{n}{n+1}\frac{2}{n+2} \tag{A.41}$$

$$\leq 1 \tag{A.42}$$

for all $n \geq 9$. □