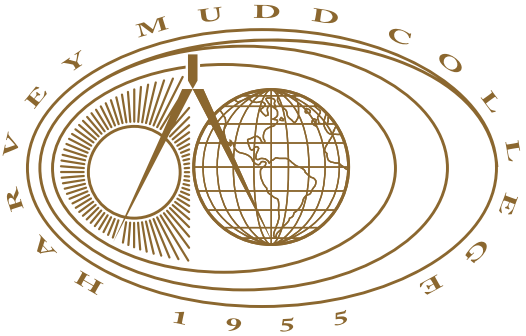
2012

Algebraic Methods for Log-Linear Models

Aaron Pribadi
Harvey Mudd College



Algebraic Methods for Log-Linear Models

Aaron Pribadi

Michael Orrison, Advisor

Weiqing Gu, Reader

May, 2012

HARVEY MUDD
COLLEGE
Department of Mathematics

Abstract

Techniques from representation theory (Diaconis, 1988) and algebraic geometry (Drton et al., 2008) have been applied to the statistical analysis of discrete data with log-linear models. With these ideas in mind, we discuss the selection of sparse log-linear models, especially for binary data and data on other structured sample spaces. When a sample space and its symmetry group satisfy certain conditions, we construct a natural spanning set for the space of functions on the sample space which respects the isotypic decomposition; these vectors may be used in algorithms for model selection. The construction is explicitly carried out for the case of binary data.

Contents

List of Figures

Acknowledgements

I would like to thank Michael Orrison, my advisor, for his guidance over the course of my thesis work. I would also like to thank Weiqing Gu, my second reader, for her time and help. Finally, I must mention that both the Department of Mathematics and the greater Harvey Mudd College community have been instrumental in my growth as a mathematician and as a person these past four years.

Chapter 1

Introduction

Discrete observations model the world in almost the simplest way possible. Things are labelled and put into categories: "This ice cream is green and has sprinkles." Given enough observations, one can begin to make predictions.

Though on the surface this procedure appears to be quite simple, there exists a deep well of problems that are both difficult and intensely practical. The pervasive influence of digital storage and computation has only increased the flood of readily available data. In many cases, we cannot use data as well as we would like. Problems surrounding genomic sequences, image recognition, natural language processing, voting data, product recommendation, and so forth are active areas of research driven by real-world concerns.

Questions arising from the analysis of discrete data pose challenges, but at the same time offer rich mathematical ground. Our hope is that ideas from "pure" mathematics, particularly geometry and algebra, can bring guiding principles and effective techniques for the analysis of data. In return, advances suggested by the needs of practical problems are often of intrinsic mathematical interest.

In this thesis, we examine log-linear models for discrete data. Methods from both representation theory and algebraic geometry have been employed for model selection with log-linear models. Our ideas are similar to those in the book by Diaconis (1988), which also discusses decomposing a space of log-linear models with help from a group action. The inspiration for our approach, however, was the recent appearance of what has been termed "algebraic statistics"; that is, the effort to apply techniques from algebraic geometry to statistics. The lecture notes by Drton et al. (2008) are currently the best introduction to the subject.

The approach to statistics taken here is influenced by the machine learning community; it emphasizes algorithmic methods, large data sets, and models which make few assumptions about the underlying random process. An excellent introduction to machine learning from the statistical point of view is the book by Hastie et al. (2009).

Log-linear models conform to a rigid structure while at the same time encompassing standard techniques for the analysis of discrete data. As is often the case, the presence of a linear structure makes many questions more tractable. The linear structure also allows concerns of symmetry and invariance to be more readily exploited. Objects that are important across different data sets on the same sample space should be invariant under the symmetries of the sample space; it is easier to describe invariant objects that are linear.

We pay special attention to models on the sample space $\{0,1\}^n$ of binary strings of a fixed length. This space is a straightforward way to represent finite data compactly, and as such is often employed as a generic encoding (e.g., in digital electronics). The space of binary strings also has a large amount of symmetry, and is a homogeneous space with respect to its natural automorphism group. We introduce a parametrization for log-linear models on a finite homogeneous space $\mathcal{X}$ under several assumptions, including that the representation $L(\mathcal{X})$ is multiplicity-free. This parametrization respects the isotypic decomposition of $L(\mathcal{X})$. We then construct this parametrization explicitly for the case $\mathcal{X} = \{0,1\}^n$. Such a parametrization is one of several for $\{0,1\}^n$, and can be used in algorithms for finding sparse models.

While the constructions described here for models, algorithms, and bases offer a procedure for the analysis of binary data, we hope that the greater importance of our investigations will be to illuminate the terrain where algebra and statistics interact.

Chapter 2

Discrete Models

In this chapter, we outline an approach to the statistics of discrete data. We then introduce log-linear models, a class which is central to our investigation, and highlight the importance of sparsity. From this exposition, it will become clear that statistical procedures require a few somewhat arbitrary choices. Algebraic concerns, introduced in the following chapter, can help clarify such decisions.

2.1 Discrete Data

A random variable is (from an elementary standpoint) either

- *Discrete*, when its sample space is countable, or
- *Continuous*, when its sample space is a subset of $\mathbb{R}^n$.

We consider the first case. In particular, we usually take the relevant sample space to be finite. Because we can then deal with finite-dimensional spaces, this assumption simplifies many ideas.

Another fundamental assumption is that repeated observations are independent and identically distributed random variables. That is, the observations $X^{(1)}, \ldots, X^{(m)}$ are random variables, and each is distributed according to the same underlying probability distribution, $X^{(i)} \sim p$. Given an observed set of values for the $X^{(i)}$, a basic objective is to estimate the underlying probability distribution p.

Let each $X^{(i)}$ take an observed value from the finite sample space $\mathcal{X}$. Because the order of the observations does not matter, we can summarize

the series of observations with the counts

$$u(x) = (\text{the number of } i \text{ for which } X^{(i)} = x)$$

for $x \in \mathcal{X}$. If the number of possible outcomes $|\mathcal{X}|$ is small relative to the number of samples m, then the empirical distribution

$$p_{\text{emp}}(x) = \frac{u(x)}{m} \tag{2.1}$$

is a useful estimate of the true distribution $X^{(i)} \sim p$.

It is sometimes the case that $|\mathcal{X}|$ is very large. Without a correspondingly large number of samples, the empirical distribution p_{emp} may not adequately capture the underlying distribution p. Such sample spaces arise naturally in a variety of situations, especially when combinatorial processes are involved. Some examples follow.

- Multivariate data occurs when multiple things are observed at once. The sample space is a Cartesian product,

  $$\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k,$$

  of several finite sets. If k labels are assigned to an observation, then the component space $\mathcal{X}_i$ could be the set of possible values for the ith label. For example, a person can have an eye color, a handedness, a gender, a political affiliation, and many more characteristics. In that case, $\mathcal{X}_1$ could be a set of colors, and so forth.

  If each variate has approximately the same number of possible values, then the size of the sample space is exponential in the number of variates.

  A comprehensive reference for the statistical analysis of multivariate data is the book by Bishop et al. (2007).

- Group-valued data occurs when discussing arrangements of something. For example, voters in an election could be asked to rank n candidates. The sample space is then the symmetric group $\mathcal{X} = S_n$. Clearly, it is rather easy to construct simple examples with large numbers of possible outcomes; for example, $n!$ for the symmetric group.

For the empirical distribution to produce a good estimate of the underlying distribution, a prohibitively large number of samples is required.

```
1) n y y n y y n n n n n y y y y
2) n y n y y y n n n n n y y y n y
3) y y y n n n y y y n y n n n y y
4) y y y n n n y y y n n n n n y y
5) y n y n n n y y y y n n n n y y
```

Figure 2.1 A few lines from the Congressional Voting Records data set.

In order to analyze data with a large number of states, it is often fruitful to restrict which distributions we consider. Ideally, we would use the structure of the underlying sample space in order to decide what restrictions to make.

Example 1 (Binary Multivariate Data)**.** The UCI database contains a large number of data sets useful for the evaluation of machine learning techniques (Frank and Asuncion, 2010). The Congressional Voting Records data set contains Congressional voting records on a number of key issues (example shown in Figure 2.1).

For this data set, each variate has two possible values, either the set $\{\texttt{yes},\texttt{no}\}$ or $\{\texttt{democrat},\texttt{republican}\}$. (This ignores missing values; e.g., where a representative did not vote.) The sample space may be written as $\mathcal{X} = \{0,1\}^{16}$, the set of binary strings with 16 bits.

In the above example, the size of the sample space is quite large, with $|\mathcal{X}| = 2^{16} = 65536$. There are 435 samples in the data set. In order to make sense of the data, some assumptions about potential distributions are needed. One rather simplistic assumption would be to assume that that each variate is independent from the others.

2.2 The Simplex and Statistical Models

We take a geometric view of statistical models. The formulation of statistical objects in terms of ideas borrowed from other parts of mathematics allows statistical problems to be attacked with a large range of useful tools. The adoption of this perspective is in large part influenced by the lecture notes by Drton et al. (2008), which summarize recent progress in using algebraic geometry for statistics.

A probability distribution on a finite set $\mathcal{X}$ can be thought of as a real-valued function $p \in L(\mathcal{X})$ subject to the restrictions that $\sum_{x\in\mathcal{X}} p(x) = 1$ and $p(x) \geq 0$ for all $x \in \mathcal{X}$. The function p is the probability mass function.

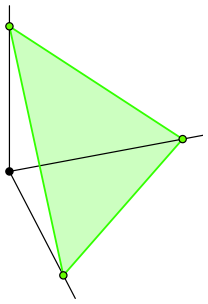


Figure 2.2 The 2-simplex.

Finite sets are particularly convenient because we can always work with a probability mass function, and because the probability mass function is always embedded in an ambient finite-dimensional vector space. The space of all distributions on $\mathcal{X}$ is a geometric object.

Definition 2. The *standard simplex* of dimension n is the subset

$$\Delta_n = \left\{ (p_1, \ldots, p_{n+1}) \in \mathbb{R}^{n+1} \, : \, \sum_{i=1}^{n+1} p_i = 1, p_i \geq 0 \right\}$$

of $\mathbb{R}^{n+1}$. If the appropriate dimension is either clear in context or irrelevant, then we may write Δ, omitting the subscript.

The simplex is a generalization of an equilateral triangle. Low-dimensional simplices are familiar shapes: Δ_0 is a point, Δ_1 is a line segment, Δ_2 is a triangle (see Figure 2.2), and Δ_3 is a tetrahedron.

There is a one-to-one correspondence between probability distributions on the set $\mathcal{X} = \{x_1, \ldots, x_{n+1}\}$ and points in the simplex Δ_n; the probability $p(i)$ is equal to the value of the coordinate p_i. We therefore identify the two concepts with each other. Because the space of all probability distributions is a geometric object, it is easy to talk about specific families of probability distributions.

Definition 3. A *statistical model* is a subset $\mathcal{M} \subset \Delta$ of the probability simplex. A *parametrized model* $\mathcal{M}$ with parameter space Θ is specified by a surjective map $\Theta \to \mathcal{M}$. We usually write p_θ, with $\theta \in \Theta$, to denote a distribution from a parametrized model.

Many statistical models have been studied and employed for the analysis of data, and the selection of an appropriate model is a delicate question. In Section 2.4 we introduce a family of models with convenient properties.

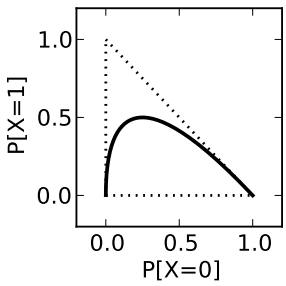


Figure 2.3 The binomial model for $N = 2$.

In an ideal situation, there is a hypothesis for an underlying mechanism producing the observable results. In that case, a particular choice of model is easy to justify.

Example 4 (Binomial Model)**.** The distribution $\text{Binom}(N, \alpha)$ models the number of heads produced by N independent "coin tosses", where $0 \leq \alpha \leq 1$ is the probability that a single toss produces a head. It is defined by the probability density function

$$p_\alpha(k) = \binom{N}{k} \alpha^k (1-\alpha)^{n-k}.$$

for $k \in \{0, \ldots, N\}$.

The map $\alpha \mapsto p_\alpha$ determines a parametrized statistical model. The binomial model is a curve—a one-dimensional subset of the simplex. In the case that $N = 2$, the model $\text{Binom}(2, \alpha)$ simulates two coin flips. The three coordinates of a point in the model measure the probabilities that zero, one, and two heads will occur, respectively. (See Figure 2.3.) As the parameter α varies over $[0, 1]$, the statistical model traces out the curve

$$\alpha \longmapsto \left((1-\alpha)^2, 2\alpha(1-\alpha), \alpha^2\right)$$

in the simplex Δ_2 (see Figure 2.3). If we knew that a coin was being flipped twice and did not know the odds of the coin, then the model $\alpha \mapsto \text{Binom}(2, \alpha)$ would be an appropriate choice for the situation.

Example 5 (Multivariate Independence)**.** The independence model on a sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ assumes that each variate is independent. For a distribution p on $\mathcal{X}$, the variates are said to be independent if p factors as

$$p(x) = p_1(x) \cdots p_k(x) \qquad \text{where} \qquad p_i(x) = \sum_{y_i = x_i} p(y)$$

for all $x = (x_1, \ldots, x_k) \in \mathcal{X}$. Each factor $p_i(x)$ is a function only of the ith component of x.

Despite this simplistic assumption, the independence model is surprisingly effective in many situations. The map

$$\begin{aligned} \Delta_{|\mathcal{X}_1|-1} \times \cdots \Delta_{|\mathcal{X}_k|-1} &\to \Delta_{|\mathcal{X}|-1} \\ (p_1, \ldots, p_k) &\mapsto p_1 \cdots p_k \end{aligned}$$

gives a one-to-one parametrization of the independence model. Counting dimensions indicates that the independence model imposes strict limitations on potential distributions. The parameter space has dimension

$$\dim(\Delta_{|\mathcal{X}_1|-1} \times \cdots \times \Delta_{|\mathcal{X}_k|-1}) = |\mathcal{X}_1| + \cdots + |\mathcal{X}_k| - k,$$

whereas the space of all distributions on $\mathcal{X}$ has dimension

$$\dim(\Delta_{|\mathcal{X}|-1}) = |\mathcal{X}_1| \cdots |\mathcal{X}_k| - 1.$$

If the variates are binary, as in Example 1, then the model has dimension k and the whole simplex has dimension $2^k - 1$. The latter grows much more rapidly with increasing k than the former.

2.3 Likelihood and Model Complexity

In the analysis of data, there is a tension between how well an explanation fits the observed data, and how well such an explanation can be expected to generalize to new data. While we do not explore all nuances of this trade-off, we do introduce a few fundamental concepts.

One way to measure how well a distribution matches data is to ask the question, "How likely is the data, given the distribution?" A maximum likelihood estimate quantifies whether a model contains distributions suitable in that way.

Let $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ be a parametrized statistical model. Suppose that some data $Z = \{z_1, \ldots, z_m\}$ are observed. It is generally assumed that the data has been drawn from independent and identically distributed samples following an unknown true distribution from the model.

Definition 6. The *likelihood function* of $\theta \in \Theta$ given the data Z is

$$L(\theta; Z) = \prod_{i=1}^{n} p_\theta(z_i).$$

It is the probability that the observed data Z would occur if the samples followed p_θ. Oftentimes, the *negative log-likelihood*,

$$-l(\theta; Z) = -\log L(\theta; Z) = -\sum_{i=1}^{n} \log p_\theta(z_i), \tag{2.2}$$

is used in lieu of the likelihood. The negative log-likelihood is useful because the contributions to $-l(\theta; Z)$ from the observations z_i are additive and each acts as a "loss" or "cost". In machine learning, various considerations for decisions are often phrased as costs; this makes them easy to combine.

Definition 7. The *maximum likelihood estimate* of the true parameter is the parameter

$$\theta_{\text{mle}} = \arg\max_{\theta \in \Theta} L(\theta; Z)$$

that maximizes the likelihood of the data, or, equivalently that minimizes the negative log-likelihood of the data. In some cases, we identify the parameter θ with the distribution p_θ. The term "maximum likelihood estimate" then refers to the distribution p_{mle} that maximizes the likelihood, given a choice of model.

The fact that a maximum likelihood estimate might not exist or might not be unique is tacitly ignored. Indeed $\{L(\theta; Z) \mid \theta \in \Theta\}$ might be an open set, and the map $\theta \mapsto L(\theta; Z)$ might not be injective. In applications with real data, it is often the case that a precise maximum likelihood estimate is not necessary, and that an approximate value is sufficient.

The distribution p_{mle} from the maximum likelihood estimate is the best one possible from the model. Notice that if $\mathcal{M} \subset \mathcal{N}$ are two nested statistical models, then the maximal likelihood from $\mathcal{M}$ is less than or equal to that from $\mathcal{N}$. Thus the size of the model determines how good of a distribution is possible.

Example 8. Suppose that the model does not restrict distributions at all—the model is the whole simplex. Then the maximum likelihood estimate is the empirical distribution p_{emp} as in Equation 2.1.

Example 9 (Multivariate Independence)**.** Recall the independence model from Example 5. The maximum likelihood estimate for the independence model is

$$p_{\text{mle}}(x) = \frac{u_1(x_1)}{m} \times \cdots \times \frac{u_k(x_k)}{m},$$

where $x = (x_1, \ldots, x_k)$, m is the number of samples, and $u_i(x_i)$ is the number of times that the ith variate of the observation is x_i. In other words, we get a maximum likelihood estimate of each variate separately with its empirical distribution, and take the product distribution.

A larger model is deemed to contain more "complex" distributions. The trade-off is that a larger model allows for probability distributions that fit the observed data better, but might also allow for over-fitting. When a model is over-fit, the resulting distribution matches the training data (i.e., the data with which the model was fit) very well, but generalizes (i.e., explains subsequently observed data) poorly. The essential problem is that a small number of sample observations do not contain enough information to fit a complex model properly. This trade-off between model complexity and predictive power can be approached in a number of ways, and is explored in the standard literature. Chapter 7 of the book by Hastie et al. (2009) is one reference.

There are at least two common methods to limit model complexity:

- We can require that the estimated distribution p is contained within some small model $\mathcal{M} \subset \Delta$.
- We can minimize $-l(p) + \pi(p)$, the negative log-likelihood with an additional penalty term measuring the complexity of p.

When the penalty term forces estimated distributions into smaller models, these two methods can produce similar results. With a well-selected penalty term, however, the latter method can be scaled easily to favor more or less complex models by adjusting the penalty.

For example, two criteria with complexity penalties are the Akaike information criterion and the Bayesian information criterion, defined as

$$\begin{aligned} \text{AIC} &= -2\, l(p_{\text{mle}}) + 2d \\ \text{BIC} &= -2\, l(p_{\text{mle}}) + (\log m)d, \end{aligned}$$

where $-l(p_{\text{mle}})$ is the negative log-likelihood of the maximum likelihood estimate as in Equation 2.2, d is the dimension of the parameter space Θ, and m is the number of samples (Hastie et al., 2009). These two criteria have different motivations, which are described in the reference. The important point to notice, however, is that the dimension of the model is used as the measure of its size.

2.4 Log-Linear Models

The selection of distributions can be treated as a problem in function estimation. From the data we construct the empirical distribution $p_{\text{emp}} \in L(X)$, and we wish to find another function $p \in L(X)$ that approximates p_{emp} subject to some set of restrictions.

The structure of $L(X)$ as a linear space suggests one method of approximation. One can expand p_{emp} in terms of a basis $B = \{v_1, \ldots, v_n\}$ of $L(X)$, $p_{\text{emp}} = \sum_{i=1}^{n} \lambda_i v_i$. Any subset of the basis elements yields an approximation of p_{emp} from the projection to the subspace spanned by our subset. One can select the terms with the largest coefficients λ_i, or, if the basis has some natural ordering, one can simply truncate the series.

This method of approximation has at least one drawback, namely that negative probabilities are possible. Dealing instead with log-probabilities is often fruitful. In fact, log-linear models (i.e., discrete exponential families) are especially prevalent in the analysis of discrete multivariate data; see, for example, Bishop et al. (2007).

Definition 10. A *log-linear* model $\mathcal{M}_{V,h}$ is a statistical model of the form

$$\mathcal{M}_{V,h} = \{p \in \Delta_{n-1} \ : \log p = (\log p_1, \ldots, \log p_n) \in V + h\},$$

where $h \in \mathbb{R}^n$, V is a linear subspace of $\mathbb{R}^n$, and $V + h$ is an affine subspace.

In many useful situations $h = 0$, so the log-linear model is associated with a vector subspace of $L(\mathcal{X})$. The usual definition of an exponential family (which need not be over a finite sample space) is as follows.

Definition 11. An *exponential family* over a sample space $\mathcal{X}$ parametrized by Θ contains distributions of the form

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp\big(\eta(\theta) \cdot T(x) + h(x)\big),$$

where $\eta : \Theta \to \mathbb{R}^d$, $T : \mathcal{X} \to \mathbb{R}^d$, and $h : \mathcal{X} \to \mathbb{R}$ are known functions, and $Z : \Theta \to \mathbb{R}$ is a normalizing constant known as the *partition function*.

An exponential family is a log-linear model when $\mathcal{X}$ is finite and η is the identity map $\mathbb{R}^d \to \mathbb{R}^d$, as p_θ is constrained to lie in the affine space

$$\text{span}\{(T_i(x_1), \ldots, T_i(x_n)) \ : i \in \{1, \ldots, d\}\} + h$$

where h is interpreted as a vector in $\mathbb{R}^n$.

Example 12 (Binary Independence)**.** The independence model with strictly positive probabilities is a log-linear model. We compute this explicitly for the case where the sample space is $\mathcal{X} = \{0,1\}^2$. One can verify that the two bits are independent for a distribution p on $\mathcal{X}$ if and only if the row span of the matrix

$$\begin{array}{c} \begin{array}{cccc} 00 & 01 & 10 & 11 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

contains $\log p$. Notice that the matrix has rank 3, so that the independence model is a proper subset of the full simplex.

There is another form in which log-linear models are often presented. Define an "energy" function $H : \mathcal{X} \to \mathbb{R}$. The corresponding *Boltzmann* distribution is defined

$$p(x) = \frac{1}{Z} \exp(-H(x))$$

where

$$Z = \sum_{x \in \mathcal{X}} \exp(-H(x))$$

is the normalizing constant. If the possible values for H in a model are all linear combinations of some basis functions, then the model is log-linear. This form is motivated by examples from statistical mechanics, where the probability of a state is proportional to $e^{-E/(k_B T)}$, where E is the energy of the state, T is the temperature, and k_B is the Boltzmann constant.

Example 13. A *Boltzmann machine* on $\{0,1\}^n$ has a distribution of the form

$$p(x) \propto \exp\left(-\sum_{i<j} \beta_{ij} x_i x_j - \sum_i \gamma_i x_i \right).$$

The Boltzmann machine and derived models have recently had many important machine learning applications (Hinton, 2006).

2.5 Sparsity and Regularization

Why are we so keen to cut down the number of dimensions? The reason, which arises in optimization and statistics, is colorfully termed the *curse of dimensionality*. It is easiest to illustrate the problem in another closely related setting.

Example 14 (Sampling and the Curse of Dimensionality)**.** We describe a classification problem with p real-valued features. Concretely, say that we want to label points in the p-dimensional cube $[0,1]^p$ either `blue` or `orange`. We can sample some points in the cube to check their color. A reasonable assumption is that if a point u is `blue`, then another nearby point v (where $\|u - v\|$ is small) is probably also `blue`. This is known as a nearest-neighbor classification algorithm.

When p becomes large, the volume of a neighborhood around a point becomes very small relative to the total volume of the space. If we want every point of the space to contain a sampled point within a fixed radius $\delta > 0$, then we require a number of samples exponential in the dimension.

In statistical learning, this is also known as the "$p \gg N$" problem, where p is the number of features and N is the number of samples. One method for battling the curse of dimensionality is to encourage *sparsity*. Roughly, this means that in a linear problem with many parameters, we want the majority of the parameters to be zero.

Example 15 (Linear Regression)**.** The term "sparse" arises from linear regression. In that setting one wishes to find a relation $y = Ax$, where $x \in \mathbb{R}^N$, $y \in \mathbb{R}$, and A is a linear transformation. A sparse solution is one where many entries of the matrix A are zero; that is, where A is a sparse matrix.

Methods that encourage sparsity have been found to perform well in practice. In what they call the "bet on sparsity", Hastie et al. (2009) introduce a heuristic that explains the importance of sparse methods:

> Use a procedure that does well in sparse problems, since no procedure does well in dense problems.

We must assume that we can find important low-dimensional spaces, because otherwise the estimation of parameters becomes very difficult.

We outline one approach to sparse modeling, the colorfully named Lasso, first introduced in a paper by Tibshirani (1996). Consider a model with real parameters $\beta_1, \ldots, \beta_n$. As before the setting is one of linear regression, fitting $y = \sum_{i=1}^{n} \beta_i x_i$. The objective is to minimize

$$\sum_{j=1}^{m} \left(y^{(j)} - \sum_{i=1}^{n} \beta_i x_i^{(j)} \right)^2 + \lambda \sum_{i=1}^{n} |\beta_i|$$

given data $\{(x^{(j)}, y^{(j)})\}_{j=1}^{m}$ and some hyperparameter $\lambda > 0$. That is, it minimizes the residual sum of squares with an additional penalty term which

is a multiple of the L_1-norm of the parameter vector. As the value of λ is increased, this approach causes some parameters β_i to go to exactly zero, yielding a sparse model. The usage of the L_1 norm as a penalty is also known as L_1 regularization. The procedure derived from using L_2 regularization instead is known as ridge regression; in contrast, it does not bring select parameters to zero, but instead reduces all parameters at the same time.

The approach of L_1 regularization has also been applied directly to log-linear models, in the paper by Buchman et al. (2012). With data $\{z_1, \ldots, z_m\}$ and functions $\{f_1, \ldots, f_n\}$ that span $L(\mathcal{X})$, they minimize

$$-\sum_{j=1}^{m} \log p(z_j; \beta) + \lambda \sum_{i=1}^{n} |\beta_i|$$

where λ is a hyperparameter and the probability distributions are in a log-linear model

$$\log p(z_j; \beta) = \sum_{i=1}^{n}$$

with parameters $\beta_1, \ldots, \beta_n$. That is, they minimize the negative log-likelihood with an L_1 penalty term. This, again, yields a sparse representation.

One should keep in mind that methods for finding sparse models often depend on selecting certain subspaces of $L(\mathcal{X})$ beforehand; L_1 regularization relies on the selection of a full set of basis vectors. Because there are many possible choices for such subspaces, one would like to know that the outcome of the procedure is not too dependent on choices made arbitrarily, and that the selection of subspaces is somehow natural. Under certain circumstances, especially when the sample space is highly structured, such natural choices are possible.

Chapter 3

Decompositions of $L(\mathcal{X})$

In this chapter, we describe algebraic ideas that can help make natural choices from among options posed by the statistical procedures described in the previous chapter. Specifically, we use tools from representation theory to find natural subspaces and vectors in $L(\mathcal{X})$ that can be used for model selection in a log-linear framework. As mentioned in the introduction, this approach is close to that explicated in the book by Diaconis (1988).

3.1 An Invariance Principle

The array of possible models and regularization procedures is large enough that some guiding principles are urgently needed. Symmetries of the sample space, if they exist, suggest some constraints. We shall say that we wish our statistical procedures to be invariant under reparametrization.

More precisely, depending on the nature of the sample space $\mathcal{X}$, there are permutations $\sigma : \mathcal{X} \to \mathcal{X}$ that preserve the structure of $\mathcal{X}$. Such permutations form the automorphism group $\mathrm{Aut}(\mathcal{X})$ of the sample space. Essentially, the automorphism group describes the admissible ways to re-label the data. A statistical procedure is invariant under $\sigma \in \mathrm{Aut}(\mathcal{X})$ if both

- Applying σ to the data, then running the statistical procedure, and
- Running the statistical procedure, then applying σ to the results

produce the same output.

Example 16 (Binary Multivariate Data)**.** Suppose that as in Example 1 the data consist of voting records of individuals on n binary issues. The sample

space is $\{0,1\}^n$. A natural choice for the automorphism group is the hyperoctahedral group. The hyperoctahedral group and its interaction with binary data will be examined more closely in Section 4.1, as it is our primary example.

Under the log-linear framework described in Section 2.4, a model consists of a subspace of $L(\mathcal{X})$ that is sufficient to adequately describe the data. Suppose that one wishes to say that a particular subspace of $L(\mathcal{X})$ is important in general; that is, for the majority of data one encounters for the sample space $\mathcal{X}$, this subspace is necessary. This claim occurs in practice.

Example 17 (Binary Multivariate Data)**.** Let $\mathcal{X} = \{0,1\}^n$. The space $L(\mathcal{X})$ contains subspaces of kth-order effects. For any subset $S \subset \{1,\ldots,n\}$ and function $h : S \to \{0,1\}$, define the function $f_{S,h} \in L(\mathcal{X})$ as

$$f_{S,h}(b_1 \cdots b_n) = \begin{cases} 1 & \text{if } b_i = h(i) \text{ for all } i \in S \\ 0 & \text{otherwise} \end{cases}$$

where the b_i are bits. The function $f_{S,h}$ is an indicator function for bit strings for which the bits indexed by S take on particular values. The space of kth-order effects is the subspace

$$\operatorname{span}\{f_{S,h} : S \subset \{1,\ldots,n\}, |S| = k, f : S \to \{0,1\}\}$$

of $L(\mathcal{X})$. The log-linear model associated with the space of kth-order effects only "cares" about the interaction between at most k variates. Zeroth-order effects produce only the uniform distribution. First-order effects produce the independence model (Example 5). Second-order effects produce the Boltzmann machine (Example 13). The low order spaces are considered more essential than higher order ones.

Every permutation from the automorphism group of $\mathcal{X}$ induces a linear automorphism of $L(\mathcal{X})$. Thus the automorphism group has an induced group action on $L(\mathcal{X})$. If a subspace of $L(\mathcal{X})$ is truly generically important, then it must be invariant under the action of the automorphism group. The good news is that one can get a handle on the invariant subspaces of $L(\mathcal{X})$.

3.2 The Isotypic Decomposition

The action of the automorphism group of $\mathcal{X}$ on $L(\mathcal{X})$ is a permutation representation. In order to explain this statement and its implications, we introduce a few ideas from representation theory.

The goal of representation theory is to better understand the structure of a group (or other algebraic object) by transferring the problem to the domain of linear algebra. The field of representation theory is very rich, so we limit ourselves to a few foundational concepts from the representation theory of finite groups.

Definition 18. A *representation* of a group G is a group homomorphism $\rho : G \to GL(V)$ from G to the group of invertible linear transformations on a vector space V.

This group homomorphism "represents" group elements as invertible matrices. Here we make the simplifying assumptions that G is a finite group and that V is a finite-dimensional vector space. It is common to identify the group representation ρ with the vector space V; one then says that V is the group representation.

One subtlety is that it is possible to make certain strong statements only for complex representations, whereas for probabilities distributions we wish to deal with real vector spaces. In some cases of practical interest, however, real representations behave similarly nicely.

One source of representations, including the representations of interest to us for statistics, is through permutation groups.

Definition 19. A *permutation representation* of a group $G \leq S_n$ on the space $L(\mathcal{X})$ of functions on $\{x_1, \ldots, x_n\}$ is defined by the action

$$(\sigma f)(x_k) = f(x_{\sigma^{-1}k})$$

for all $\sigma \in G$.

Representations of finite groups are well-behaved in large part because they can be broken up into constituent parts.

Definition 20. An *irreducible representation* is a nonzero representation V with no nontrivial (i.e., not $\{0\}$ or V) subspaces invariant under the group action.

Theorem 21 (Maschke)**.** *A complex finite-dimensional representation of a finite group can be written as a direct sum of irreducible representations.*

Theorem 22. *A finite group has only finitely many irreducible complex representations, up to isomorphism.*

As a result, every complex representation can be written in a particular canonical form.

Definition 23. The *isotypic decomposition* of a complex representation V of a finite group G is the direct sum

$$V = \bigoplus_{\rho \in \widehat{G}} m_\rho W_\rho$$

where $\widehat{G}$ is the collection of irreducible representations of G and $m_\rho W_\rho$ is the direct sum of m_ρ copies of the representation ρ. If each m_ρ is either 0 or 1, then the representation is called *multiplicity-free*.

Thus the automorphism group $\mathrm{Aut}(\mathcal{X})$ defines a decomposition of $L(\mathcal{X})$ into a direct sum of invariant complex spaces. For many cases of interest, including binary data, this decomposition also holds over the real field.

3.3 Homogeneous Spaces and K-Invariant Vectors

Sometimes $\mathcal{X}$ exhibits an exceptional amount of symmetry. In the best case, "every point in $\mathcal{X}$ is the same". The following formulation makes this last statement precise.

Definition 24. A *homogeneous space* is a space $\mathcal{X}$ together with a transitive action of a group G on $\mathcal{X}$.

Homogeneous spaces usually turn up in the context of the action of a Lie group on a smooth manifold, in which case the group is required to act by diffeomorphisms. For our purposes, we take $\mathcal{X}$ to be finite and G to be the automorphism group of $\mathcal{X}$.

Given a choice of any point $x_0 \in \mathcal{X}$, let K be the stabilizer of x_0. Then we can identify the set $\mathcal{X}$ with the coset space G/K. If a different point $y = gx_0$ is selected, with $g \in G$, then the stabilizer of y is related to K by an inner automorphism. Specifically, the stabilizer of y is gKg^{-1}.

Definition 25. Let $\mathcal{X} = G/K$ be a homogeneous space. If $L(\mathcal{X})$ is multiplicity-free, then we call (G, K) a *Gelfand pair*.

There are in fact several equivalent characterizations of Gelfand pairs, the statements of which we do not fully explain (Ceccherini-Silberstein et al., 2008).

- Convolution in the space $L(K\backslash G/K)$ of K-bi-invariant functions is commutative.

- $\mathrm{Hom}_G(L(\mathcal{X}), L(\mathcal{X}))$, the algebra of operators intertwining the permutation representation, is commutative.
- For all irreducible representations (ρ, V) of G, we have $\dim V^K \leq 1$, where V^K is the subspace of all K-invariant vectors in V.

In this setting, finding subspaces of $L(\mathcal{X})$ falls under the purview of the field of harmonic analysis. We make use of the paper by Scarabotti and Tolli (2010) on the subject. Specifically, given this setup, we can find distinguished vectors in the isotypic spaces of $L(\mathcal{X})$. The following theorem is a result in Section 2.5 of their paper:

Theorem 26 (Scarabotti and Tolli (2010))**.** *Let (G, K) be a Gelfand pair, so that $L(G/K)$ is multiplicity-free. Let V be an irreducible representation of G contained in $L(G/K)$. Let $V^K = \{v \in V : kv = v \text{ for all } k \in K\}$ be the space of K-invariant vectors in V. Then* $\dim V^K = 1$.

Thus each bijection $\mathcal{X} \cong G/K$ yields a unique K-invariant vector for each irreducible subspace, up to scaling.

Recall that different selections of base point x_0 yield different stabilizers $\mathrm{fix}_G(x_0)$, related by inner automorphisms. We claim that the vectors invariant under some $\mathrm{fix}_G(x_0)$, form spanning sets for each irreducible representation in $L(\mathcal{X})$. We state this precisely, and give a proof.

Theorem 27. *Let G be a group acting transitively on a set $\mathcal{X}$. Let $L(\mathcal{X})$ be the space of complex-valued functions on $\mathcal{X}$; it is a representation of G with the induced action. Fix $x_0 \in \mathcal{X}$, and let $K = \mathrm{fix}_G(x_0)$ be the stabilizer of x_0. Assume that $L(\mathcal{X})$ is multiplicity-free; that is, that (G, K) is a Gelfand pair. Let $L(\mathcal{X}) = W_1 \oplus \cdots \oplus W_m$ be the isotypic decomposition of $L(\mathcal{X})$. Fix any $1 \leq d \leq m$. For each $g \in G$, there exists a gKg^{-1}-invariant vector $f_{d,g}$ in W_d, unique up to scaling. Then $\{f_{d,g} : g \in G\}$ is a spanning set for W_d.*

Proof. Let f be a K-invariant vector in W_d. Because W_d is an irreducible representation of G, it is sufficient to show that

$$Gf = \{gf : g \in G\} \subset \mathrm{span}\{f_{d,g} : g \in G\}.$$

But for any $k \in K$,

$$\begin{aligned}((gkg^{-1})(gf))(x) &= f(g^{-1}(gkg^{-1})^{-1}x) = f(g^{-1}gk^{-1}g^{-1}x) \\ &= f(k^{-1}g^{-1}x) \qquad = f(g^{-1}x) \\ &= (gf)(x)\end{aligned}$$

so gf is gKg^{-1}-invariant. The result follows. □

The significance of this theorem is that we have a spanning set of $L(\mathcal{X})$ which depends (up to scaling) only on $\mathcal{X}$ and its automorphism group, and which respects the isotypic decomposition. From such a spanning set, a sparse log-linear model may be recovered using, for instance, the L_1 regularization procedure explained in Section 2.5.

Chapter 4

Binary Data

The constructions of the previous chapter are somewhat general. In this chapter, we apply those constructions to the case of binary multivariate data. It is first necessary to examine the structure of the hyperoctahedral group, the automorphism group of the set $\mathcal{X} = \{0,1\}^n$ of binary strings of a fixed length. We then derive the isotypic decomposition for $L(\mathcal{X})$ and find the spanning set of K-invariant vectors.

4.1 The Hyperoctahedral Group

Let us devote our attention to the case of binary multivariate data. As before, let the sample space be $\mathcal{X} = \{0,1\}^n$, the set of length n binary strings.

$$\left\{ \underbrace{00\cdots00}_{n \text{ bits}},\ 00\cdots01,\ 00\cdots10,\ \ldots,\ 11\cdots11 \right\}$$

Figure 4.1 2^n binary strings

If we do not assign particular meanings to any of the bits beforehand, then certain re-labellings of the sample space do not change its structure. In particular, there are two natural operations,

- Reordering the bits, and
- Flipping a bit.

Each operation specifies a permutation of $\mathcal{X} = \{0,1\}^n$. If $\mu \in S_n$ is a reordering of the bits, the corresponding permutation of the $\mathcal{X}$ is

$$b_1 \cdots b_n \mapsto b_{\mu(1)} \cdots b_{\mu(n)}.$$

A flip of the kth bit is the permutation

$$b_1 \cdots 0 \cdots b_n \mapsto b_1 \cdots 1 \cdots b_n$$
$$b_1 \cdots 1 \cdots b_n \mapsto b_1 \cdots 0 \cdots b_n$$

where the kth position is modified.

The permutations specified by reorderings and flips of bits are elements of an ambient group, the symmetric group on $\{0,1\}^n$. They generate the subgroup known as the hyperoctahedral group $S_2 \wr S_n$. This group may be defined as the automorphism group of the n-hypercube or as the wreath product (the notation $S_2 \wr S_n$ comes from the second definition). The first formulation is more concrete, so we use it.

First we define a metric on the set of binary strings.

Definition 28. The *Hamming distance* between two binary strings is the number of positions in which the two strings differ.

For example, the following pairs of binary strings have Hamming distances of 0, 1, and 2, respectively:

$$0010 \xleftrightarrow{\text{distance } 0} 0010$$
$$0010 \xleftrightarrow{\text{distance } 1} 0110$$
$$0010 \xleftrightarrow{\text{distance } 2} 0100.$$

It is straightforward to verify that the Hamming distance is in fact a metric. The edges of the hypercube graph are determined by the Hamming metric.

Definition 29. The *n-hypercube graph* has as its vertices the set $\{0,1\}^n$. Two vertices form an edge if and only if their Hamming distance from each other is exactly 1; that is, if the two binary strings differ in exactly one bit.

The graph distance between any two vertices of the hypercube graph—the minimal number of edges in a walk between those two vertices—is the same as their Hamming distance. A look at a picture of a hypercube graph (Figure 4.2) makes the origin of the graph's name clear.

The structure of a graph is determined entirely by its edges, so we get the following definition of a graph automorphism:

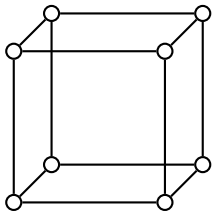


Figure 4.2 The 3-hypercube graph

Definition 30. A *graph automorphism* is a permutation of the vertices of a graph that preserves its edges. That is, if a graph has vertices S and edges E, where each edge is a two-element subset of S, a permutation $\sigma : S \to S$ is a graph automorphism if for every edge $\{u, v\} \in E$, we have $\{\sigma(u), \sigma(v)\} \in E$. The automorphisms of a graph form a group under composition.

Definition 31. The *hyperoctahedral group* H_n is the automorphism group of the n-hypercube.

4.2 Constructing the Isotypic Decomposition

The hypercube graph has a number of nice properties. In particular, it is distance-transitive.

Definition 32. A *distance-transitive* graph is a graph such that given any two vertices u_1 and v_1 at distance d and any other two vertices u_2 and v_2 also at distance d, there exists some automorphism σ of the graph with $\sigma(u_1) = u_2$ and $\sigma(v_1) = v_2$.

The fact that the hypercube graph is distance-transitive allows us to compute the isotypic decomposition of $L(\mathcal{X})$. Recall that $\mathcal{X} = \{0, 1\}^n$ is the set of vertices of the hypercube graph and that the relevant group $S_2 \wr S_n$ is the automorphism group of the graph. Furthermore, $L(\mathcal{X})$ is a representation of $S_2 \wr S_n$ with an action induced by the action on $\mathcal{X}$.

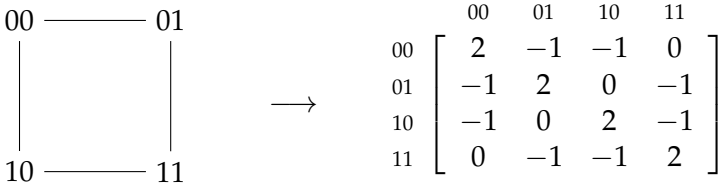
The representation $L(\mathcal{X})$ for a distance-transitive graph decomposes into the eigenspaces of a particular matrix.

Definition 33. The *Laplacian matrix* of a graph with vertices $\{v_1, \ldots, v_m\}$ is the $m \times m$ matrix $L = (l_{i,j})_{i,j=1}^{m}$ defined by

$$l_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

It is also the difference between the degree matrix and the adjacency matrix of the graph.

The Laplacian matrix of the 2-hypercube graph is



Of course, the Laplacian matrix may be viewed as a linear transformation on $L(\mathcal{X})$. Objects like the Laplacian matrix arise in the field of *spectral graph theory*.

The following theorem makes precise the relationship between the Laplacian matrix and the isotypic decomposition of $L(\mathcal{X})$ in the case of a distance-transitive graph. It is a close corollary of a result by Stanton, produced by translating a result for the adjacency matrix into a result for the Laplacian matrix. Because a distance transitive graph is regular, the eigenspaces of the Laplacian matrix and of the adjacency matrix are the same. If each vertex has degree d, then an eigenspace with eigenvalue λ with respect to the adjacency matrix has eigenvalue $d - \lambda$ with respect to the Laplacian matrix.

Theorem 34 (Stanton (1984))**.** *Let $\mathcal{X}$ be the vertices of a distance transitive graph, and let G be the automorphism group of the graph, so that $L(\mathcal{X})$, the space of complex-valued functions on $\mathcal{X}$, is a representation of G. Then the isotypic decomposition of $L(\mathcal{X})$ is multiplicity-free, and is given by*

$$L(\mathcal{X}) = W_{\lambda_1} \oplus \cdots \oplus W_{\lambda_k}$$

where the λ_i are distinct eigenvalues of the Laplacian matrix and W_{λ_i} is the eigenspace corresponding to λ_i. In particular, the eigenspaces of the Laplacian matrix are irreducible representations of G.

The eigenspaces for the Laplacian and adjacency matrices of the n-hypercube graphs have a recursively defined basis.

Theorem 35 (Cook and Wolfe (2006))**.** *Let Q_n be the adjacency matrix of the n-hypercube graph. If v is an eigenvector of Q_{n-1} with eigenvalue λ, then the concatenated vectors $\langle v_1, \ldots, v_{2^{n-1}}, v_1, \ldots, v_{2^{n-1}} \rangle$ and $\langle v_1, \ldots, v_{2^{n-1}}, -v_1, \ldots, -v_{2^{n-1}} \rangle$ are eigenvectors of Q_n with eigenvalues $\lambda + 1$ and $\lambda - 1$ respectively. All eigenvectors of Q_n can be written in this form for some eigenvector of Q_{n-1} when $n \geq 1$.*

Proof. Let Λ_n be the set of eigenvalues of the adjacency matrix Q_n of the n-hypercube graph. Because it has no edges, the adjacency matrix of the 0-hypercube graph is $Q_0 = \begin{bmatrix} 0 \end{bmatrix}$, so $\Lambda_0 = \{0\}$. From Theorem 35,

$$\Lambda_n = \{\lambda + 1 \, : \lambda \in \Lambda_{n-1}\} \cup \{\lambda - 1 \, : \lambda \in \Lambda_{n-1}\}$$

for all $n \geq 1$, so

$$\Lambda_n = \{-n, -n+2, -n+4, \ldots, n\}$$

for all $n \geq 0$. Each vertex of the n-hypercube graph has n neighbors, so the eigenvalues of the Laplacian matrix are

$$\{n - \lambda \, : \lambda \in \Lambda_n\} = \{0, 2, 4, \ldots, 2n\}$$

as desired. □

These results give an explicit algorithm to compute the isotypic decomposition of $L(\mathcal{X})$ for $\mathcal{X} = \{0,1\}^n$. Each irreducible representation contained in $L(\mathcal{X})$ is an eigenspace W_{2k}, where $2k$ is its eigenvalue with respect to the Laplacian matrix.

Begin with the vector $v = (1)$, which is an eigenvector of the matrix for the 0-hypercube. For n steps, we replace v with either (v, v) for $(v, -v)$. To get a vector with eigenvalue $2k$, in $n - k$ steps we need to replace v with (v, v) and in k steps we need to replace v with $(v, -v)$. This procedure yields $\binom{n}{k}$ orthogonal eigenvectors in W_{2k}, and 2^n eigenvectors from all eigenspaces.

The eigenvectors generated from this procedure are of a particular form.

Corollary 36. *An eigenbasis of $L(\mathcal{X})$ with respect to the Laplacian matrix of the n-hypercube graph is given by the so-called Walsh functions on $\{0,1\}^n$.*

$$W_0 = \left\langle \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\rangle \qquad W_2 = \left\langle \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \right\rangle$$

$$W_4 = \left\langle \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \end{bmatrix} \right\rangle \qquad W_6 = \left\langle \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} \right\rangle$$

Figure 4.3 Eigenbasis of $L(\{0,1\}^3)$.

Example 37. The eigenspaces of the Laplacian matrix of the 3-hypercube graph, or equivalently the irreducible representations in $L(\{0,1\}^3)$, have the bases shown in Figure 4.3. A function $f : \{0,1\}^n \to \mathbb{R}$ is represented as a column vector containing the values of f evaluated on each binary string in lexicographic order.

4.3 A Parametrization for Binary Log-Linear Models

We now give a parametrization for $L(\mathcal{X})$ in accordance with Theorem 27. Our explicit computations are for the case $n = 3$, but the procedure is general.

We can construct the isotypic decomposition

$$L(\mathcal{X}) = W_0 \oplus W_2 \oplus \cdots \oplus W_{2n}$$

by the algorithm described in Section 4.2. Each W_{2k} is the eigenspace of the Laplacian matrix with eigenvalue $2k$, and has a basis as in Example 37.

Let the base point in $\mathcal{X}$ be $x_0 = 00\cdots0$. Then the stabilizer $K = \mathrm{fix}_G(x_0)$ in $S_2 \wr S_n$ is the copy of S_n in $S_2 \wr S_n$ that rearranges the bits but does not flip any bit. Notice that under the action of S_n, vectors in any one of the above

$$W_0 = \left\langle \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\rangle \qquad W_2 = \left\langle \begin{bmatrix} 3 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \\ -1 \\ -3 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \\ -1 \\ 1 \\ -1 \\ 1 \\ -3 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 3 \\ 1 \\ -1 \\ -3 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ -1 \\ -3 \\ 3 \\ 1 \\ 1 \\ -1 \end{bmatrix} \right\rangle$$

$$W_4 = \left\langle \begin{bmatrix} 3 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \\ -1 \\ -1 \\ -1 \\ -1 \\ 3 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 3 \\ -1 \\ -1 \\ 3 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ -1 \\ 3 \\ 3 \\ -1 \\ -1 \\ -1 \end{bmatrix} \right\rangle \qquad W_6 = \left\langle \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} \right\rangle$$

Figure 4.4 Spanning set of invariant vectors for $L(\{0,1\}^3)$.

bases are mapped to another vector in the basis. It follows that the sum of the vectors in any basis is K-invariant. Thus the vectors

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \in W_0, \quad \begin{bmatrix} 3 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \\ -1 \\ -3 \end{bmatrix} \in W_2, \quad \begin{bmatrix} 3 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 3 \end{bmatrix} \in W_4, \quad \text{and} \quad \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} \in W_6$$

are all K-invariant. Acting on the vectors with $S_2 \wr S_n$, and identifying vectors which are scalar multiples of each other, we get the spanning sets shown in Figure 4.4.

4.4 Other Parametrizations

Some of the motivation for identifying a spanning set comes from the paper by Buchman et al. (2012). In that paper, the authors consider a number of spanning sets for $L(\mathcal{X})$ where $\mathcal{X} = \{0,1\}^n$. With these spanning sets, the authors fit log-linear models to binary data using L_1 regularization as described in Section 2.5 of this paper.

The spanning sets considered in the paper by Buchman et al. (2012), along with their names for them, are as follows.

The *full parametrization* contains functions of the form

$$f_{S,g}(b_1 \cdots b_n) = \begin{cases} 1 & \text{when } b_i = g(i) \text{ for all } i \in S \\ 0 & \text{otherwise,} \end{cases}$$

where S is any subset of $\{1,\ldots,n\}$ and g is any function $S \to \{0,1\}$. This parametrization contains 3^n functions.

The *canonical parametrization* contains functions of the form

$$f_S(b_1 \cdots b_n) = \begin{cases} 1 & \text{when } b_i = 1 \text{ for all } i \in S \\ 0 & \text{otherwise,} \end{cases}$$

for all $S \subset \{1,\ldots,n\}$. This parametrization contains 2^n functions.

The *spectral parametrization* contains the 2^n Walsh functions on $\{0,1\}^n$ as in Corollary 36.

Our spanning set in Figure 4.4 shares with the spectral parametrization the properties that the set is invariant (as a set) under the action of $S_2 \wr S_n$, and that appropriate subsets of the set generate the irreducible representations in $L(\mathcal{X})$.

Additionally, the vectors generated by this abstract procedure seem to be amenable to admitting interpretation. Each vector is invariant under some stabilizer subgroup, and heavily weights the outcomes fixed under that subgroup. In this case, each stabilizer actually fixes two points of the sample space. In the space of first-order effects, vectors concentrate probability around a single outcome. In the space of second-order effects, vectors have a "polarizing" effect; an outcome and its opposite (e.g., 001 and 110) are both heavily weighted.

Chapter 5

Conclusion

We have put forward an approach to finding sparse log-linear models on sample spaces which contain significant structure. The advantage of an algebraic approach is that it reduces the number of arbitrary choices necessary for a model selection algorithm.

The effect of Theorem 27 is that under certain conditions, we can construct a spanning set of $L(\mathcal{X})$ that is natural, in the sense that it depends only on $\mathcal{X}$ and its automorphism group. In the case of binary data, these vectors also seem to admit interpretation. An algorithm such as that in Section 2.5 can select a small subset of these vectors that together describe the observed data well, restricting model complexity.

5.1 Future Work

Future work could more closely examine the results of this approach on other sample spaces. For example, the symmetric group and quotient spaces of it have applications in voting theory. While the interaction between representation theory and statistics has been examined before in this setting, applications to model construction from the statistical learning perspective have not been, to the author's knowledge.

It is possible other settings are superior to $\{0,1\}^n$ for this type of construction. If the stabilizer of a base point is a smaller group, then there are more K-invariant vectors. In that case, a chain of groups the Gelfand condition yields a basis for the K-invariant vectors; this is a *Gelfand-Tsetlin basis* (Scarabotti and Tolli, 2010). It is possible even that the stabilizer is the trivial group, in which case a basis for all of $L(\mathcal{X})$ is recovered. One potential

issue with such an approach is that the basis will depend on the choice of subgroup chain.

The investigation into binary models was launched by questions about the geometry of a certain mixture model of log-linear binary models, known as the *restricted Boltzmann machine*. It is the model presented in Example 13, with the additional stipulations that some of the variates are hidden and that there are no interactions between visible and hidden variates. The visible distribution is the marginalization over all hidden states. There have recently been advances in understanding the geometry of this model with tools from algebraic geometry and tropical geometry (Cueto et al., 2010), but some questions remain open. In general, the interaction between mixture models and log-linearity remains incompletely understood (Drton et al., 2008). It is possible that concerns of symmetry and sparsity could play a role in better understanding and using such models.

Bibliography

Bishop, Yvonne M., Stephen E. Fienberg, and Paul W. Holland. 2007. *Discrete Multivariate Analysis: Theory and Practice*. Springer.

Buchman, David, Mark Schmidt, Shakir Mohamed, David Poole, and Nando de Freitas. 2012. On sparse, spectral and other parametrizations of binary probabilistic models. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics* .

Ceccherini-Silberstein, Tullio, Daniele D'Angeli, Alfredo Donno, Fabio Scarabotti, and Filippo Tolli. 2008. Finite Gelfand pairs: examples and applications. *Ischia Group Theory 2008 (Proceedings of the Conference)* .

Cook, Michael, and William J. Wolfe. 2006. The hypercube graph and the inhibitory hypercube network. URL http://repository.library.csuci.edu/bitstream/handle/10139/418/Wolfe2006TheHypercubeAuthor~%2b.pdf?sequence=1.

Cueto, Maria, Jason Morton, and Bernd Sturmfels. 2010. Geometry of the Restricted Boltzmann Machine. *Contemporary Mathematics* 516. arXiv:09084425.

Diaconis, Persi. 1988. *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics.

Drton, Mathias, Bernd Sturmfels, and Seth Sullivant. 2008. *Lectures on Algebraic Statistics*. Birkhäuser Basel.

Frank, A., and A. Asuncion. 2010. UCI machine learning repository. URL http://archive.ics.uci.edu/ml.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer, 2nd ed.

Hinton, Geoffrey. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* .

Scarabotti, Fabio, and Filippo Tolli. 2010. Harmonic analysis on a finite homogeneous space. *Proceedings of the London Mathematical Society* .

Stanton, Dennis. 1984. Orthogonal polynomials and Chevalley groups. *Special Functions: Group Theoretical Aspects and Applications, 87-128* .

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* .