Math 197: Senior Thesis

Algebraic Methods for Log-Linear Models

Aaron Pribadi

Harvey Mudd College

Introduction

Problems concerning genetics, image recognition, natural language processing, voting, and product recommendation hinge on making sense of discrete data.

Such problems pose challenges, but also offer rich mathematical structures. We investigate log-linear models, a type that interacts particularly well with methods from geometry and algebra.

Discrete Data

Let $\mathcal{X}$ be a finite sample space.

Observations $X_1, X_2, \ldots, X_m$ are i.i.d. random variables from a true distribution $X_i \sim p$.

```
1) democrat    n y y n y y n n n n n n y y y y
2) republican  n y n y y y n n n n n y y y n y
3) democrat    y y y n n n y y y n y n n n y y
4) democrat    y y y n n n y y y n n n n n y y
```

Figure 1: UCI Congressional Voting Records Data

If $\mathcal{X}$ is large, we need many samples or a model.

Models

The **simplex** Δ_{n-1} contains as points all distributions over $\mathcal{X} = \{x_1, \ldots, x_n\}$.

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

A **statistical model** $\mathcal{M}$ is a subset of the simplex.
A **parametrization** is a surjective map $\Theta \to \mathcal{M}$.
A **maximum likelihood estimate** p_{mle} minimizes

$$-l(\theta; Z) = -\sum_{i=1}^{m} \log p_\theta(z_i).$$

given data $Z = \{z_1, \ldots, z_m\}$.

A model can be selected with a criterion

$$-l(p_{\text{mle}}) + \pi(\mathcal{M})$$

that penalizes 'complex' models.

Log-Linear Models

A **log-linear** model is of the form

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

that is, its log-probabilities fall in an affine space. Sparse models with $\dim V$ small are effective.

A model can be specified with a matrix.

$$\mathcal{M}_{\mathcal{A}} = \{ p \in \Delta_{N-1} \,:\, \log p \in \text{rowspace}(\mathcal{A}) \}$$

Example With $\mathcal{X} = \{0,1\}^2$

$$\mathcal{A} = \begin{array}{c} \begin{matrix} 00 & 01 & 10 & 11 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

$\mathcal{M}_{\mathcal{A}}$ makes the two bits independent.

L_1 Regularization

Let $\{b_1, \ldots, b_n\}$ be a basis for $L(\mathcal{X})$. Choose

$$\log p = \sum_{j=1}^{n} \beta_j b_j$$

to minimize

$$-\sum_{i=1}^{m} \log p(z_i) + \lambda \sum_{j=1}^{n} |\beta_j|$$

where $\lambda > 0$ is a hyperparameter.

The L_1 penalty encourages a sparse structure. This algorithm depends on the choice of basis.

Conditional Monte Carlo Sampling

The vanishing ideal of $\mathcal{M}_{\mathcal{A}} \subset \mathbb{C}^n$ is generated by

$$I(\mathcal{M}_{\mathcal{A}}) = \langle p^u - p^v \,:\, u, v \in \mathbb{N}^d \text{ and } \mathcal{A}u = \mathcal{A}v \rangle.$$

Theorem (Markov Basis) A set of moves $\{b_1, \ldots, b_k\}$ is a Markov basis iff $p^{b_i^+} - p^{b_i^-}$ generates the ideal $I(\mathcal{M}_{\mathcal{A}})$.

The Metropolis-Hastings algorithm takes samples conditioned on the statistic $\mathcal{A}U = \mathcal{A}u$. With samples, a χ^2 test can be applied for the model.

Invariance Principles

Permutations $\sigma : \mathcal{X} \to \mathcal{X}$ rename outcomes.

Under certain σ, we want the results of statistical procedures be invariant. The **automorphism group** G of $\mathcal{X}$ consists of the σ that 'preserve its structure'.

Isotypic Decomposition of $L(\mathcal{X})$

Say that $V \subset L(\mathcal{X})$ is important for the majority of the data that one encounters for $\mathcal{X}$. Then V should be invariant under the action of G.

As a representation, $L(\mathcal{X})$ decomposes into **isotypic** components

$$L(\mathcal{X}) = \bigoplus_{\rho \in \widehat{G}} m_\rho W_\rho.$$

If $L(\mathcal{X})$ is multiplicity-free, then the decomposition of $L(\mathcal{X})$ into irreducibles is unique.

Homogeneous Spaces

Let the action of G on $\mathcal{X}$ be transitive. Then $\mathcal{X}$ is a **homogeneous space** and $\mathcal{X} \cong G/K$ where K is the stabilizer of some $x_0 \in \mathcal{X}$.

If (G, K) is a Gelfand pair, then $L(\mathcal{X})$ is multiplicity free. The series of groups $K \leq G$ induces **Gelfand-Tsetlin** bases for the K-invariant vectors in each isotypic compnent.

Binary Models

Let $\mathcal{X} = \{0,1\}^k$. We want invariance under

- flips of bits, and
- rearrangements of bits.

These make the **hyperoctahedral** group $S_2 \wr S_n$.

The irreducibles of $S_2 \wr S_n$ in $L(\mathcal{X})$ are the eigenspaces of the Laplacian matrix of the hypercube graph. They represent kth-order effects.

Boltzmann Machines

A **Boltzmann Machine** is a model on $\{0,1\}^n$ contained in the eigenspaces for $\lambda = 0, 2, 4$. More concretely, it is a log-linear model governed by at most 2nd-order interactions.

$$H(x) = -\sum_{i<j} \beta_{ij} x_i x_j - \sum_i \gamma_i x_i$$

$$p(x) \propto \exp(-H(x))$$

A derived model used in machine learning is the **Restricted Boltzmann Machine**.

$$H(x,h) = -\sum_{i<j} \beta_{ij} x_i h_j - \sum_i \gamma_i x_i - \sum_j \delta_j h_j$$

$$p(x) \propto \sum_{h \in \{0,1\}^k} \exp(-H(x,h))$$

It contains hidden and visible nodes, with interactions only between hidden and visible nodes. The visible distribution is the marginalization.

In general, making **mixture models** is a powerful way to combine log-linear models. The geometry of the situation, where weighted sums occur before and after the exponential map, is incompletely understood.

Acknowledgments

I would like to acknowledge Professor Michael Orrison for the introduction to the field of algebraic statistics, and for his encouragement and forbearance while I worked on my thesis.

For Further Information

- Contact the author at `aaron.pribadi@gmail.com`.
- The poster and thesis document are available electronically at `http://www.math.hmc.edu/~apribadi/thesis/`.

Advisor: Michael Orrison **Reader:** Weiqing Gu