Autologistic Regression Models, With Application to Segmentation of Hyperspectral Satellite Imagery

Mark Wolters, Shanghai Center for Mathematical Sciences
Charmaine Dean, Western University

November 27, 2015
Shanghai Jiao Tong University
Institute of Natural Sciences
The technical content:

Autologistic regression for binary variables $C$, with complex association, and covariate information $X$.
- There are a few model variants.
- Claim: *it matters* which one you choose.

The application:

Where is the smoke in this picture?
Motivating application
Remote sensing for smoke monitoring

Earth-orbiting satellites help study large-scale environmental phenomena.

Our interest: smoke from forest fires.

**Data:** MODIS images
- 1 per day, 143 days
- 1.2 Mp each
- Centered at Kelowna, BC
- Hand-drawn smoke areas

**Goal:** classify pixels into smoke/nonsmoke

**Why?**
- Health studies
- Model input or validation
- Monitoring & archiving
Data characteristics

- **Binary responses** (smoke/nonsmoke).
- Spectra at each pixel are covariates for predicting smoke.
- **Hyperspectral images**: a high-dimensional predictor space.
- Expect **spatial association**.
The true class label for pixel $i$ is $C_i$. The full set of $N$ class labels (the true scene) is $C$. The image features (predictors) for pixel $i$ are $x_i$. The full set of features (the observed image) is $X$. 
Autologistic regression models
Spatial Associations

Image segmentation = pixel classification.
If independent pixels ⇒ Use standard classification technology.
But smoke/nonsmoke regions are *spatially smooth*.
Many *ad hoc* ways to let pixels influence each other.

Model-based approach: *Markov random fields (MRFs).*

- graphical model
- popular in computer vision

**Generative model:**
\[ p(C|X, \theta) \propto p(X|C, \theta)p(C, \theta) \]

**Discriminative model:**
\[ p(C|X, \theta) = \eta(X, \theta) \quad \Leftarrow \text{regression} \]

We will use the discriminative approach.

Model \( p(C|X, \theta) \) directly as a MRF.
Hammersley–Clifford theorem

Joint PMF can be expressed as a product of potential functions, one for each maximal clique.

\[ M = \text{the set of maximal cliques.} \]
\[ C_m = \text{the variables in the } m^{\text{th}} \text{ clique.} \]

Then

\[ p(c) = \frac{1}{Z} \prod_{m \in M} \phi_m(c_m) \]

Customary to write the joint density as Gibbs distribution form,

\[ p(c) \propto e^{Q(c)} \]

where \( Q(\cdot) \) is the negpotential function.
• A Markov random field of binary random variables.

• For now, use zero/one coding: $C_i \in \{0, 1\}$.

• The graph is a regular, square grid.

• Nothing new
  Physics: Ising model
The Autologistic Model (2)

**Joint PMF:**

\[
\Pr(C = c|\alpha, \lambda) = \frac{1}{Z(\alpha, \lambda)} \exp \left( \sum_{i \in V} \alpha_i c_i + \sum_{(i,j) \in E} \lambda_{ij} c_i c_j \right)
\]

- Positive \(\alpha_i\) values favor +1 class.
- Setting \(\lambda_{ij} > 0\) favors *locally smooth* configurations \((C_i = C_j)\).
- Typically set \(\lambda_{ij} = \lambda, \forall i, j\)

**Conditional distributions:**

Let \(\pi_i = \Pr(C_i = 1|\text{all other } C)\). Then can show:

\[
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \alpha_i + \sum_{j \sim i} \lambda_{ij} c_j
\]

\(j\) is a neighbour of \(i\)
Autologistic Regression

Put covariates in the unary part: $\alpha_i = x_i^T \beta$.

Pairwise coefficients: $\lambda_{ij} = \lambda$.

Then:

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = x_i^T \beta + \lambda \sum_{j \sim i} c_j.$$  

Interpretation:

- Unary part is a linear predictor.
- $x_i^T \beta$ determines conditional log-odds of $C_i = +1$ in the absence of spatial effects.
- Pairwise $\lambda$ determines strength of neighbour effects.
- Setting $\lambda = 0$ reverts to standard logistic regression.
The neighbour sum $\sum c_j$ increases log-odds unless all neighbours are zero. 
$\implies$ Estimates of $\beta, \lambda$ are strongly coupled

Strongly recommended a centered autologistic model:

$$
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = x_i^T \beta + \lambda \sum_{j \sim i} (c_j - \mu_j)
$$

where $\mu_j = E[Y_j | \lambda = 0]$ is the independence expectation.
Estimation Issues

Estimation is made hard by the normalizing constant.

Existing possibilities

1. Ignore spatial association (logistic regression, large $n$, large $p$).

2. Pseudolikelihood (PL):  
\[
L(\beta, \lambda) \approx \prod_{\text{img}} \prod_{i=1}^{n} \logit(\pi_i)
\]

3. Monte Carlo ML

4. Bayesian approach

Problems

- We have $\sim 10^8$ pixels
- We have thousands of predictors, need model selection
- We’re still developing models—rapid evaluation of candidates is beneficial

Hughes et al. use perfect sampling; recommend PL for large $n$. 
Some claims about coding and centering
Model variants

- Binary variables do not have to take values 0 and 1.
- In general, let them have coding \( \{ \ell, h \} \)
- We’re most interested in
  \( \{0, 1\} \), used in statistics and sometimes in computer vision
  \( \{-1, 1\} \), used in physics and sometimes in computer vision
- And we have two possibilities:
  centered, or
  standard (not centered)

*Are all these model variants just parameter transformations, or are they distinct models?*

*If not the same, what are the differences?*

In the following, say \( f_1(z; \theta_1) \) and \( f_2(y; \theta_2) \) are *equivalent* if for any \( \theta_2 \) there exists a \( \theta_1^* \) such that the two models assign the same probability whenever \( z \) and \( y \) represent the same configuration.
A general form of the model

If we let the coding be \( \{\ell, h\} \) and define the centering adjustment

\[
\mu_\alpha = \begin{cases} 
0 & \text{for a standard model} \\
[\mu_1^\alpha \cdots \mu_n^\alpha]^T & \text{with } \mu_i^\alpha = \frac{\ell e^{\ell \alpha_i} + h e^{h \alpha_i}}{e^{\ell \alpha_i} + e^{h \alpha_i}} & \text{for a centered model}
\end{cases}
\]

Then the autologistic (AL) PMF is

\[
f_C(c; \alpha, \lambda) \propto \exp(c^T \alpha - \lambda c^T A \mu_\alpha + \frac{\lambda}{2} c^T A c),
\]

where \( A \) is the adjacency matrix.

And the logit form is

\[
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = (h - \ell) \left( \alpha_i + \lambda \sum_{j \sim i} (c_j - \mu_j^\alpha) \right)
\]

To obtain the autologistic regression (ALR) model, just plug in \( \alpha = X\beta \)
Autologistic models are equivalent

Theorem 1
All AL models are equivalent, irrespective of coding or centering.

Sketch of proof:
• If $z$ and $y$ have different coding, they are related by
  \[ z = ay + b1 \]
  whenever they represent the same configuration.
• Use this to equate PMFs of $z$ and $y$
• Obtain an explicit parameter transformation from one model to the other
Autologistic regression models are NOT equivalent

**Theorem 2**

Different ALR models are not, in general, equivalent.

**Sketch of proof:**

- Follow logic of Theorem 1.
- Transformation between model only exists if an overdetermined system ($n$ equations, $p$ unknowns) can be solved for $\beta$.
- Coefficients of the system depend on arbitrary $X$.
- System is linear for standard models, nonlinear for centered ones.

This means

- standard, zero/one
- standard, plus/minus
- centered, zero/one
- centered, plus/minus

Are *four different probabilistic structures.*
Advantages of standard, plus/minus model

**Claim:** the standard model with \{−1, 1\} coding is the best choice.

**Why?**

1. It resolves the asymmetry of the standard model without the awkward $\mu_\alpha$ term:

$$
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = 2 \left( x_i^T \beta + \lambda \sum_{j \sim i} c_j \right)
$$

   Sign of pairwise term depends on majority vote.

2. It “decouples” $\beta$ and $\lambda$ better than centering (evidence to follow).

2. It allows a convenient *plug-in estimation* of $\lambda$ ...
An estimation shortcut for $\lambda$

Proposal: plug-in estimation

a) Use independence (logistic) to get $\hat{\beta}$
   - Including model selection
   - Sample pixels if necessary to reduce $n$ to manageable size

b) Choose $\hat{\lambda}$ to optimize predictive power

Rationale

Treat $\lambda$ as a smoothing parameter.

- Assuming independence, $\hat{\beta}$ captures how information in $X$ can be used to predict $C$.
- For fixed $\hat{\beta}$, tuning $\lambda$ will optimally reduce noise in the predicted probabilities.
Checking the claims
A small example

$n = 9$ variables, square graph.

One predictor plus intercept.

- Small problem.
- Can compute probabilities directly.

Linear predictor vector is

$$\alpha = X\beta = \begin{bmatrix} 1 & x_o \\ 1 & x_o \\ 1 & x_o \\ 1 & x_o \\ 1 & x_5 \\ 1 & x_o \\ 1 & x_o \\ 1 & x_o \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

- Choose values of $x_o, x_5, \beta_0, \beta_1$
- Get marginal $\Pr(C_5 = \text{high})$ as a function of $\lambda$, for four models:
  - standard, zero/one
  - standard, plus/minus
  - centered, zero/one
  - centered, plus/minus
Small example: case 1

\( \beta = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \). Linear predictors: 0 0 0

\[
\begin{array}{cccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
0.5 & 1 & 1.5 & 2 & 2.5 & 3 \\
0.7 & 0.8 & 0.9 & 1 & 1 & 1 \\
0.6 & 0.7 & 0.8 & 0.9 & 1 & 1 \\
0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

Marginal probability of center point vs. \( \lambda \)

Shows the asymmetry of the standard 0/1 model.
Small example: case 2

\[ \beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x_o = 1, \quad x_5 = 1. \]

Linear predictors:

\[
\begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{bmatrix}
\]

Standard and centered models give opposite behaviour as \( \lambda \to \infty \).

Contradicts the idea that centered models have “interpretable parameters.”
Small example: case 3

\[ \beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ x_o = 1, \ x_5 = -1. \]

Linear predictors:

\[
\begin{array}{ccc}
1 & 1 & 1 \\
1 & 0 & 1 \\
1 & 1 & 1 \\
\end{array}
\]

Again, standard and centered models disagree.

Does the centered model behaviour make sense?
Small example: case 4

\[ \beta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x_o = 0, \quad x_5 = -1. \]

Linear predictors:

\[
\begin{bmatrix}
0 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & 0 
\end{bmatrix}
\]

Standard plus/minus very different from the other three.

Limiting probability not equal to 1 or 0.

\( C_5 \) probability constant wrt \( \lambda \).
Generated RGB images with characteristics similar to the smoke data.

- 5 sizes: $100^2$, $200^2$, $400^2$, $600^2$, $800^2$ pixels.
- 90 images at each size.
- Training, validation, and test groups.
Parameter estimates and prediction error: plug-in vs. pseudolikelihood (plus-minus coding).

<table>
<thead>
<tr>
<th>pixels</th>
<th>method</th>
<th>$\hat{R}$</th>
<th>$\hat{G}$</th>
<th>$\hat{B}$</th>
<th>$\hat{\lambda}$</th>
<th>error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$100^2$</td>
<td>plug-in</td>
<td>-2.21</td>
<td>-2.02</td>
<td>1.91</td>
<td>0.90</td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td>PL</td>
<td>-2.04</td>
<td>-1.99</td>
<td>2.06</td>
<td>0.99</td>
<td>20.4</td>
</tr>
<tr>
<td>$200^2$</td>
<td>plug-in</td>
<td>-1.64</td>
<td>-1.35</td>
<td>1.71</td>
<td>1.00</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>PL</td>
<td>-1.61</td>
<td>-1.30</td>
<td>1.70</td>
<td>1.19</td>
<td>17.7</td>
</tr>
<tr>
<td>$400^2$</td>
<td>plug-in</td>
<td>-2.05</td>
<td>-1.42</td>
<td>1.63</td>
<td>1.60</td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td>PL</td>
<td>-2.08</td>
<td>-1.40</td>
<td>1.68</td>
<td>1.36</td>
<td>20.1</td>
</tr>
<tr>
<td>$600^2$</td>
<td>plug-in</td>
<td>-1.91</td>
<td>-1.22</td>
<td>1.76</td>
<td>1.95</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td>PL</td>
<td>-1.97</td>
<td>-1.36</td>
<td>1.79</td>
<td>1.51</td>
<td>20.4</td>
</tr>
<tr>
<td>$800^2$</td>
<td>plug-in</td>
<td>-1.55</td>
<td>-1.44</td>
<td>1.58</td>
<td>1.95</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>PL</td>
<td>-1.57</td>
<td>-1.43</td>
<td>1.49</td>
<td>1.59</td>
<td>18.6</td>
</tr>
</tbody>
</table>

- Parameter estimates similar.
- Error rates similar.
Example predictions, 800 × 800 image.
What about plugin with different model variants?
Analysis of the smoke data
Analysis flowchart

- training images
  - sample
  - training pixels
    - logistic regression
    - model selection
      - fitting
        - performance evaluation
          - best independence model
- validation images
  - sample
  - validation pixels
    - choose to minimize prediction error
      - autologistic regression
        - final performance results
- test images
  - best autologistic model
    - final performance results
Results 1

Use a logistic GAM

- Each variable or interaction is a piecewise linear function
- Model search by genetic algorithm

<table>
<thead>
<tr>
<th>Predictor set</th>
<th>Selected variables (MODIS band numbers)</th>
<th>plug-in $\hat{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>main effects</td>
<td>1 6 7 8 14 16 17 18 21 23 25 26 30 31 32 36</td>
<td>1.85</td>
</tr>
<tr>
<td>main effects &amp;</td>
<td>7 30 2:3 5:26 6:11 7:36 8:20 8:22 8:25 8:31</td>
<td>1.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nonsmoke pixels</td>
</tr>
<tr>
<td>main effects, logistic</td>
<td>21.1</td>
</tr>
<tr>
<td>interactions, logistic</td>
<td>20.0</td>
</tr>
<tr>
<td>main effects, autologistic</td>
<td>17.6</td>
</tr>
<tr>
<td>interactions, autologistic</td>
<td>16.2</td>
</tr>
</tbody>
</table>
Qualitative results

- Reasonable results in many cases
- Mixed smoke + cloud is still a problem
- Data quality issues (mis-labelled training/test data)
Conclusions and future directions
Summary

• ALR is an interesting option for binary-response regression problems with complex associations.

• Thus far, different communities appear to have used different codings by default.
  – But this yields different models!
  – Plus/minus coding is best?

• The centered model has been put forth as the “new default” ALR model
  – Our work casts doubt on this choice.

• We’ve proposed a computationally-feasible analysis scheme for ALR with large sets of hyperspectral images.
Further work

- First priority: finish assessments of model variants and formalize

- ALR extensions
  - Let the pairwise parameter be $\lambda(x_i, x_j)$: adaptive smoothing.
  - Autobinomial model

- Related models
  - MRF of Beta RVs to model probabilities directly?

- Other applications
  - Ecological data
  - Network data
  - … suggestions?