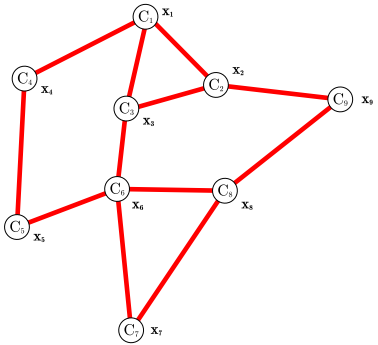
Autologistic Regression Models, With Application to Segmentation of Hyperspectral Satellite Imagery

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

November 27, 2015
Shanghai Jiao Tong University
Institute of Natural Sciences

Outline

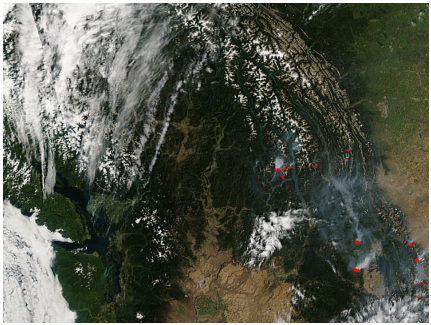
The technical content:



Autologistic regression for binary variables $\mathbf{C}$, with complex association, and covariate information $\mathbf{X}$.

- There are a few model variants.
- Claim: *it matters* which one you choose.

The application:



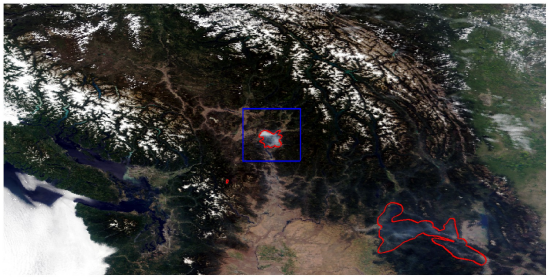
Where is the smoke in this picture?

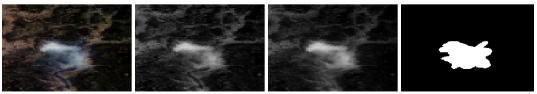
Motivating application

Remote sensing for smoke monitoring

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

Our interest: smoke from forest fires.





Original image | Green | Blue | True class

Data: MODIS images

- 1 per day, 143 days
- 1.2 Mp each
- Centered at Kelowna, BC
- Hand-drawn smoke areas

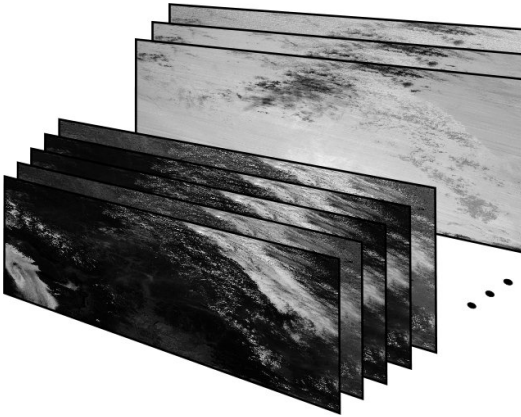
Goal: classify pixels into *smoke*/*nonsmoke*

Why?

- Health studies
- Model input or validation
- Monitoring & archiving

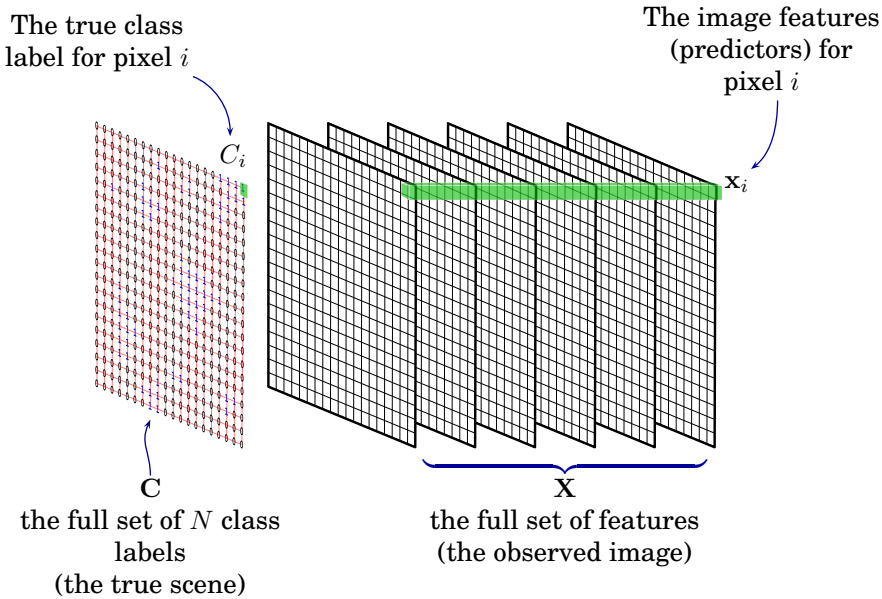
Data characteristics

- *Binary responses* (smoke/nonsmoke).
- Spectra at each pixel are *covariates* for predicting smoke.
- *Hyperspectral* images: a *high-dimensional* predictor space.
- Expect *spatial association*.



35 image planes
+
higher order terms

Notation


The true class label for pixel i
The image features (predictors) for pixel i
C_i
$\mathbf{x}_i$
$\mathbf{C}$
the full set of N class labels
(the true scene)
$\mathbf{X}$
the full set of features
(the observed image)

Autologistic regression models

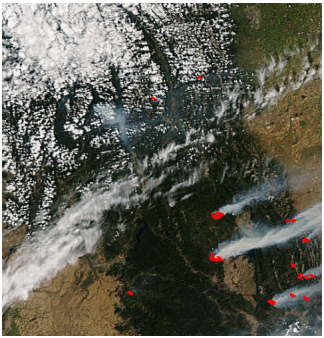
Spatial Associations

Image segmentation = pixel classification.

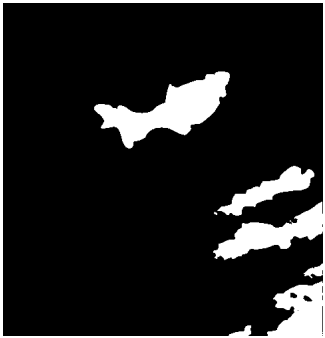
If independent pixels ⇒ Use standard classification technology.

But smoke/nonsmoke regions are *spatially smooth*.

RGB working image



The true scene



Model-Based Approaches

Many *ad hoc* ways to let pixels influence each other.

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

- graphical model
- popular in computer vision

Image model (how to do?)

MRF model for the true scene

Generative model: $p(\mathbf{C}|\mathbf{X},\theta) \propto p(\mathbf{X}|\mathbf{C},\theta)p(\mathbf{C},\theta)$

Discriminative model: $p(\mathbf{C}|\mathbf{X},\theta) = \eta(\mathbf{X},\theta)$ $\Leftarrow$ *regression*

"conditional random field"

We will use the discriminative approach.

Model $p(\mathbf{C}|\mathbf{X},\theta)$ directly as a MRF.

Hammersley–Clifford theorem

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

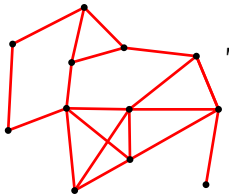
$\mathcal{M}$ = the set of maximal cliques.
$\mathbf{C}_m$ = the variables in the m^{th} clique. Then

$$p(\mathbf{c}) = \frac{1}{Z} \prod_{m \in \mathcal{M}} \phi_m(\mathbf{c}_m)$$

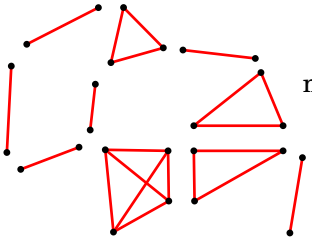
Customary to write the joint density as *Gibbs distribution* form,

$$p(\mathbf{c}) \propto e^{Q(\mathbf{c})}$$

where $Q(\cdot)$ is the *negpotential function*.



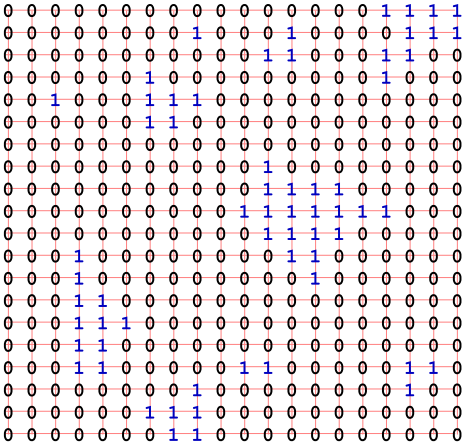
The MRF



Its maximal cliques

The Autologistic Model

- A Markov random field of **binary random variables**.
- For now, use zero/one coding: $\mathbf{C}_i \in \{0, 1\}$.
- The graph is a **regular, square grid**.
- Nothing new
  Physics: Ising model
  Statistics: Besag (1974)



The Autologistic Model (2)

Joint PMF:

$$\Pr(\mathbf{C} = \mathbf{c}|\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{\alpha}, \boldsymbol{\lambda})} \exp\left(\underbrace{\sum_{i \in \mathcal{V}} \alpha_i c_i}_{\textbf{unary terms}} + \underbrace{\sum_{(i,j) \in \mathcal{E}} \lambda_{ij} c_i c_j}_{\textbf{pairwise terms}} \right)$$

unary terms (one per vertex)

pairwise terms (one per edge)

- Positive α_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: $\alpha_i = \mathbf{x}_i^T \boldsymbol{\beta}$.

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

Then:

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

Interpretation:

- Unary part is a linear predictor.
- $\mathbf{x}_i^T \boldsymbol{\beta}$ determines conditional log-odds of $C_i = +1$ in the absence of spatial effects.
- Pairwise λ determines strength of neighbour effects.
- *Setting $\lambda = 0$ reverts to standard logistic regression*.

A centered version

Caragea & Kaiser (2009)
"Autologistic models with interpretable parameters"
Hughes, Haran, & Caragea (2011)
"Autologistic models for binary data on a lattice"

- The neighbour sum $\sum c_j$ increases log-odds unless all neighbours are zero.
  $\Longrightarrow$ Estimates of $\boldsymbol{\beta}, \lambda$ are strongly coupled

- Strongly recommended a *centered autologistic model*:

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

  where $\mu_j = \mathrm{E}[Y_j|\lambda = 0]$ 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): $L(\boldsymbol{\beta}, \lambda) \approx \prod_{\text{img}} \prod_{i=1}^{n} \text{logit}(\pi_i)$
3. Monte Carlo ML
4. Bayesian approach

} *Hughes et al. use perfect sampling; recommend PL for large n.*

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

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 $f_1(\mathbf{z}; \boldsymbol{\theta}_1)$ and $f_2(\mathbf{y}; \boldsymbol{\theta}_2)$ are *equivalent* if for any $\boldsymbol{\theta}_2$ there exists a $\boldsymbol{\theta}_1^*$ such that the two models assign the same probability whenever $\mathbf{z}$ and $\mathbf{y}$ represent the same configuration.

A general form of the model

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

$$\boldsymbol{\mu}_{\boldsymbol{\alpha}} = \begin{cases} \mathbf{0} & \text{for a standard model} \\ [\mu_1^{\alpha} \cdots \mu_n^{\alpha}]^T \text{ with } \mu_i^{\alpha} = \dfrac{\ell e^{\ell\alpha_i} + h e^{h\alpha_i}}{e^{\ell\alpha_i} + e^{h\alpha_i}} & \text{for a centered model} \end{cases}$$

Then the *autologistic (AL)* PMF is

$$f_{\mathbf{C}}(\mathbf{c}; \boldsymbol{\alpha}, \lambda) \propto \exp(\mathbf{c}^T\boldsymbol{\alpha} - \lambda \mathbf{c}^T \mathbf{A}\boldsymbol{\mu}_{\boldsymbol{\alpha}} + \frac{\lambda}{2}\mathbf{c}^T\mathbf{A}\mathbf{c}),$$

where $\mathbf{A}$ 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 $\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\beta}$

Autologistic models are equivalent

Theorem 1

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

Sketch of proof:

- If $\mathbf{z}$ and $\mathbf{y}$ have different coding, they are related by

$$\mathbf{z} = a\mathbf{y} + b\mathbf{1}$$

  whenever they represent the same configuration.
- Use this to equate PMFs of $\mathbf{z}$ and $\mathbf{y}$
- 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 $\boldsymbol{\beta}$.
- Coefficients of the system depend on arbitrary $\mathbf{X}$.
- 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 $\boldsymbol{\mu_\alpha}$ term:

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

   Sign of pairwise term depends on majority vote.

2. It "decouples" $\boldsymbol{\beta}$ and λ better than centering (evidence to follow).

2. It allows a convenient *plug-in estimation* of λ ...

An estimation shortcut for λ

Proposal: plug-in estimation

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

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

Rationale

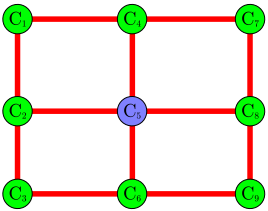
Treat λ as a smoothing parameter.

- Assuming independence, $\hat{\boldsymbol{\beta}}$ captures how information in $\mathbf{X}$ can be used to predict $\mathbf{C}$.
- For fixed $\hat{\boldsymbol{\beta}}$, tuning λ 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.

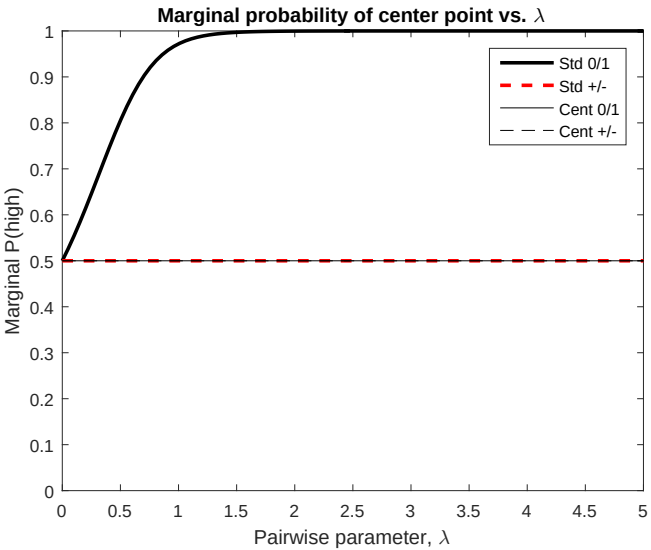
Linear predictor vector is

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

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

Small example: case 1

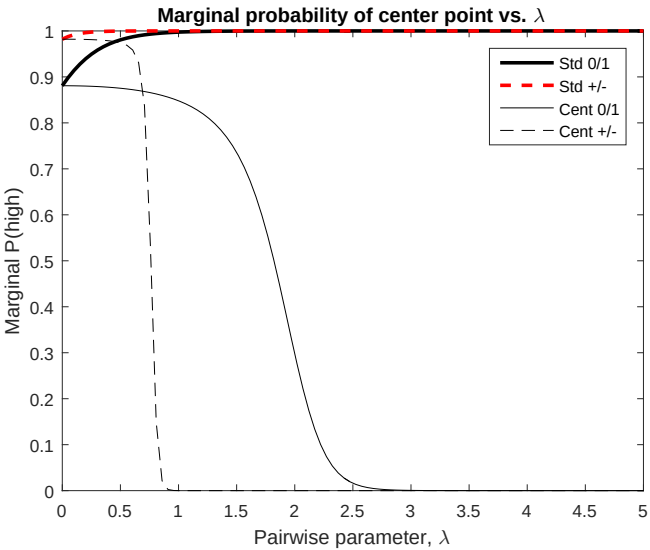
$$\boldsymbol{\beta} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \qquad \text{Linear predictors:} \quad \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}$$



Shows the asymmetry of the standard 0/1 model.

Small example: case 2

$\boldsymbol{\beta} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $x_o = 1$, $x_5 = 1$. Linear predictors: $\begin{matrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{matrix}$

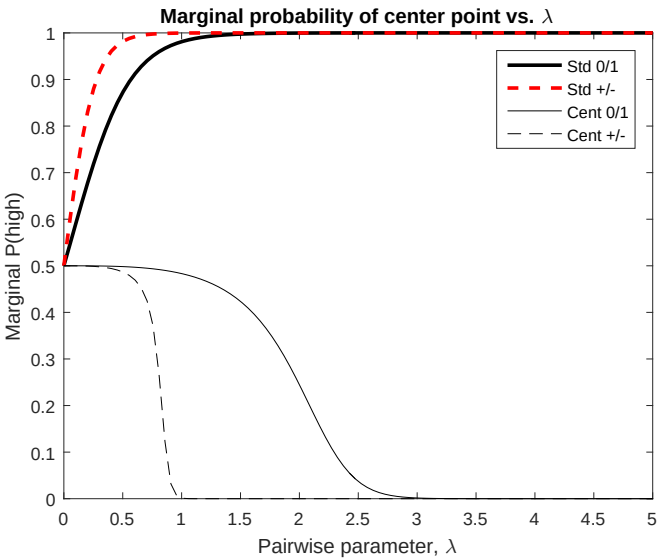


Standard and centered models give opposite behaviour as $\lambda \to \infty$.

Contradicts the idea that centered models have "interpretable parameters."

Small example: case 3

$\boldsymbol{\beta} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $x_o = 1$, $x_5 = -1$. Linear predictors: $\begin{matrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{matrix}$

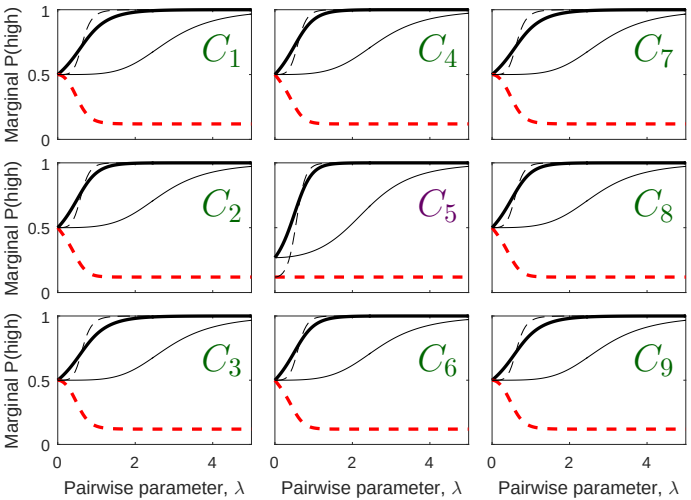


Again, standard and centered models disagree.

Does the centered model behaviour make sense?

Small example: case 4

$\boldsymbol{\beta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $x_o = 0$, $x_5 = -1$. Linear predictors: $\begin{matrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{matrix}$



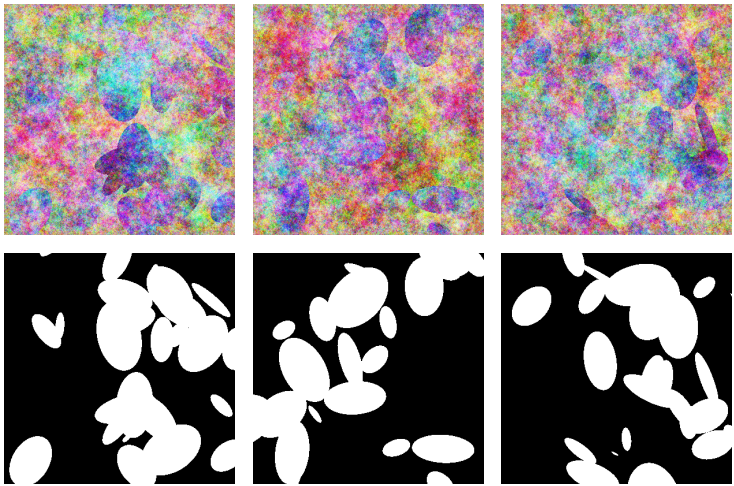
Standard plus/minus very different from the other three.

Limiting probability not equal to 1 or 0.

C_5 probability constant wrt λ.

Simulated images

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.

Simulated images–results 1

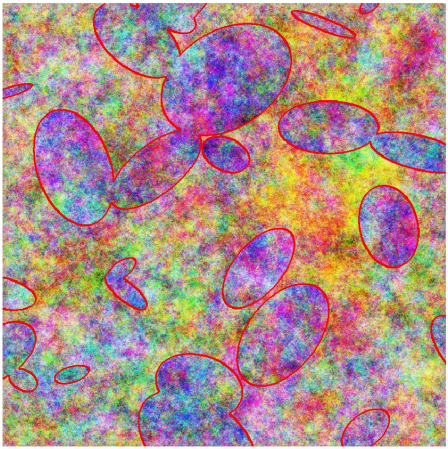
Parameter estimates and prediction error: plug-in vs. pseudolikelihood (plus-minus coding).

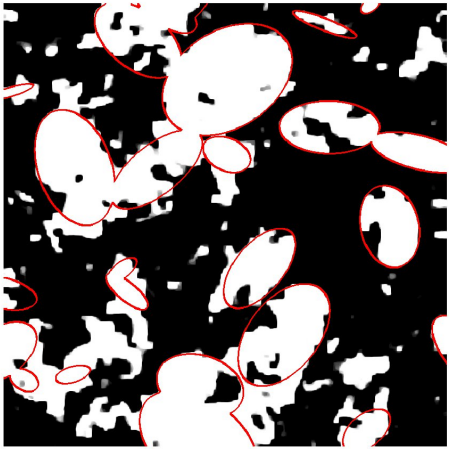
pixels	method	$\hat{R}$	$\hat{G}$	$\hat{B}$	$\hat{\lambda}$	error rate (%)
100^2	plug-in	−2.21	−2.02	1.91	0.90	20.1
	PL	−2.04	−1.99	2.06	0.99	20.4
200^2	plug-in	−1.64	−1.35	1.71	1.00	17.7
	PL	−1.61	−1.30	1.70	1.19	17.7
400^2	plug-in	−2.05	−1.42	1.63	1.60	20.1
	PL	−2.08	−1.40	1.68	1.36	20.1
600^2	plug-in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

- Parameter estimates similar.
- Error rates similar.

Simulated images–results 2

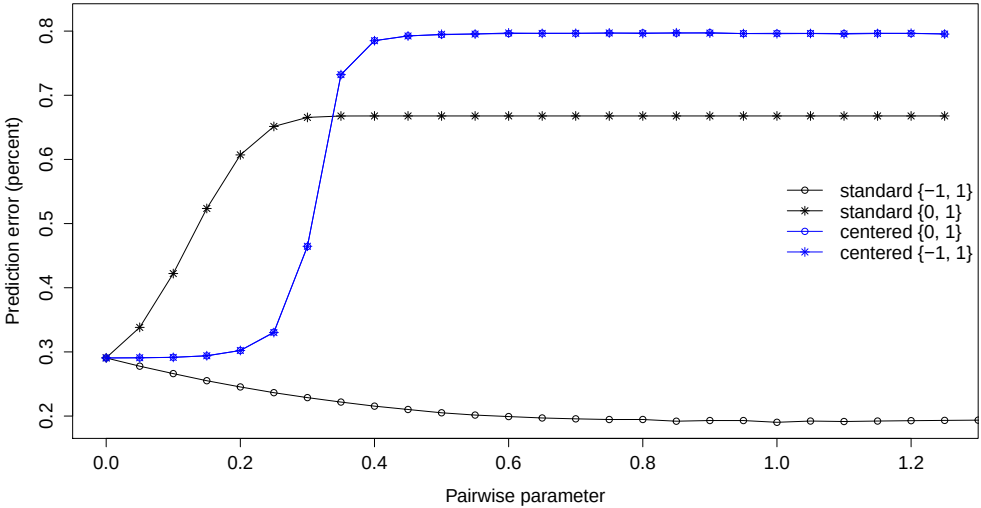
Example predictions, 800×800 image.





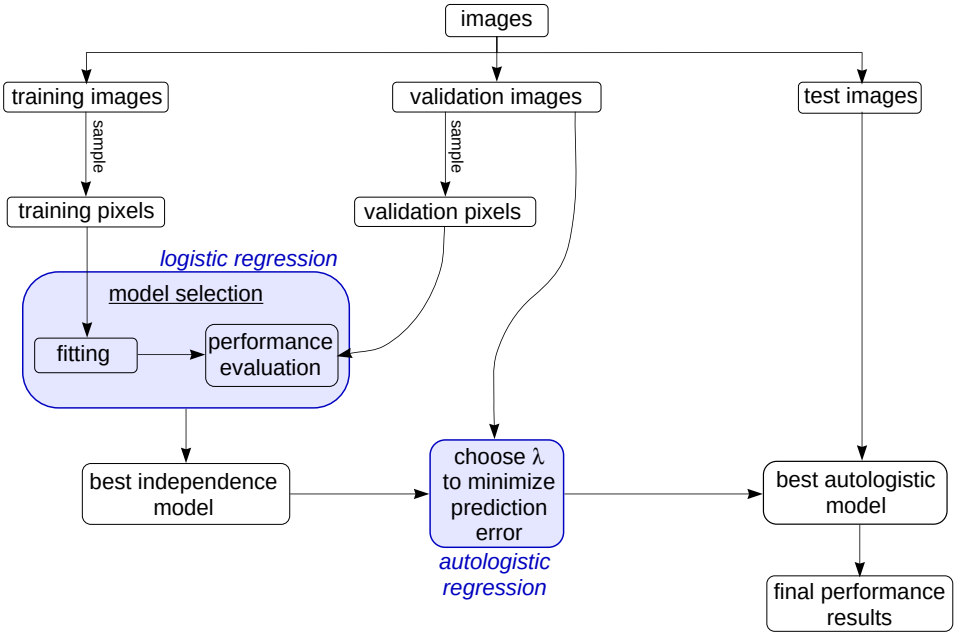
Simulated images–results 3

What about plugin with different model variants?



Analysis of the smoke data

Analysis flowchart



Results 1

Use a logistic GAM

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

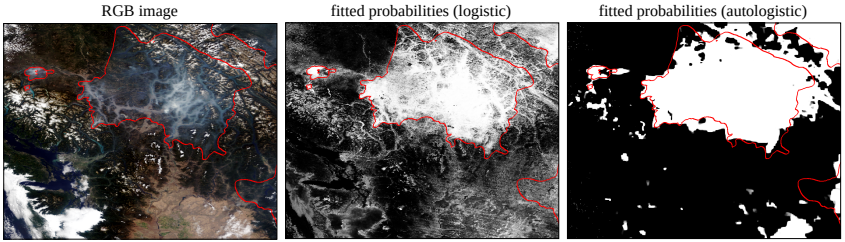
Predictor set	Selected variables (MODIS band numbers)	plug-in $\hat{\lambda}$
main effects	1 6 7 8 14 16 17 18 21 23 25 26 30 31 32 36	1.85
main effects & interactions	7 30 2:3 5:26 6:11 7:36 8:20 8:22 8:25 8:31 13:15 13:23 16:31 18:23 22:36 32:36	1.75

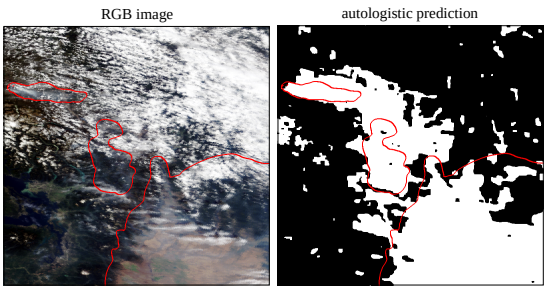
Model	Error rate (%)		
	nonsmoke pixels	smoke pixels	overall
main effects, logistic	21.1	25.9	21.6
interactions, logistic	20.0	23.3	20.3
main effects, autologistic	17.6	23.9	18.2
interactions, autologistic	16.2	21.3	16.7

Results 2

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 $\lambda(\mathbf{x}_i, \mathbf{x}_j)$: adaptive smoothing.
  - Auto*binomial* model
- Related models
  - MRF of Beta RVs to model probabilities directly?
- Other applications
  - Ecological data
  - Network data
  - . . . suggestions?