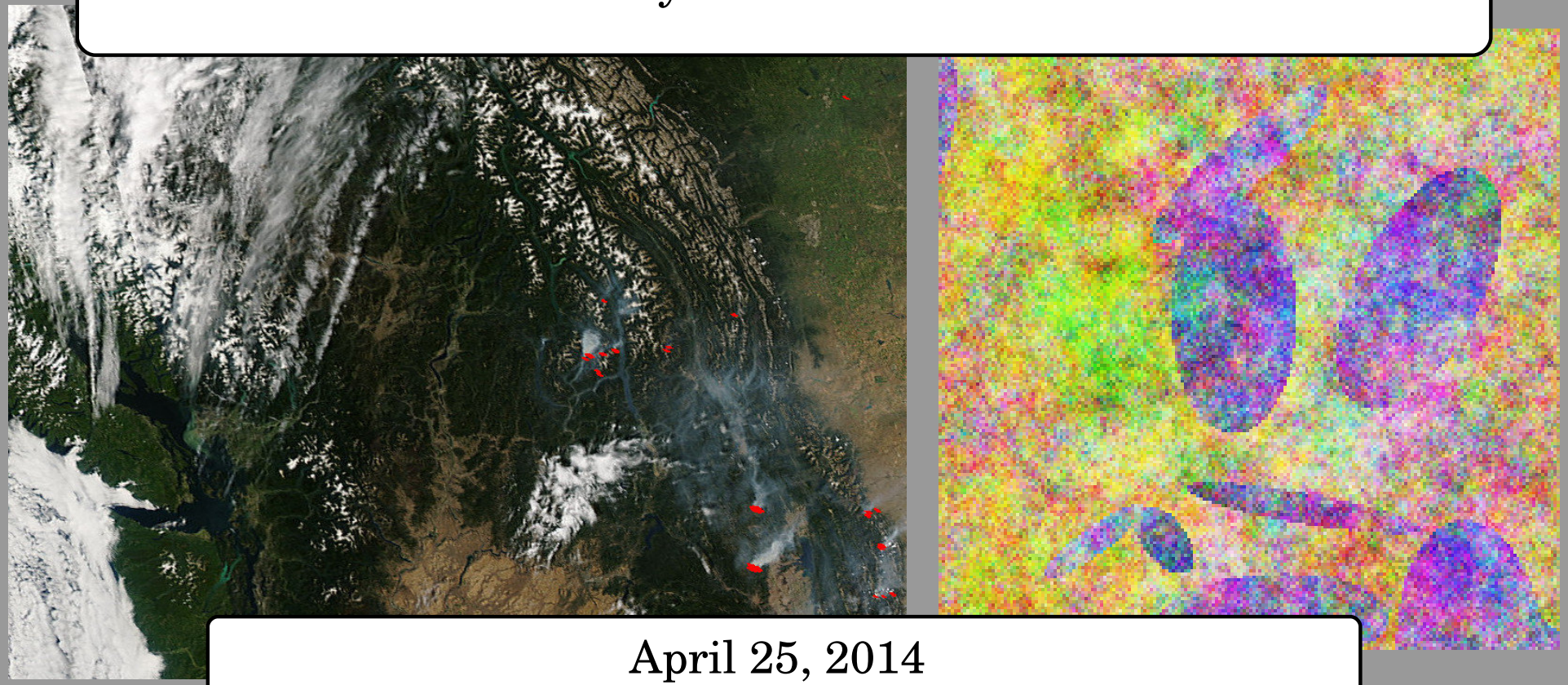
Detection of Smoke in Satellite Images Using Autologistic Regression

Mark Wolters
University of Western Ontario

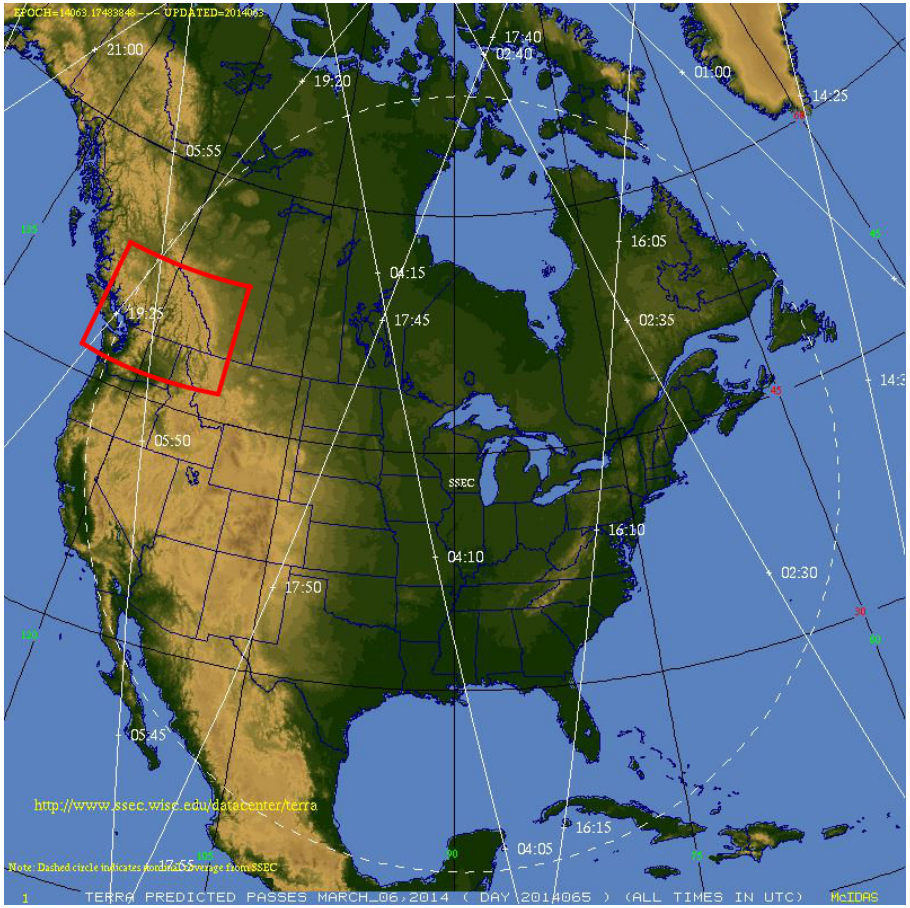


April 25, 2014
Department of Statistics and Finance
University of Science and Technology of China

The Data & the Application

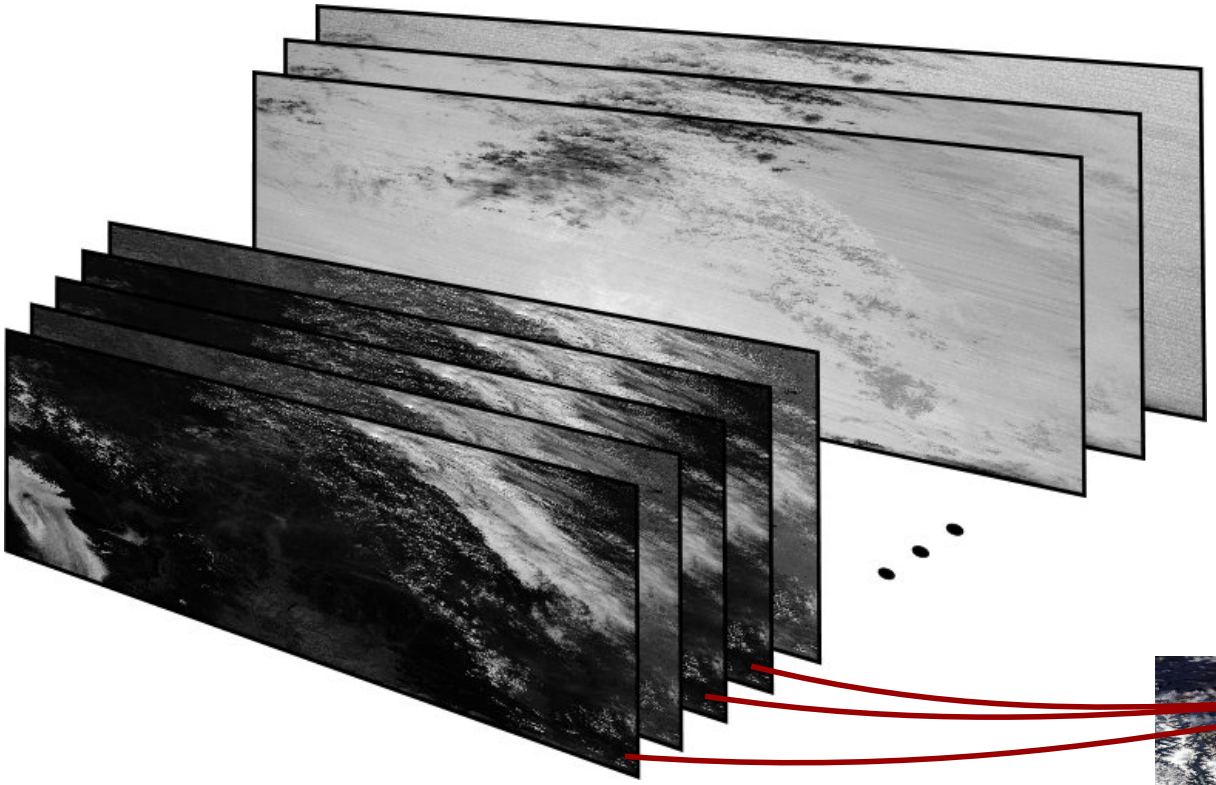
Remote Sensing Data

- Earth-Orbiting satellites Terra & Aqua
- Instrument: MODIS (moderate resolution imaging spectroradiometer)
- Big data (dimensions, volume, throughput)
- ROI: region centered on Kelowna, B.C., Canada



(http://www.ssec.wisc.edu/datacenter/terra/)

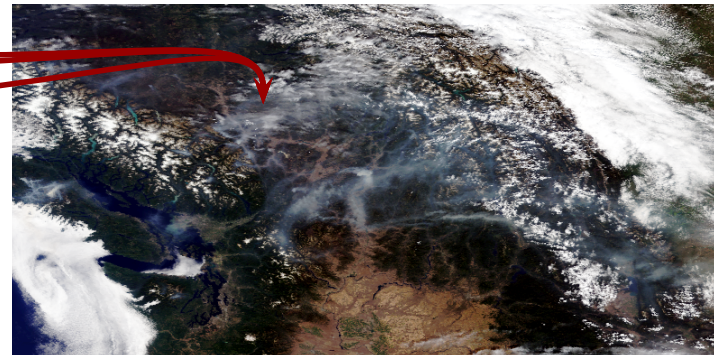
Hyperspectral Images



One image, 36 image planes

Band 36, 14μm

Band 1, 0.65μm



Bands 1, 3, 4 make RGB image for visualization

Application: Smoke Monitoring

- Smoke from forest fires has **population health relevance**
- **Large area**, hard to monitor.
- First-principles modelling efforts
  (wildfire model + weather model + dispersion model)
- Satellite data:
  - copious, freely available
  - wide geographic coverage
  - no altitude information
- Applications:
  - Retrospective studies
  - Model validation
  - Model initialization, updating

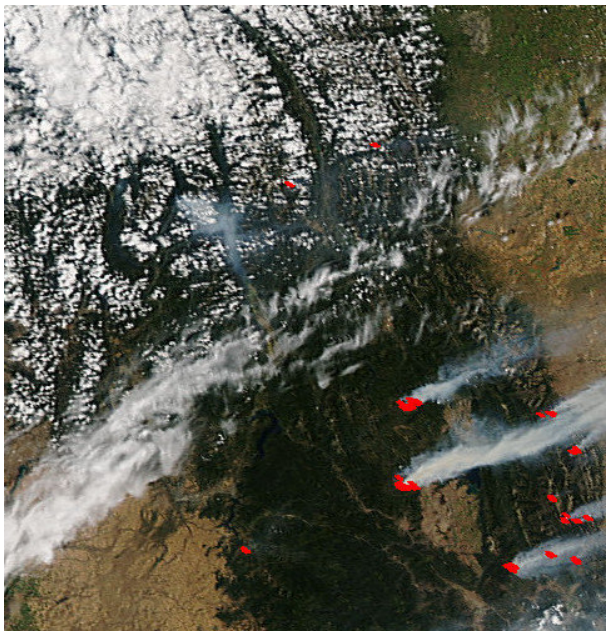
The Goal

Develop an automatic system for identifying smoke in MODIS imagery.

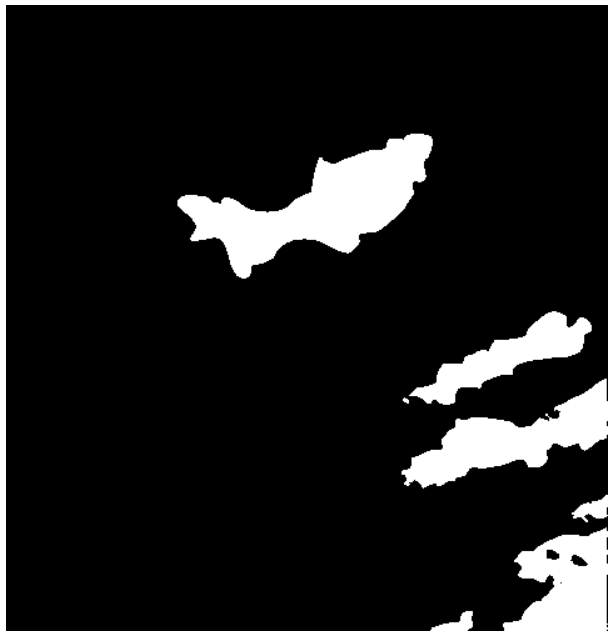
$\Rightarrow$ **binary image segmentation**

- Classes: smoke, nonsmoke
- Supervised learning, hand-drawn training images
- Methodology applicable to other applications

RGB working image



The true scene



Notation

N	Number of pixels in the image.
d	The number of spectra collected at each pixel ($d = 3$ for RGB image, $d = 36$ for hyperspectral).
C_i	The true unknown class label of pixel i. Let c_i take values $\{-1, +1\}$.
$\mathbf{x}_i$	Image features (predictors) for pixel i.
$\mathbf{C}$	The full set of labels $\{C_1, \ldots, C_N\}$. The complete true scene.
$\mathcal{X}$	The full set of features $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. The complete information from the observed image.

Incorporating “Context”

Spatial Associations

- If independent pixels $\Rightarrow$ 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:** $p(\mathbf{C}|\mathcal{X},\theta) \propto p(\mathcal{X}|\mathbf{C},\theta)p(\mathbf{C},\theta)$
- **Discriminative model:** $p(\mathbf{C}|\mathcal{X},\theta) = \eta(\mathcal{X},\theta)$ $\Leftarrow$ *regression*

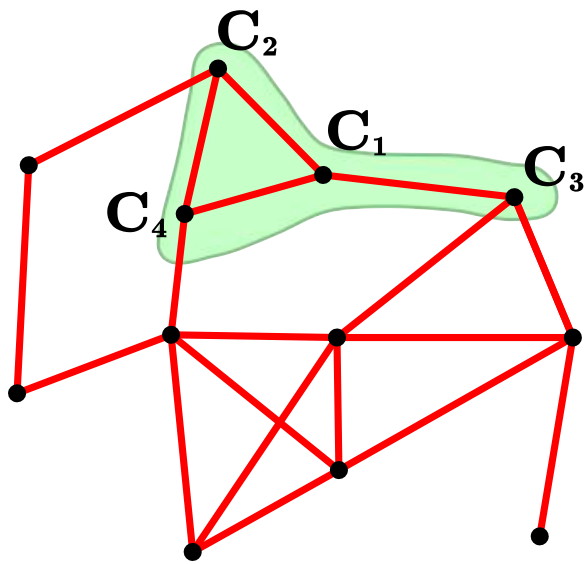
We will use the discriminative approach.

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

Aside: Markov Random Fields (1/3)

- A collection of random variables, $\mathbf{C} = [C_1, C_2, \ldots, C_n]^T$.
- Undirected graph structure.
- The graph lets us see the *Markov blanket* of any C.
- In the figure at right:

$$\Pr(C_1|C_2, \ldots, C_{12}) = \Pr(C_1|C_2, C_3, C_4)$$



- Could build a model by specifying 12 such relationships.
- But: are they consistent?

Aside Markov Random Fields (2/3)

Hammersley–Clifford theorem

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

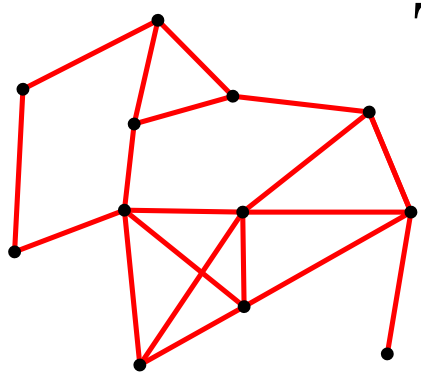
$\mathcal{M}$ = the set of maximal cliques.
$\mathbf{C}_m$ = the variables in the m^{th} clique.
Then

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

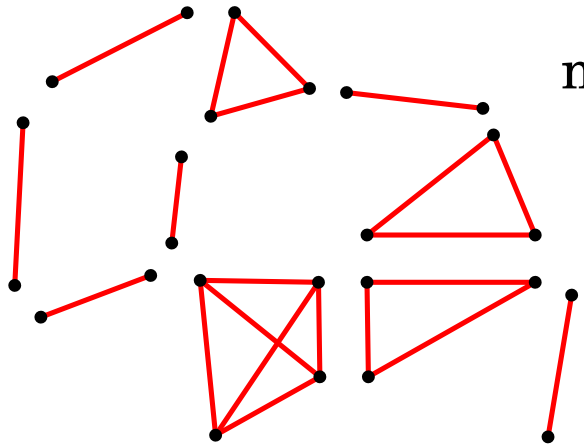
normalizing constant

m^{th} potential function

The MRF



Its maximal cliques



Aside: Markov Random Fields (3/3)

Notes:

Potential functions must be strictly positive.
So write the joint density in *Gibbs distribution* form, $p(\cdot) \propto e^{-U(\cdot)}$

$$
\begin{aligned}
p(\mathbf{c}) &= \frac{1}{Z} \prod_{m \in \mathcal{M}} \phi(\mathbf{c}_m) \\
&= \frac{1}{Z} \prod_{m \in \mathcal{M}} e^{-\psi(\mathbf{c}_m)} \\
&= \frac{1}{Z} \exp\left(- \sum_{m \in \mathcal{M}} \psi(\mathbf{c}_m) \right) \\
&= \frac{1}{Z} e^{-U(\mathbf{c})}
\end{aligned}
$$

- $U(\mathbf{c})$ is called the *energy function*.
- the $\psi(\mathbf{c}_m)$ are called *clique energies*, *clique potentials*, or *potential functions*.
- Values of $\mathbf{c}$ that give *higher* energy give *lower* probability.

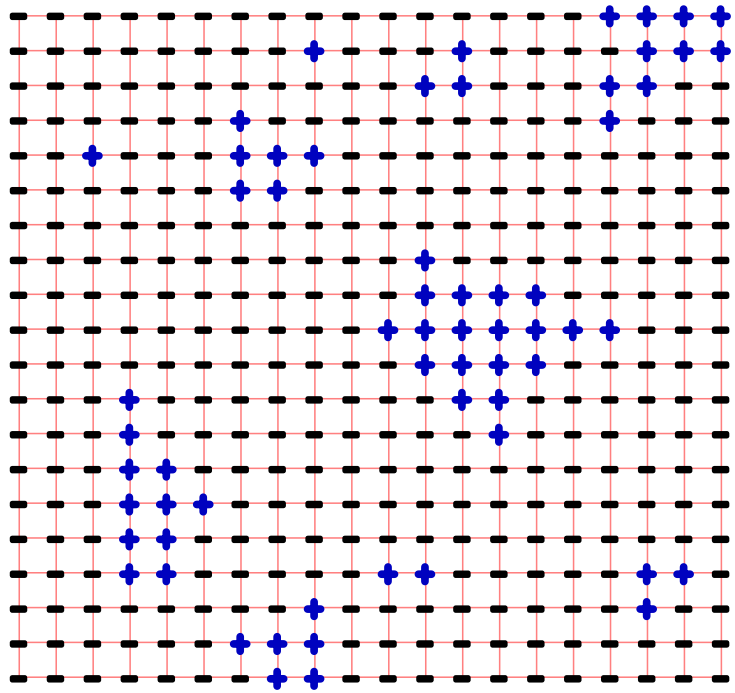
The Autologistic Regression Model

The Autologistic Model

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

(Besag, 1974)

- A MRF of **binary** random variables.
- Use plus/minus coding: $\mathbf{C}_i \in \{-1, +1\}$.
- **Pairwise** energy function
- The graph is a **regular grid**.



Parameter Interpretation

The energy function is

$$U(\mathbf{c}) = -\sum_{i \in \mathcal{V}} \alpha_i c_i - \sum_{(i,j) \in \mathcal{E}} \beta_{ij} c_i c_j.$$

So at location i:

- Positive α_i values favor $+1$ (smoke) class.
- Setting $\beta_{ij} > 0$ favors *locally smooth* configurations ($C_i = C_j$).

What about conditional distributions?

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

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 2\alpha_i + 2\sum_{j \sim i} \beta_{ij} c_j$$

$j \sim i$: j is a neighbour of i

Parameter Interpretation (continued)

If α_i, β_{ij} known, full conditionals π_i have simple form.

So can use *Gibbs sampling* to draw from $p(\mathbf{C})$.

Note: usually set $\beta_{ij} = \beta$ for all (i, j).

Open Gibbs sampler videos

Autologistic *Regression*

Unary coefficients depend on covariates: $\alpha_i = \frac{1}{2}\mathbf{x}_i^T\boldsymbol{\omega}$.

Pairwise coefficients constant, as before: $\beta_{ij} = \frac{\lambda}{2}$.

Then the conditional logits become

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

Interpretation:

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

Parameter Estimation

Challenge:

- Normalizing constant Z is intractable: need 2^N terms.
- likelihood-based inference hard.

Solutions:

- Use the independence model to estimate $\boldsymbol{\omega}$ and plug in. Choose λ directly based on prediction error.
- *Pseudolikelihood*: just maximize $\prod \pi_i$.
- Various sampling-based approximations have been proposed.

Prediction/Classification

Challenge:

- For fixed parameters $\hat{\omega}, \hat{\lambda}$ we seek a "best" configuration $\mathbf{c}$ for a given new image.
- Problematic because N too large.

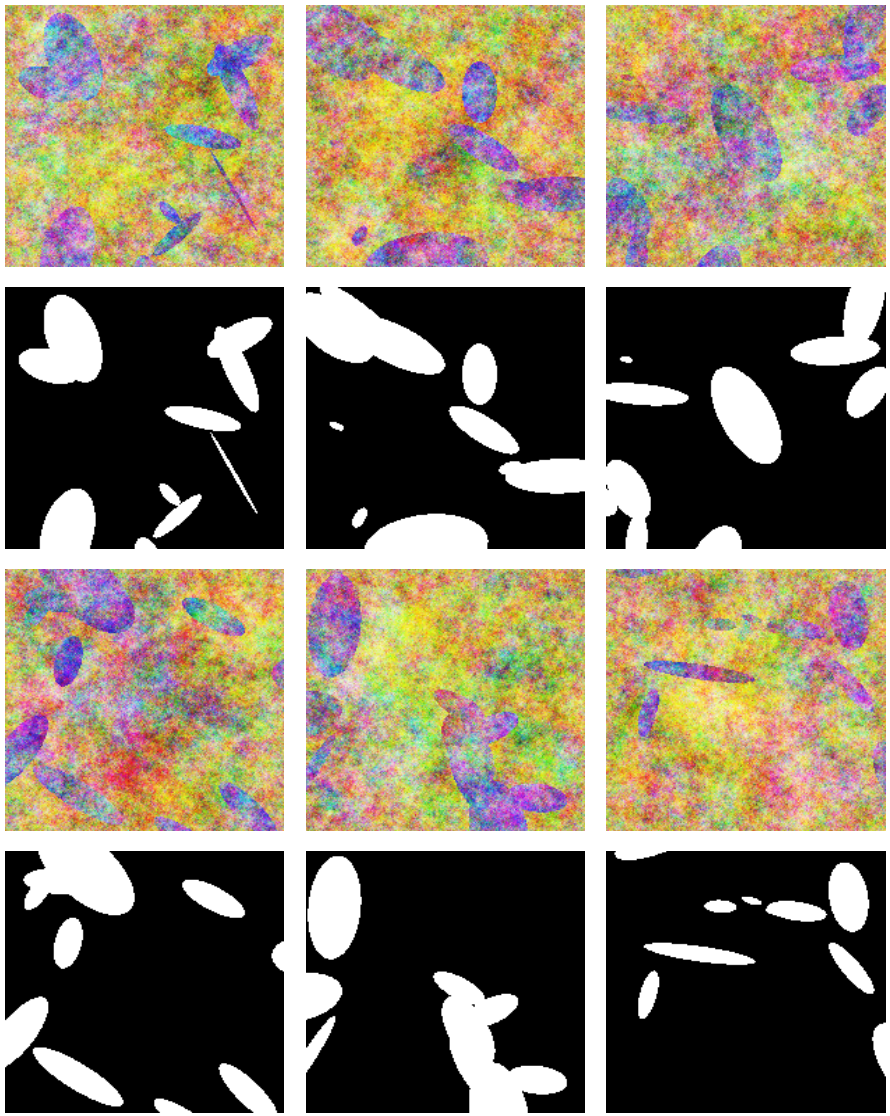
Solutions:

- Maximum A Posteriori (MAP): find the $\mathbf{c}^*$ that maximizes $p(\mathbf{C}|\mathcal{X})$
  - Attractive because don't need to know Z.
  - Hard in general: search 2^N possibilities
  - Global methods exist for some cases, using *graph cut* methods.
- Marginal probabilities:
  - Approximate marginal $\Pr(C_i = +1)$. Use Gibbs sampler.
  - Then, assign C_i to smoke class if $\Pr(C_i = +1) > 0.5$.

Model Performance

Simulated Data

- RGB images, 200×200
- Random ellipses
- Phase 1 (smoke) inside the ellipses, phase 2 (nonsmoke) outside.
- One GMRF per color per phase
- 20 training images
- 20 test images
- 20 validation images



Models Compared

- Consider five variants of the model.
- How to measure performance?

$$\text{error rate} = \begin{pmatrix} \text{proportion nonsmoke} \\ \text{pixels misclassified} \end{pmatrix} + \begin{pmatrix} \text{proportion smoke} \\ \text{pixels misclassified} \end{pmatrix}$$

1) IPLM Independent Pixel Logistic Model

2) IPLM+ Independence model with neighour information

3) AL-1a Autologistic with estimation shortcut

4) AL-1b Standard autologistic model

5) AL-2 Autologistic with adaptive pairwise coefficient

Model: IPLM

1) IPLM Independent Pixel Logistic Model

- Standard logistic classifier, with feature selection

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T \boldsymbol{\omega}$$

- $\mathbf{x}_i$ consists of features selected from among all interactions of R, G, B up to 3rd order.
- This produces 41 candidate variables:

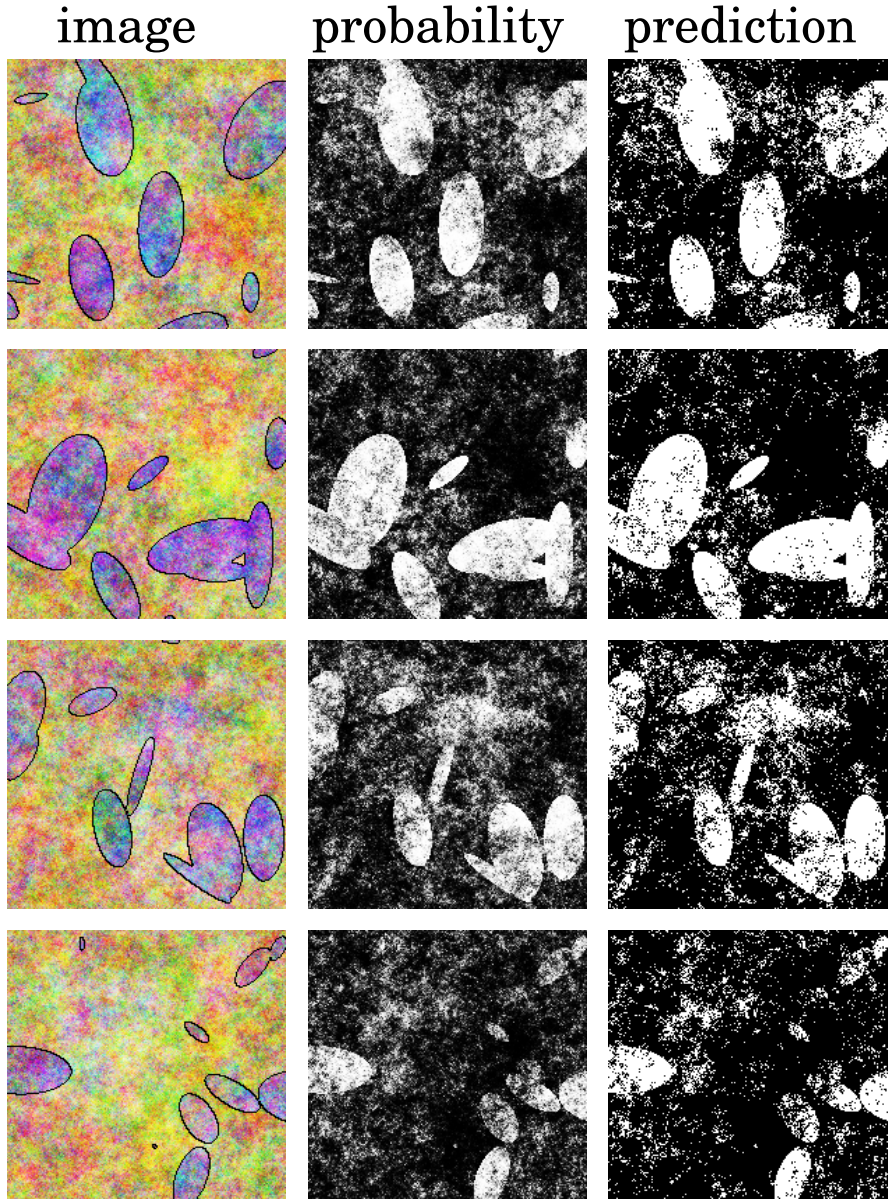
$$R, G, B, R^2, G^2, B^2,$$
$$RG, RB, RR^2, RG^2, RB^2, GB, GR^2, GG^2, GB^2, BR^2, BG^2, BB^2, R^2G^2, R^2B^2, G^2B^2$$
$$RGB, RGR^2, RGG^2, RGB^2, RBR^2, RBG^2, RBB^2, RR^2G^2, RR^2B^2, RG^2B^2$$
$$GBR^2, GBG^2, GBB^2, GR^2G^2, GR^2B^2, GG^2B^2$$
$$BR^2G^2, BR^2B^2, BG^2B^2, R^2B^2G^2.$$

- Used GA subset selection.
- Simple model (R, G, B) was best.

Sample Output—IPLM

- Error rate: 0.301
- Unary Coefficient estimates:

icept	R	G	B
4.31	-8.16	-5.53	8.29



Model: IPLM+

2) IPLM+ Independence model with neighour information

- Use predictors (R, G, B) as in IPLM.
- Add extra predictors for N, E, S, W neighbours:
  $Rn, Gn, Bn, Re, Ge, Be, Rs, Gs, Bs, Rw, Gw, Bw$

Sample Output—IPLM+

- Error rate: 0.212
- Unary Coefficient estimates:

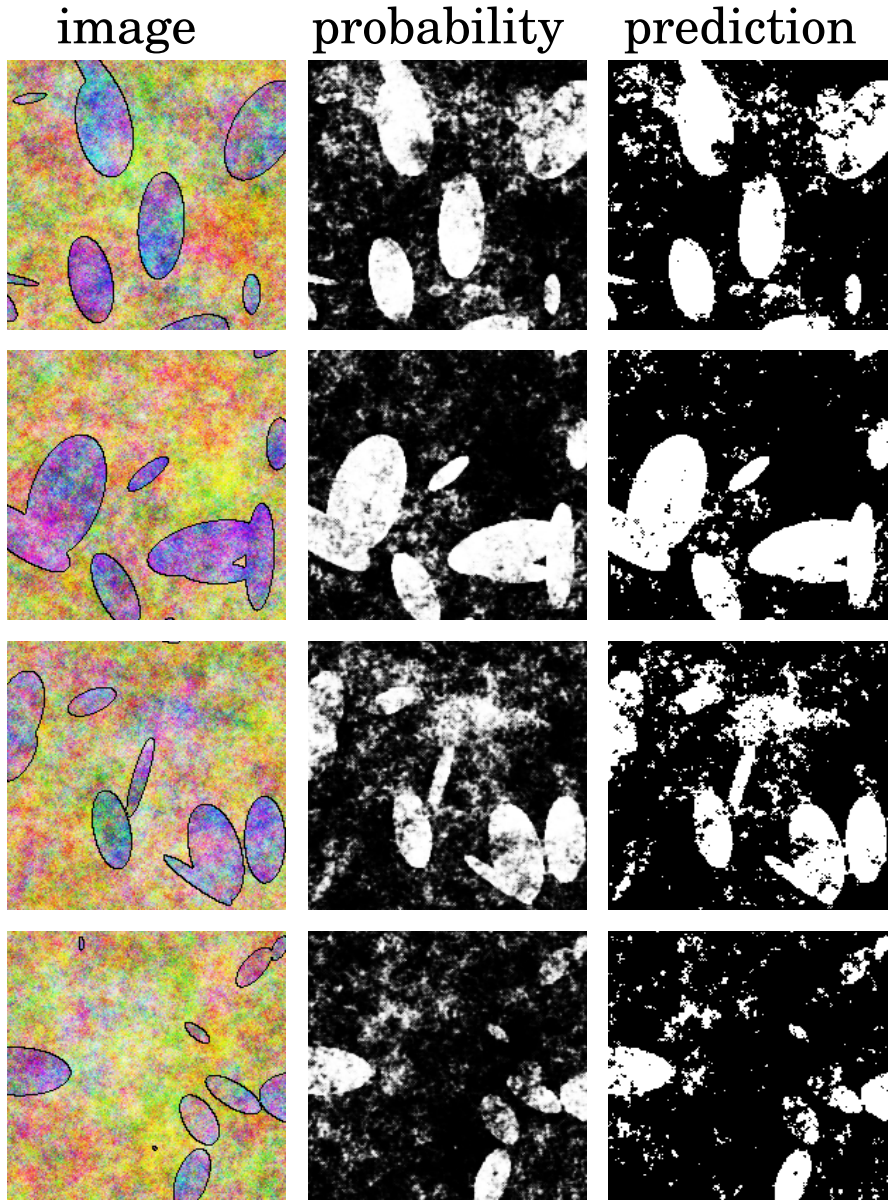
icept	R	G	B
7.28	-3.31	-2.10	3.08

Rn	Gn	Bn
-2.54	-1.66	2.05

Re	Ge	Be
-2.41	-1.59	2.43

Rs	Gs	Bs
-2.05	-1.55	2.35

Rw	Gw	Bw
-2.47	-1.90	2.39



Model: AL-1a

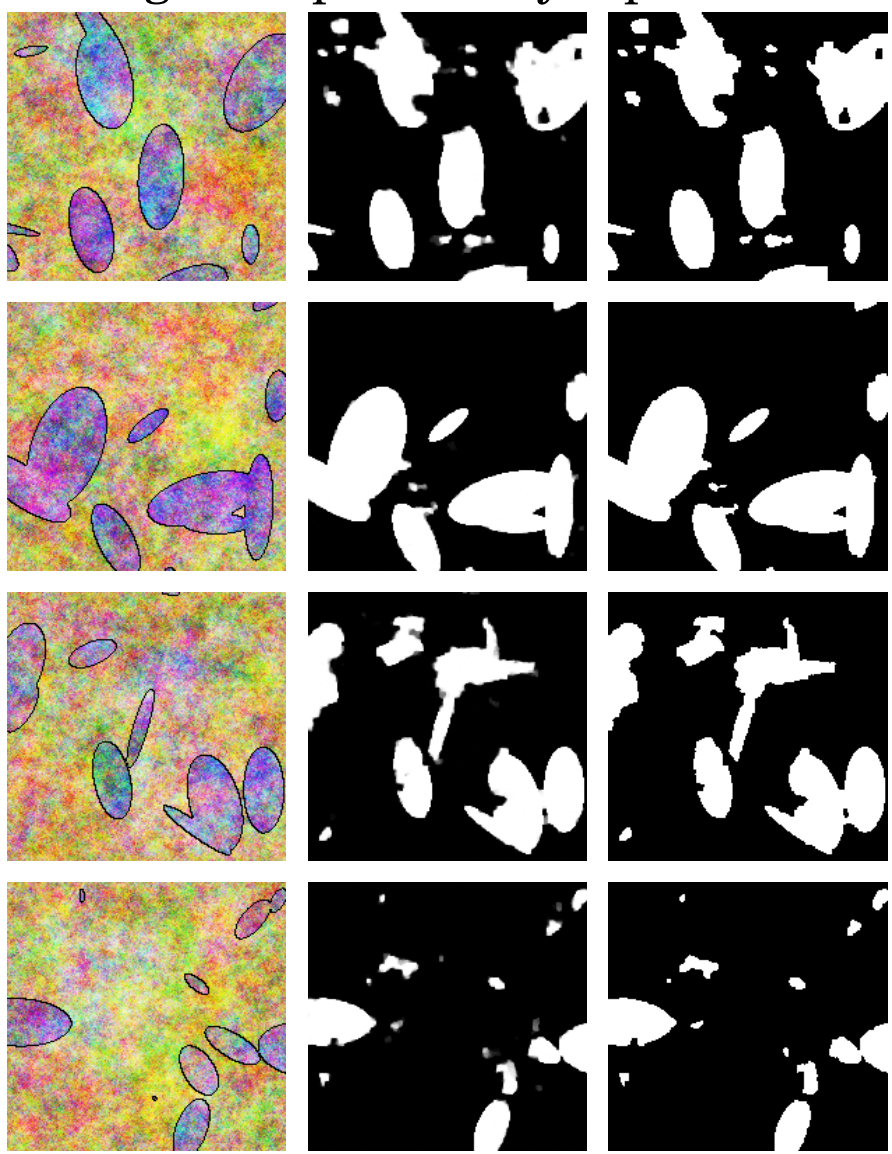
3) AL-1a Autologistic with estimation shortcut

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

- Borrow $\hat{\boldsymbol{\omega}}$ from IPLM
- Choose $\hat{\lambda}$ to minimize classification error.

Sample Output—AL-1a

- Error rate: 0.137
- Unary Coefficient estimates: *same as IPLM*.
- Pairwise coefficient estimate: $\hat{\lambda} = 2.3$.



Model: AL-1b

4) AL-1b Standard autologistic model

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

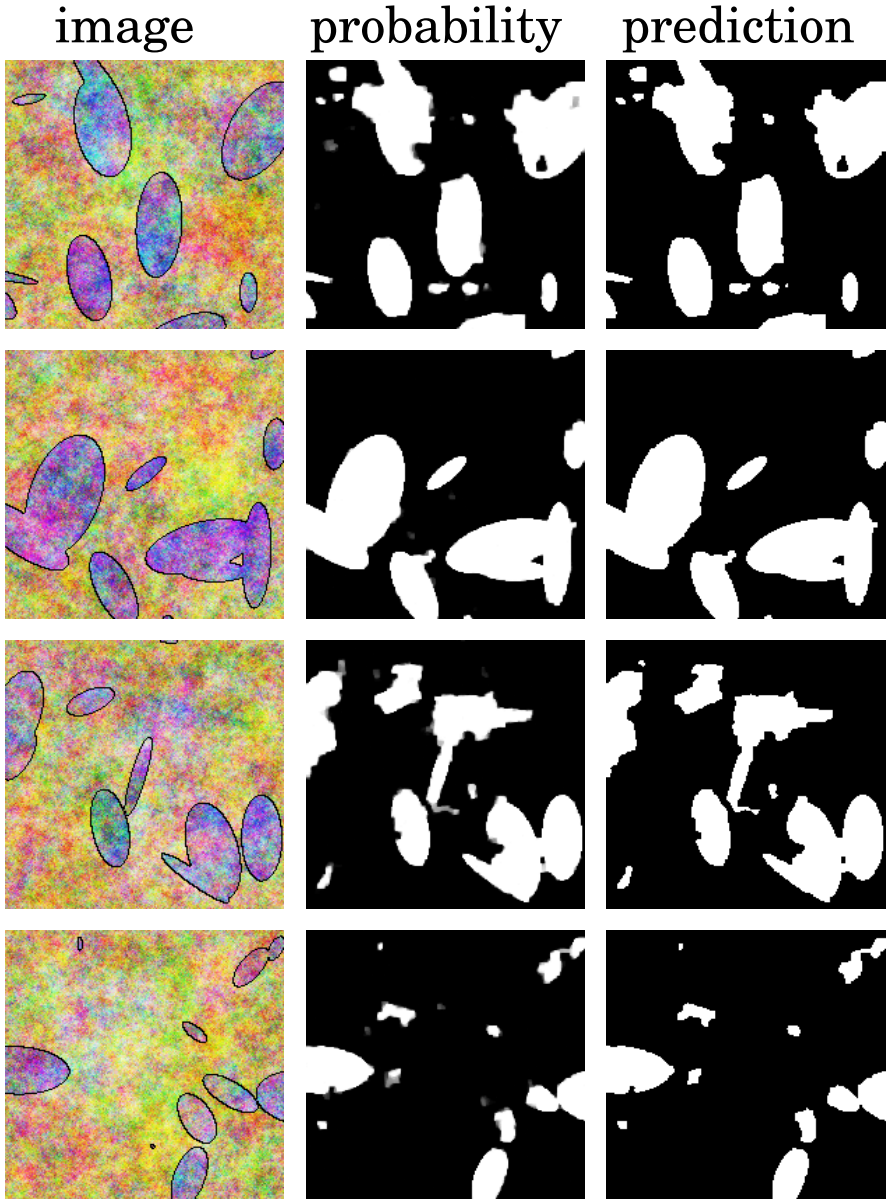
– $\hat{\omega}$, $\hat{\lambda}$ simultaneously estimated by pseudolikelihood

Sample Output—AL-1b

- Error rate: 0.128
- Unary Coefficient estimates:

icept	R	G	B
4.15	-7.77	-5.81	8.45

- Pairwise coefficient estimate: $\hat{\lambda} = 2.68$.



Model: AL-2

5) AL-2 Autologistic with adaptive pairwise coefficient

- Let pairwise β_{ij} depend on $\mathbf{x}_i$, $\mathbf{x}_j$.
- Smooth more when neigbour pixels are "similar".

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\omega} + \sum_{j\sim i} \underbrace{\boldsymbol{\phi}(d_{ij})^T\boldsymbol{\gamma}}c_j.$$

a piecewise linear function with coefficients γ

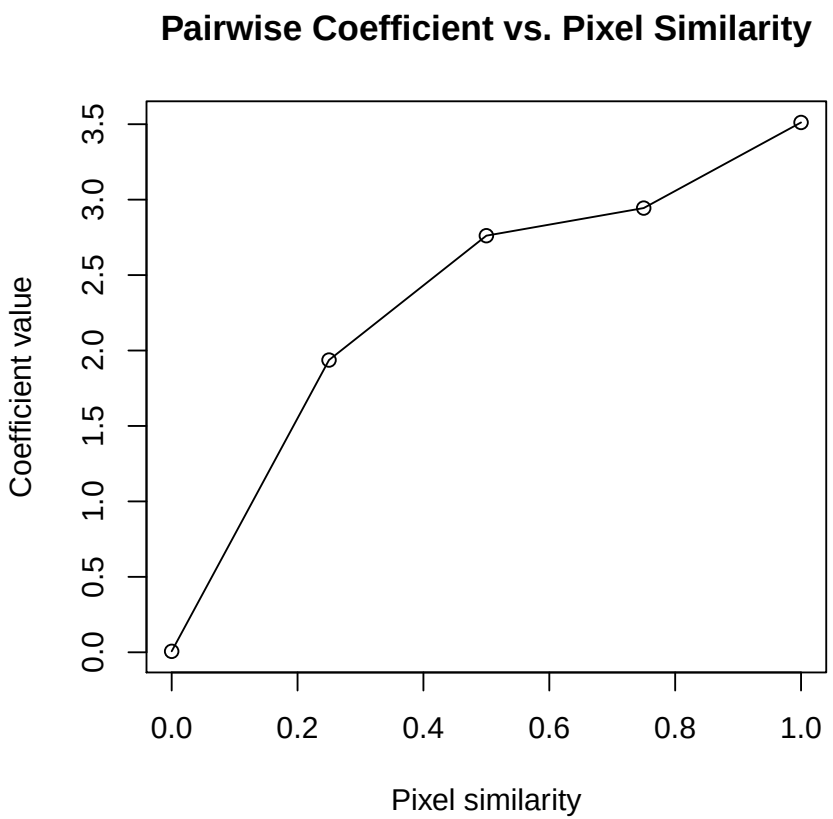
- Let $d_{ij} = 1 - |\hat{\pi}_i^I - \hat{\pi}_j^I|$,

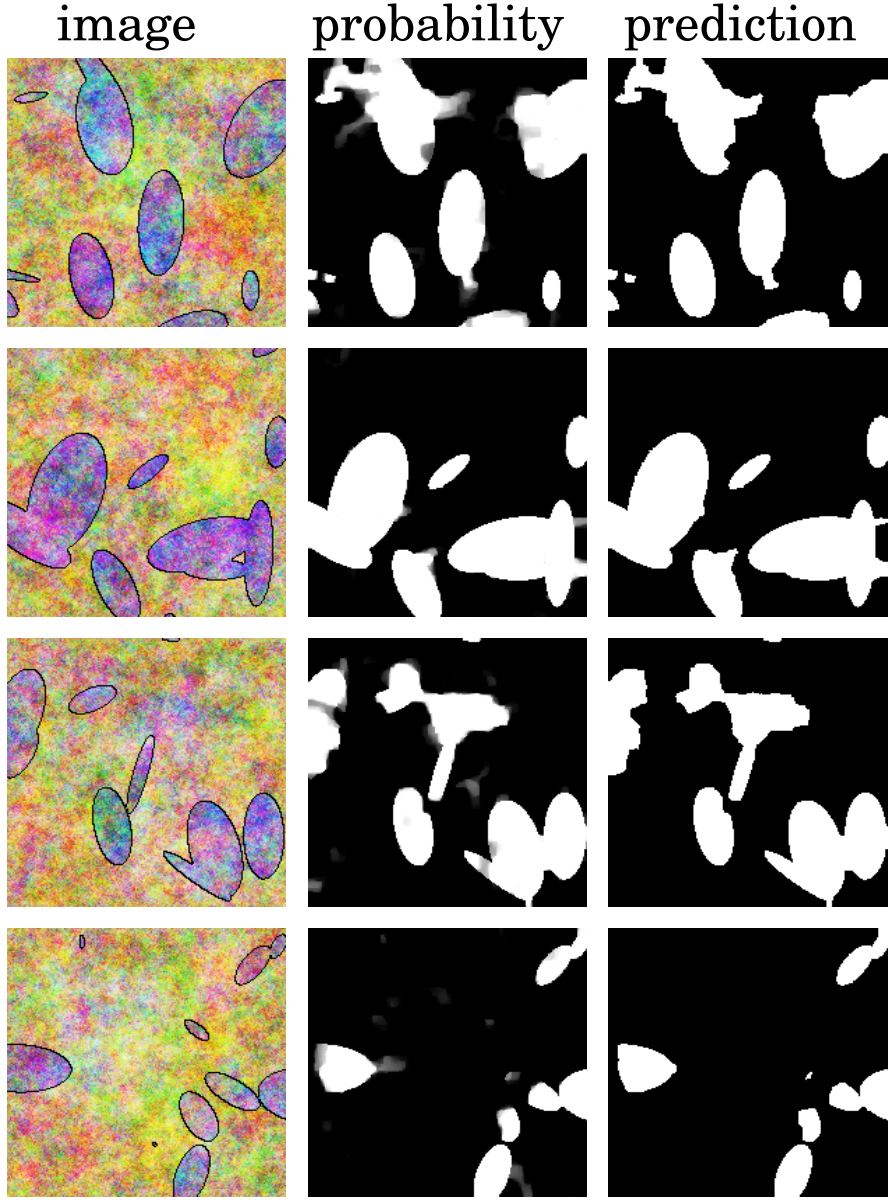
  where $\hat{\pi}_i^I$ is the IPLM fitted probability.
  $\Rightarrow$ If IPLM assigns i and j same probability, $d_{ij} = 1$
  $\Rightarrow$ As IPLM predictions diverge, $d_{ij} \to 0$

Sample Output—AL-2

- Error rate: 0.123
- Unary Coefficient estimates:

icept	R	G	B
0.826	-1.57	-1.64	2.28





Summary and Next Steps

Conclusions

- A promising approach to binary image segmentation:
  - Connection with logistic regression.
  - Different estimation approaches.
  - Simpler than generative models.
- Prediction based on marginal probability estimates (Gibbs sampling) seems to work well.
- Modelling the pairwise coefficient as a function of similarity opens potential for adaptive smoothing.
- This approach could provide hands-off classification of MODIS data as they arrive.

Future work

- Verify performance on the real data!
- Extend to *autobinomial* model (multiple classes).
- Spatial statistics applications (medicine, ecology):
  - Accurate estimation of the unary parameters is most important.
  - Plus/minus coding and adaptive smoothing both have potential to improve parameter estimation.