Science-Driven Models for Image Analysis in Astronomy and Solar Physics

David A. van Dyk

Statistics Section
Imperial College London

Chinese University Hong Kong, April 2015

Imperial College London

Introduction

Massive new data streams are opening up a world of opportunities for data scientists!

1. Astrostatistics: Data quality and quantity lead to more interesting statistical models
2. Data-driven versus Science-driven methods
3. Predictive models versus Descriptive models
4. Tradeoff: computational speed and statistical principles
5. These issues are not unique to astronomy!

Joint work with:

- CHASC International Center for Astrostatistics
  (Includes researchers at Harvard, Univ of California, NASA, Imperial, Crete, etc.)
- Imperial Centre for Inference in Cosmology

Imperial College London

Outline

1. Statistical Learning in Astronomy
2. Example I: *Mapping Thermal Structure in Solar Images*
3. Example II: *Testing Unexpected Features in X-ray Images*

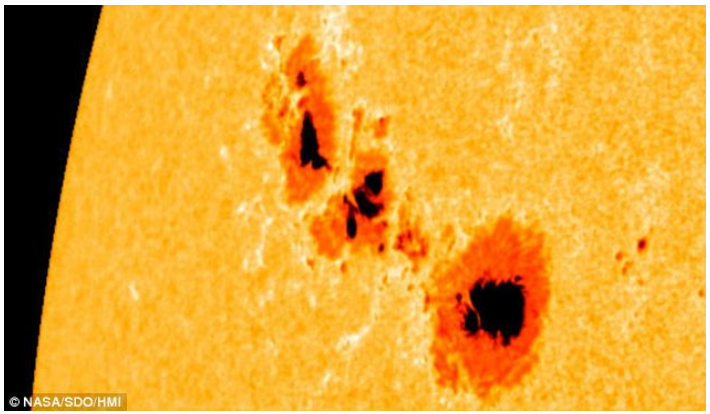
Massive Data Sets and Data Streams

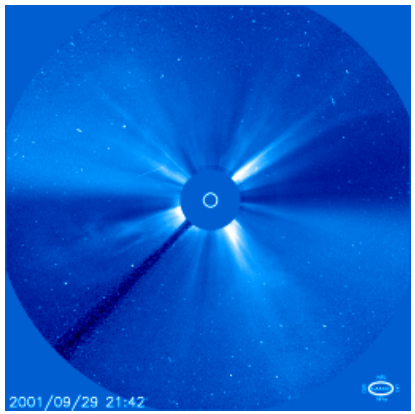
Dramatic increase in the quality and quantity of data:

- massive new surveys: catalogs containing T/PBs of data,
- high resolution spectrography and imaging across the electromagnetic spectrum,
- incredibly detailed movies of dynamic and explosive processes in the solar atmosphere,
- massive number of items and/or features,
- space-based telescopes tailored to specific scientific goals,
- data are *not just massive:* they are rich, deep, & complex.

Massive Challenges for Data Scientists!!

Example I: Thermal Structure in the Solar Corona[1]



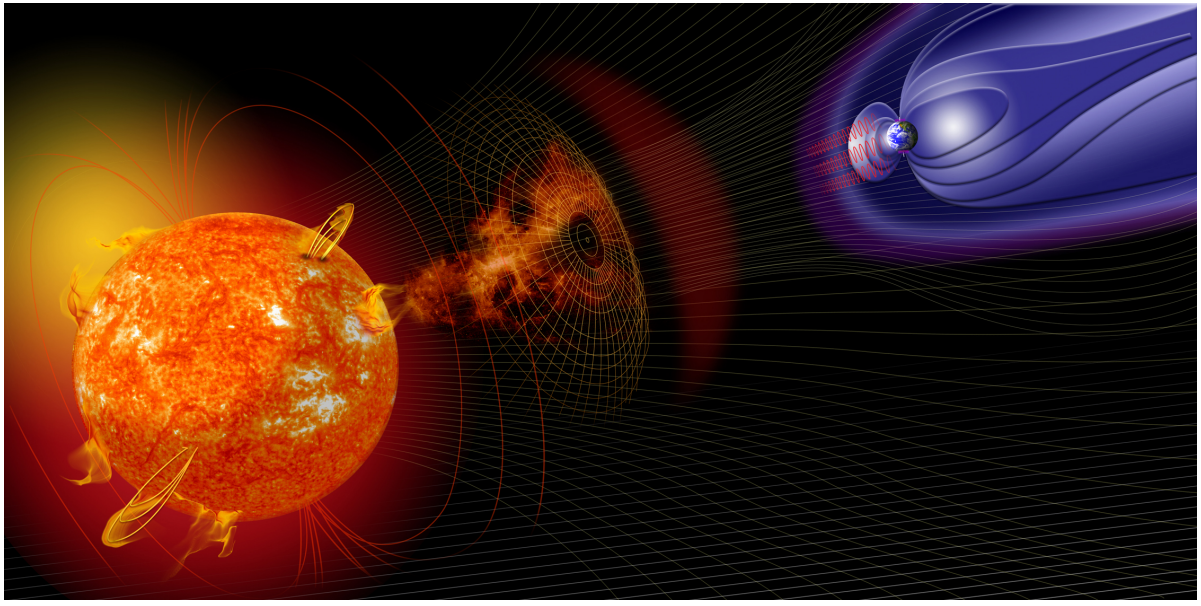


- *Solar Corona*: Highly energetic and violent, characterized by sunspots, solar flares, and coronal mass ejections.
- Solar storms can affect space weather, earth satellites, communication systems, and electric grids.
- *Goal:* Track solar activity with the aim of predicting storms and their effects on Earth.

Imperial College London

[1] N Stein, D Stenning, V Kashyap, T Lee, XL Meng, , and CHASC

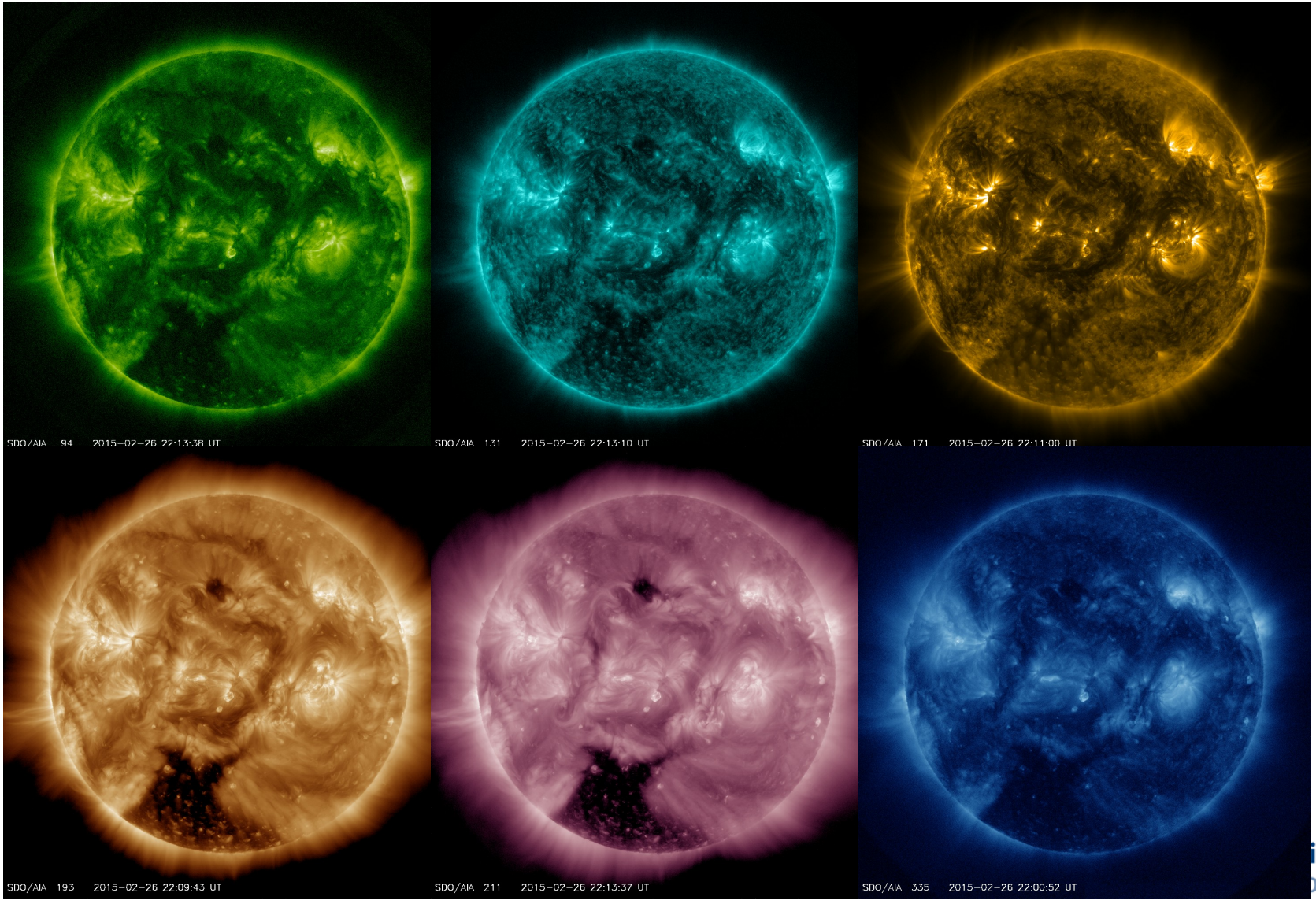
Space Weather Effects



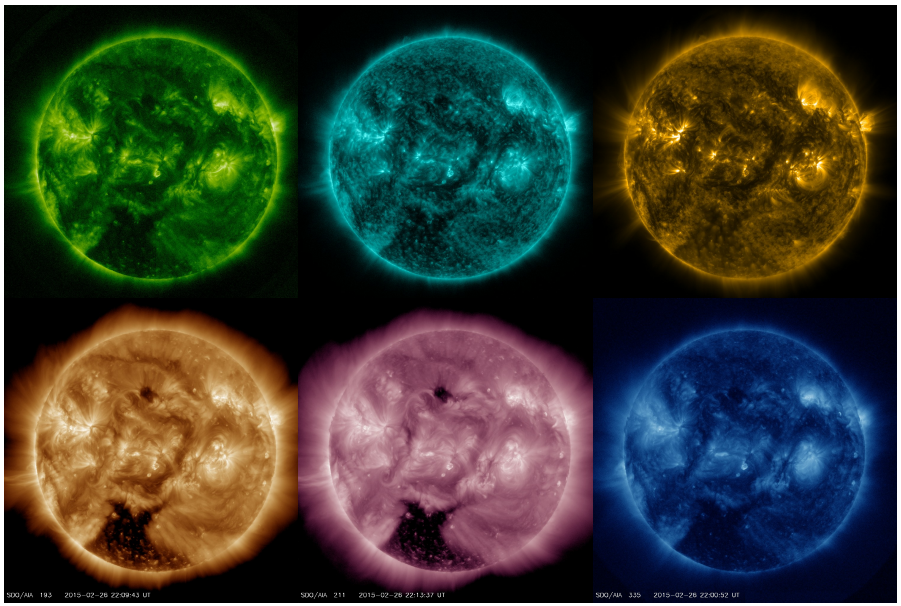
Artist illustration of events on the sun changing the conditions in Near-Earth space.
Image Credit: NASA

- The highly energetic particles released by solar flares and CMEs can impact the Earth's magnetosphere.
- These impacts cause radio interference and can damage satellites and electric power transmission.
- 1859 geomagnetic storm: Worldwide Aurorae, gold miners thought it was morning, telegraph machines failed, shocked operators, threw sparks, even if unplugged.

The Data: Pixel-by-Pixel Spectra



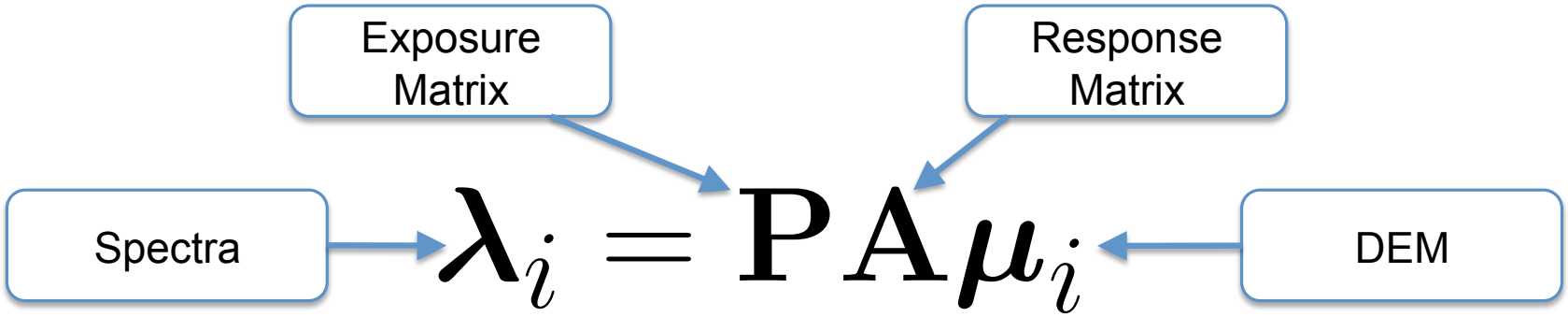
The Data: Pixel-by-Pixel Spectra



The Solar Dynamics Observatory

- NASA satellite launched February 2010
- Massive Data Stream: 1.4 TB/day of compressed data
- High Spatial and Temporal resolution
- Low Spectral resolution
- White light and magnetigram images

The Differential Emission Measure



DEM: *expected emission due to plasma of a given temperature.*
A: *expected spectra of plasmas at each temperature.*
Challenge: *Inversion is ill-posed.*

Normalized Spectra: $\pi_i = \dfrac{\mathbf{PA}\boldsymbol{\mu}_i}{\mathbf{1}^\top \mathbf{PA}\boldsymbol{\mu}_i}$, MLE: $\hat{\pi}_i = \dfrac{\boldsymbol{y}_i}{\mathbf{1}^\top \boldsymbol{y}_i}$.

Goal: *Cluster pixels with similar spectra.*

How Should we Cluster Probability Vectors?

K-means Algorithm:

Assignment: Assign units to clusters by minimizing the Euclidean distance, d_2, to the centroid.

Update: Compute new centroids by minimizing the total Euclidean distance within each cluster.

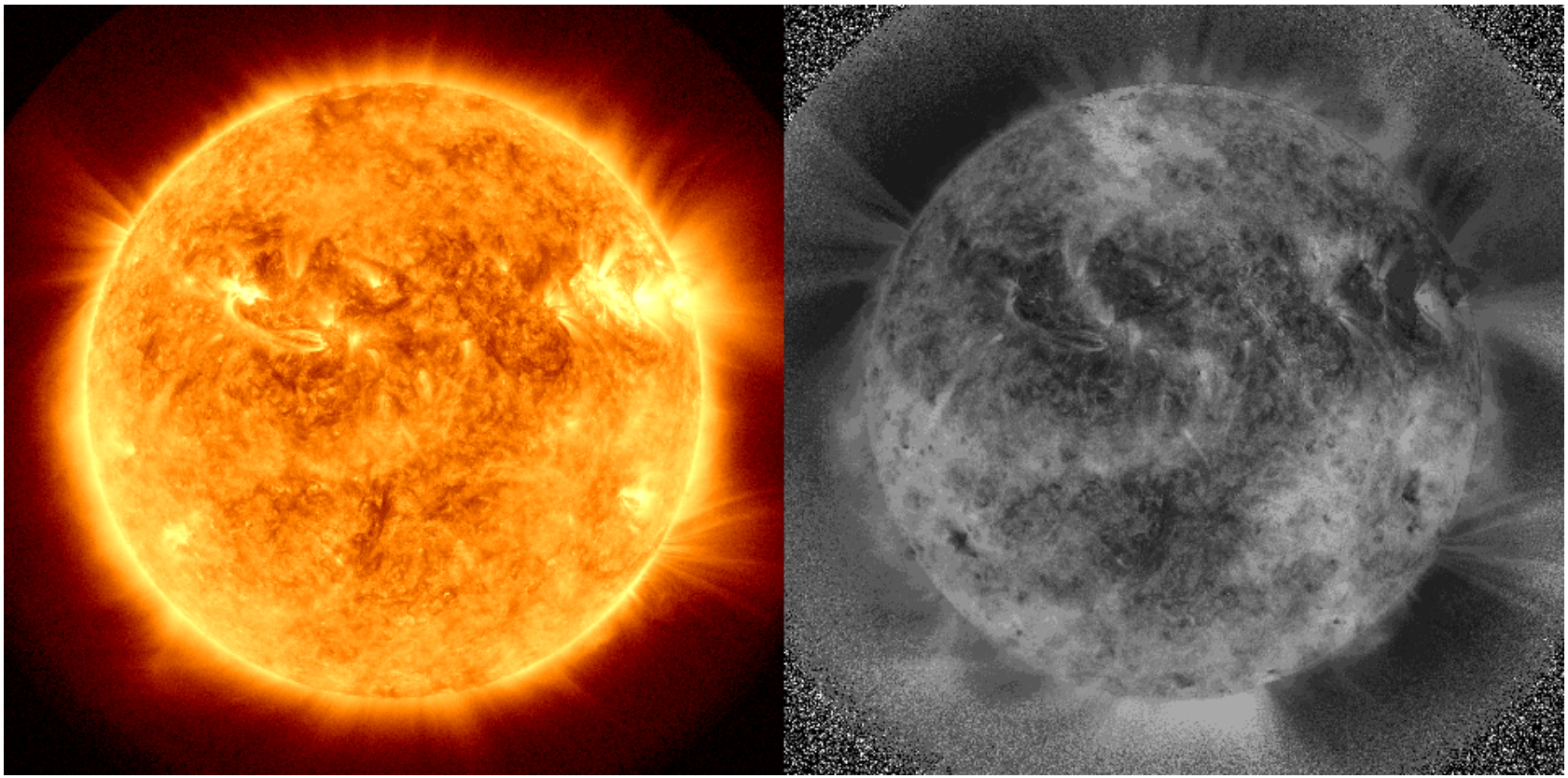
The H-means Algorithm:

- Replace Euclidean distance with **Hellinger Distance**: appropriate for probability vectors.
- Hellinger distance between $\hat{\pi}_i$ and $\boldsymbol{c}_j$:

$$d_H^2(\hat{\pi}_i, \boldsymbol{c}_j) = \frac{1}{2}\sum_k \left(\sqrt{\hat{\pi}_{ik}} - \sqrt{c_{jk}}\right)^2 = 1 - \sum_k \sqrt{\hat{\pi}_{ik} c_{jk}}.$$

- Both steps remain in closed form.

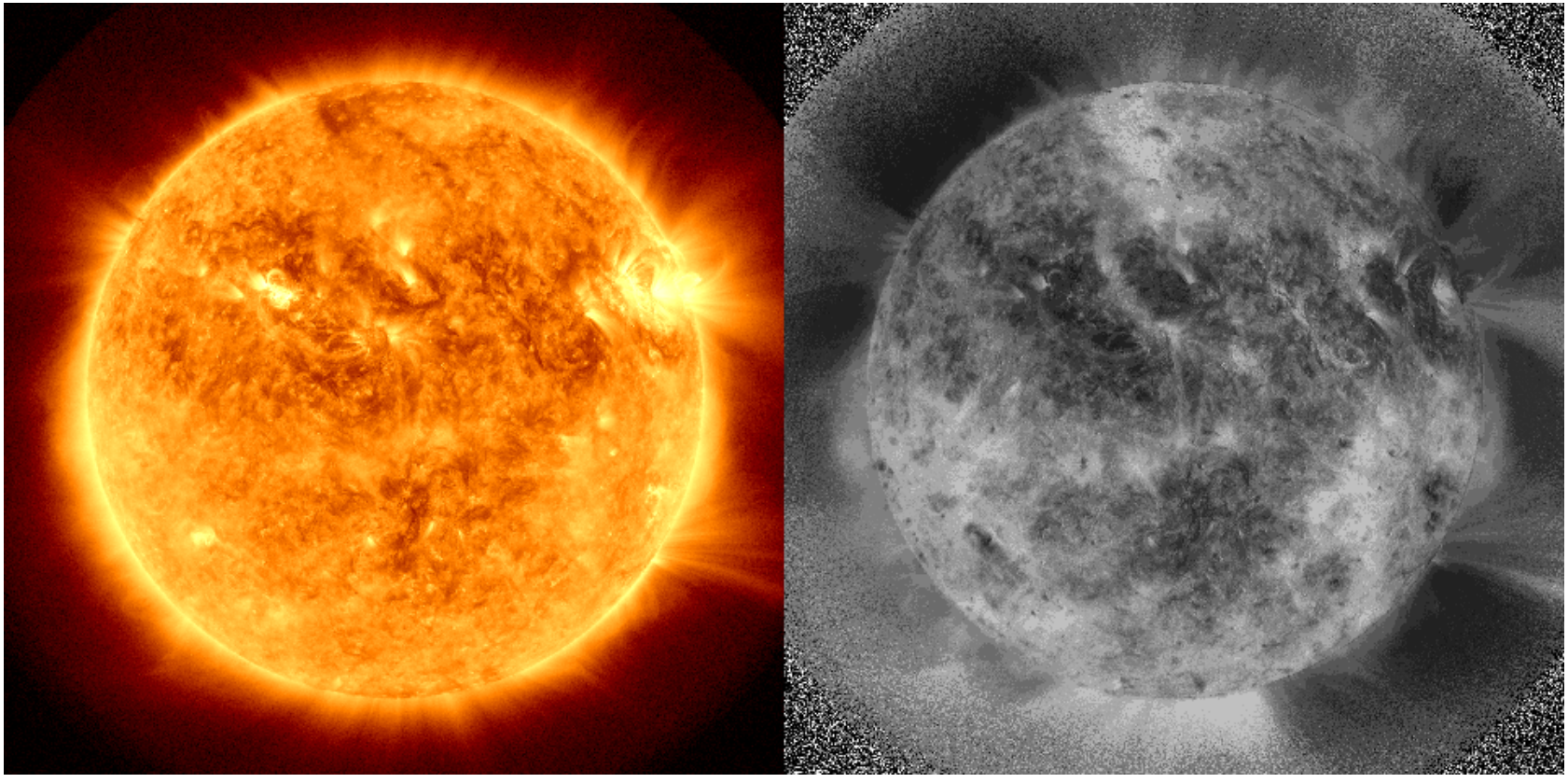
Never Before Seen Structure[2]



Grey Scale Images of Clusters: 2 Oct 2010 at 05.57

[2] Stein, N., Kashyap, V., Meng, X. L., and van Dyk, D. A. (2012). H-Means Image Segmentation to Identify Solar Thermal Features. *Proceedings of the 19th IEEE International Conference on Image Processing, ICIP 2012*

Never Before Seen Structure



Grey Scale Images of Clusters: 2 Oct 2010 at 18.43

Goal: Model, Track, & Forecast structures.

A Family of the Dissimilarity Functions[3]

Cosine Dissimilarity

$$d_{\cos}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \frac{1}{2}\, d_2^2\left(\frac{\boldsymbol{y}_i}{||\boldsymbol{y}_i||_2}, \frac{\boldsymbol{y}_j}{||\boldsymbol{y}_j||_2}\right) = 1 - \frac{\boldsymbol{y}_i^\top \boldsymbol{y}_j}{||\boldsymbol{y}_i||_2\, ||\boldsymbol{y}_j||_2};$$

(One minus cosine of angle between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$.)

- Starting with power transform: $\boldsymbol{T}_j(\boldsymbol{y}_i; \beta, \gamma) = (y_{ij} + \gamma)^\beta$, we construct a parameterized family of dissimilar functions

$$D_{\cos}(\boldsymbol{y}_i, \boldsymbol{y}_j; \beta, \gamma) = d_{\cos}\Big(\boldsymbol{T}(\boldsymbol{y}_i; \beta, \gamma), \boldsymbol{T}(\boldsymbol{y}_j; \beta, \gamma)\Big).$$

- This family includes
  - (L_1-normalized) Hellinger distance and
  - (L_2-normalized) Euclidean distance.

[3] Stein, N. M., van Dyk, D. A., and Kashyap, V. L. (2015). Tuning the Preprocessing of Solar Images to Preserve their Latent Structure. *Statistics and Its Interface*, submitted.

Optimizing the Dissimilarity Function

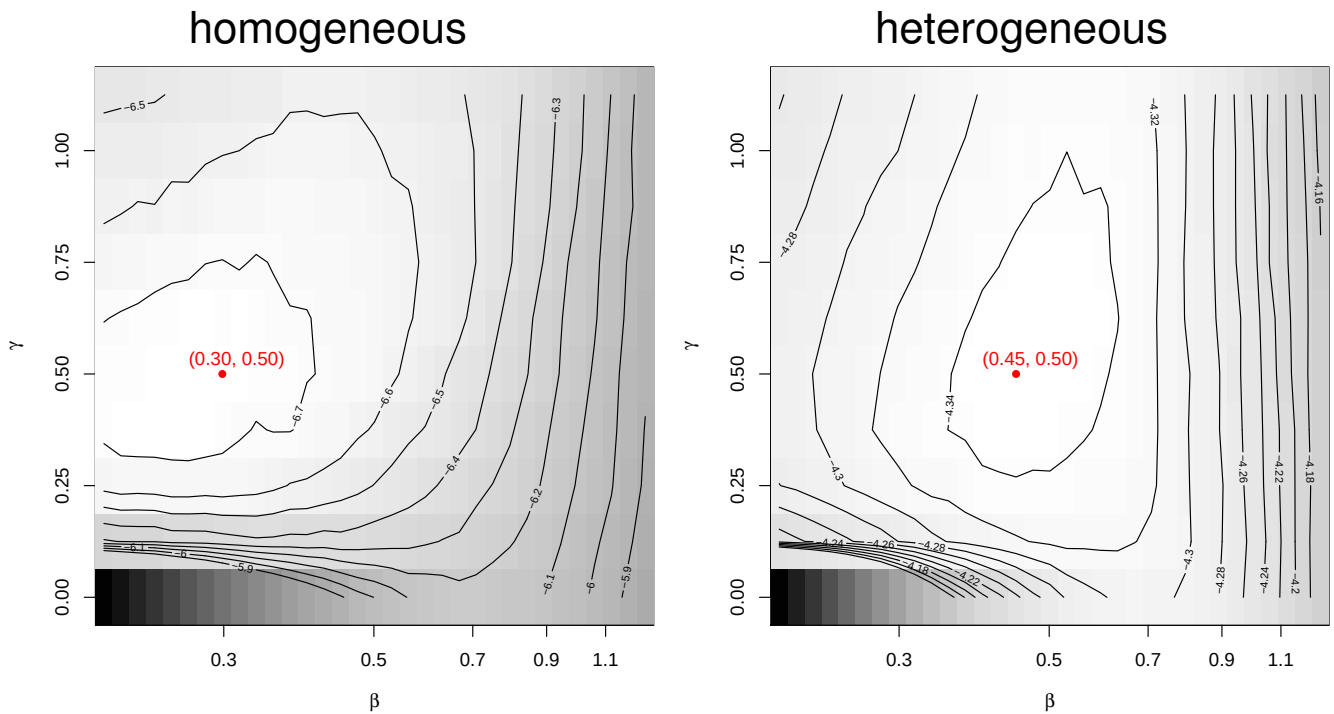
Bayesian decision theoretic approach

- L_2 Loss: $L(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}_{\beta,\gamma}) = ||\boldsymbol{\pi} - \hat{\boldsymbol{\pi}}_{\beta,\gamma}||_2^2$ ($\boldsymbol{\pi}$ is the normalized spectra)
- Risk: $R(\phi, \beta, \gamma) = E\left\{ L(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}_{\beta,\gamma}) \,\middle|\, \boldsymbol{\pi} \right\}$ [E over sampling dist'n of $\hat{\boldsymbol{\pi}}$]
- Bayes Risk: $B(\beta, \gamma) = E\left\{ R(\phi, \beta, \gamma) \right\}$ [E over prior dist'n of $\boldsymbol{\pi}$]
- Choose β and γ to minimize $B(\beta, \gamma)$.

Implementation:

- A given (β, γ) determines a partition of the pixels
- Partition can be found via K-means ($D_{\cos}$ related to d_2)
- $\hat{\boldsymbol{\pi}}_{\beta,\gamma}$ is assumed constant in each cluster of pixels
- Using prior on DEM, minimize $B(\beta, \gamma)$ via Monte Carlo

Simulation Study

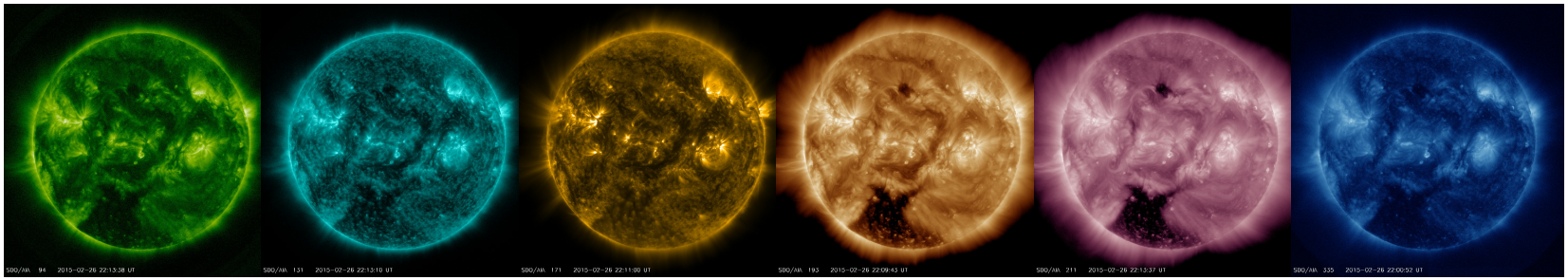


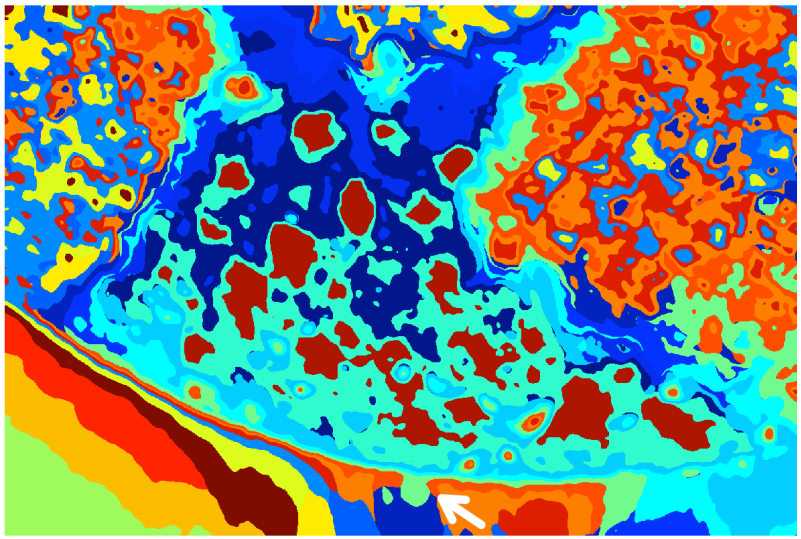
Simulation setup and results:

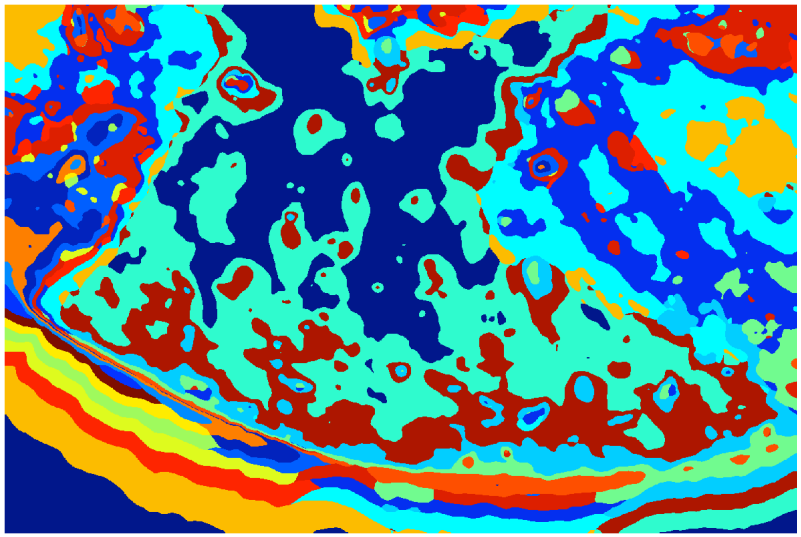
- 50 pixel images; 2 clusters (homogeneous/heterogeneous)
- Compare with $d_{\cos}$ ($\beta = 1, \gamma = 0$ and d_{H} ($\beta = 1/2, \gamma = 0$).
- Advantage of added pseudo counts with Hellinger dist
- Similar results with Rand index (measure of true partition recovery)

Imperial College London

A Closer Look at a Coronal Hole (Feb 26, 2015)

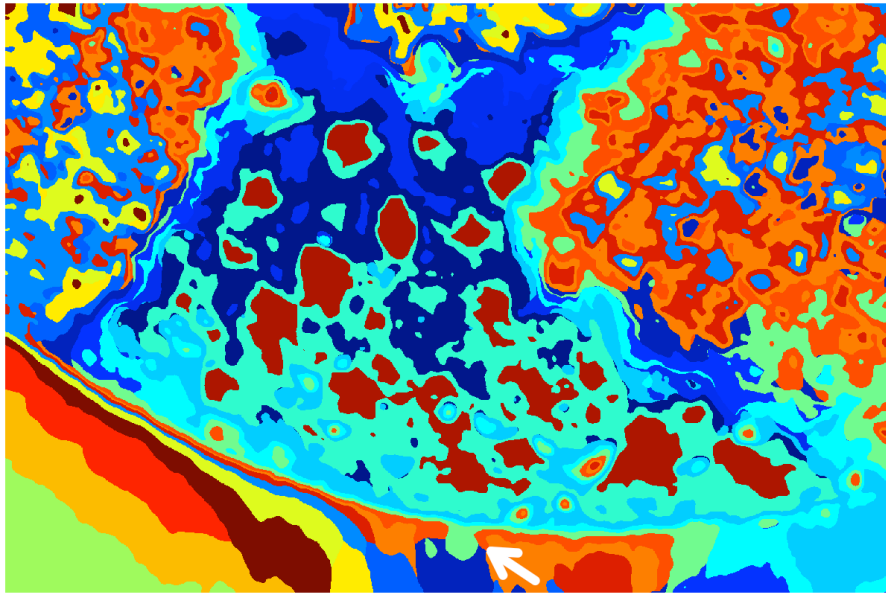


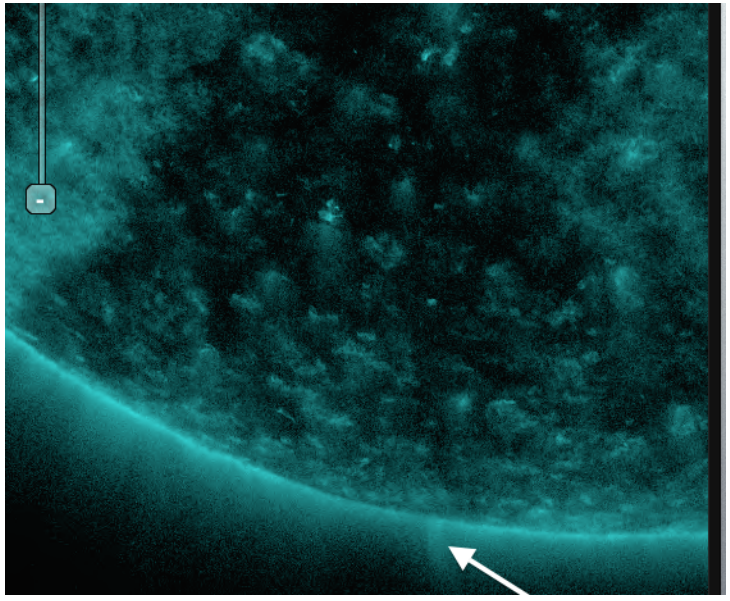




- 20 clusters with (a) $\beta = \gamma = 1/2$; (b) standard K-means
- off-limb: funnel shape structures versus horizontal striations
- boarder: gradual versus abrupt transition

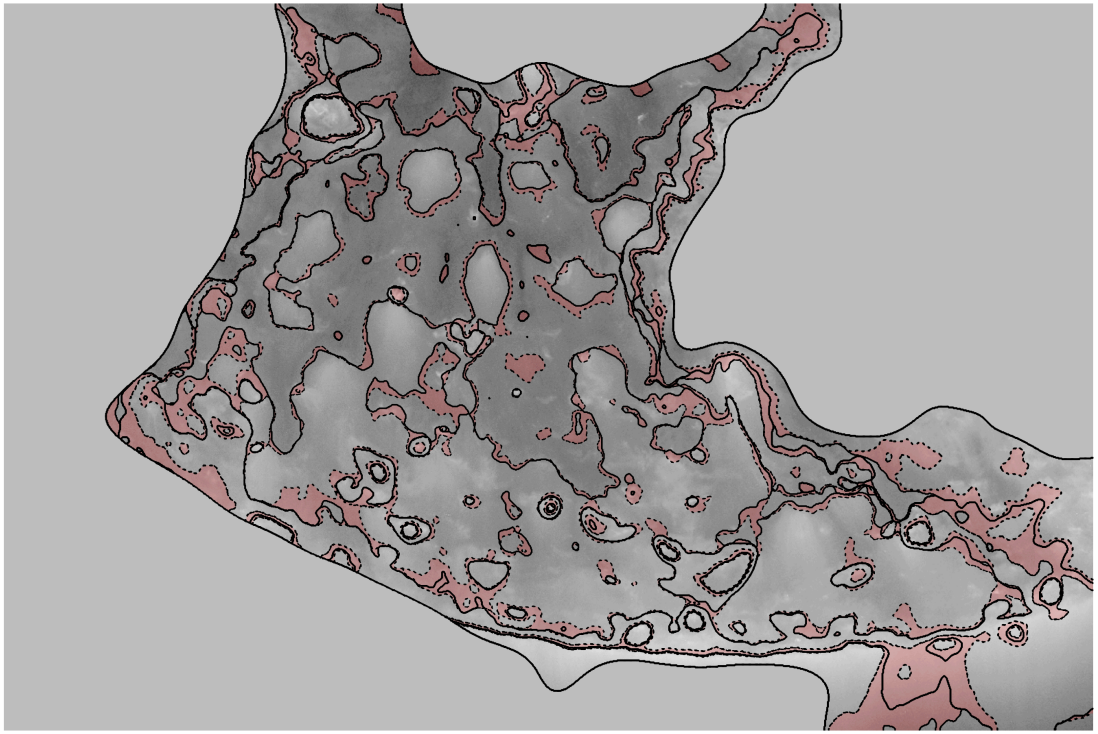
Can we Predict Coronal Activity?





- Left arrow marks bulb-shaped region of hot plasma
- Indicative of stressed magnetic field and energy release
- Right arrow: an eruption becomes visible $\sim \frac{1}{2}$ hour later

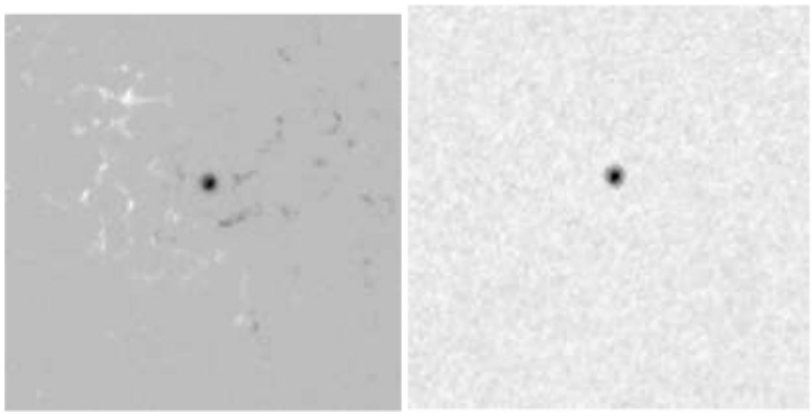
Fast Image Segmentation



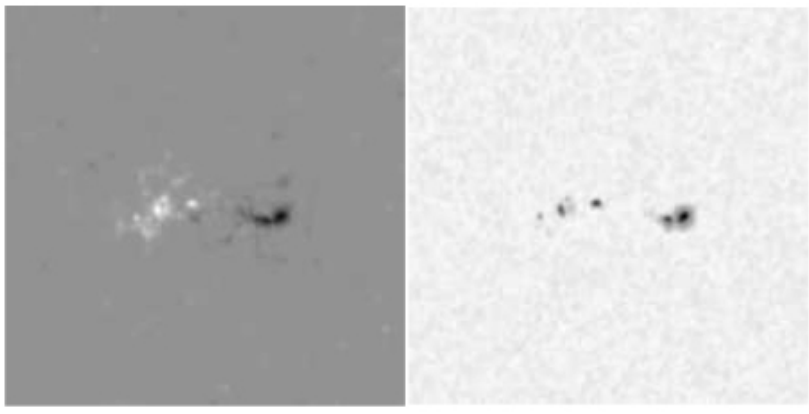
- Comparing two segmentations; $(\beta, \gamma) = (0.3, 0.5), (0.5, 0.5)$.
- Differences are highlighted in pink: adj Rand index = 0.91.
- Direct DEM reconstruction: 3-5 hours + bootstrap errors (2.5GHz machine; 1024×1024 image)
- Our segmentation: quick ($\sim$ 2 min); enhances thermal structure; spatial cohesion not imposed

Mt Wilson Classification of Sun Spots

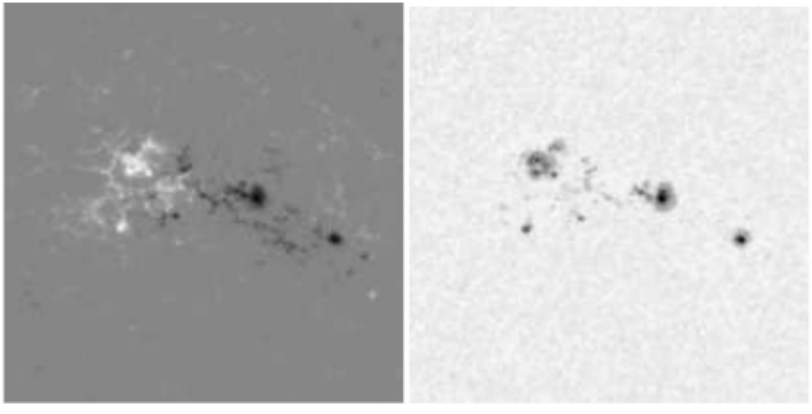
α class



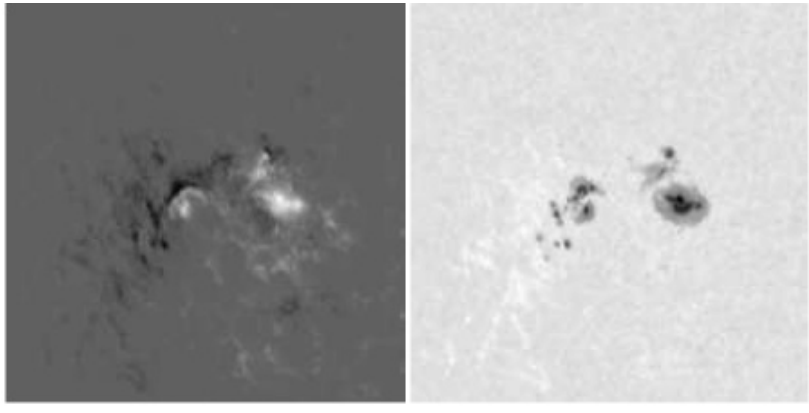
β class



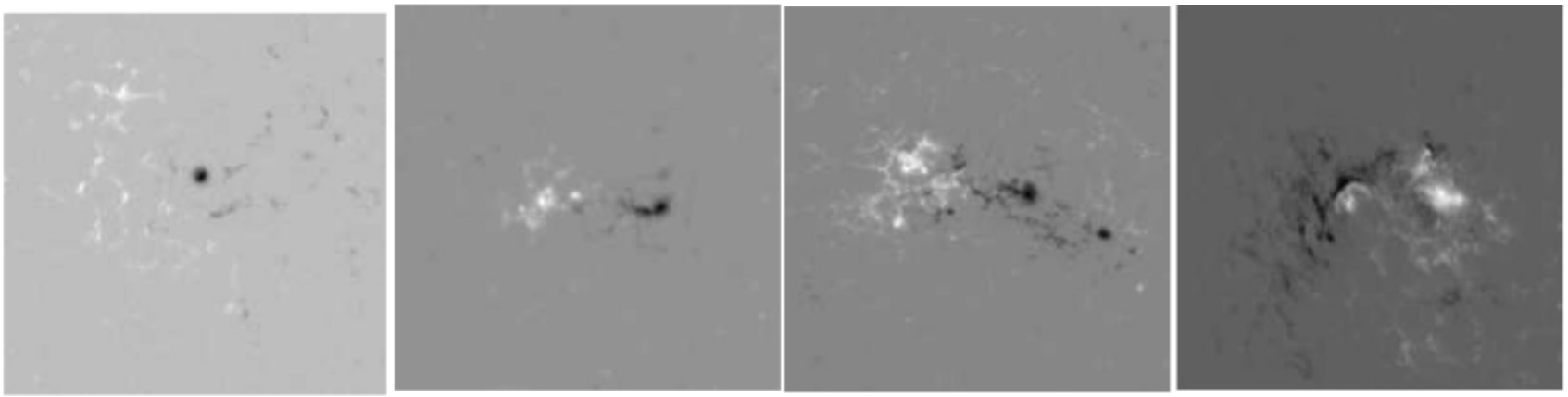
$\beta\gamma$ class



$\beta\gamma\delta$ class



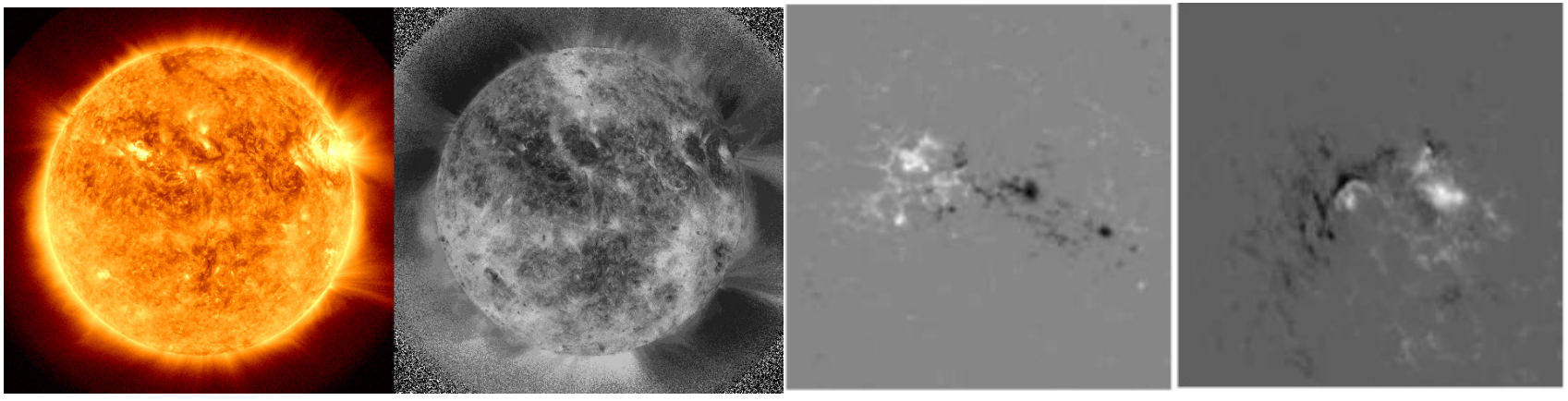
Automatic Classification of Sun Spots[4]



- Classification is predictive for activity in solar corona
- Sunspots are typically classified manually
- Automate classification: **science-driven feature selection**
- Features from mathematical morphology (e.g., area of overlap of polarities, roughness of separating line, etc.)
- Features have physical meaning beyond classification
- **Goal:** Use to model, track, and forecast evolution (movie)

[4] Stenning, D. C., Lee, T. C. M, van Dyk, D. A., Kashyap, V. L., Sandell, J., and Young, C. A. (2013). *Statistical Analysis and Data Mining*, **6**, 329–345.

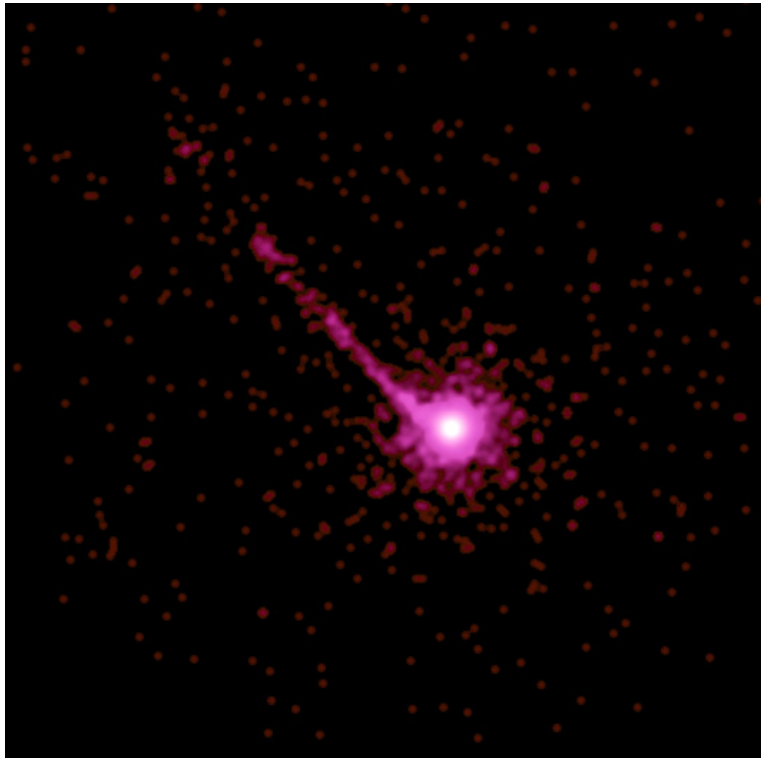
Big Data Methods in Solar Physics



- Use science-driven models to inform data-driven methods.
  - E.g., cluster spectra for DEM; optimal choice of distance.
  - E.g., tuning feature extraction to Mt. Wilson classification.
- Reduce the complexity of data to understandable features.
  - E.g., summarize images w/ feature for secondary analysis.
- More efficient to split initial and secondary analyses.

Example II: Testing for Unexpected Features

- Supermassive black holes at the center of distant galaxies appear as **Quasars**.
- **X-ray jets** extend millions of light years from some quasars
- Chandra X-Ray Observatory counts photons in pixels
- We observe a low-count images with a possible jet



Can we infer jet structure and quantify the significance of a jet detection?

How to infer jet structure?

We fit a hierarchical Bayesian model that accounts for:

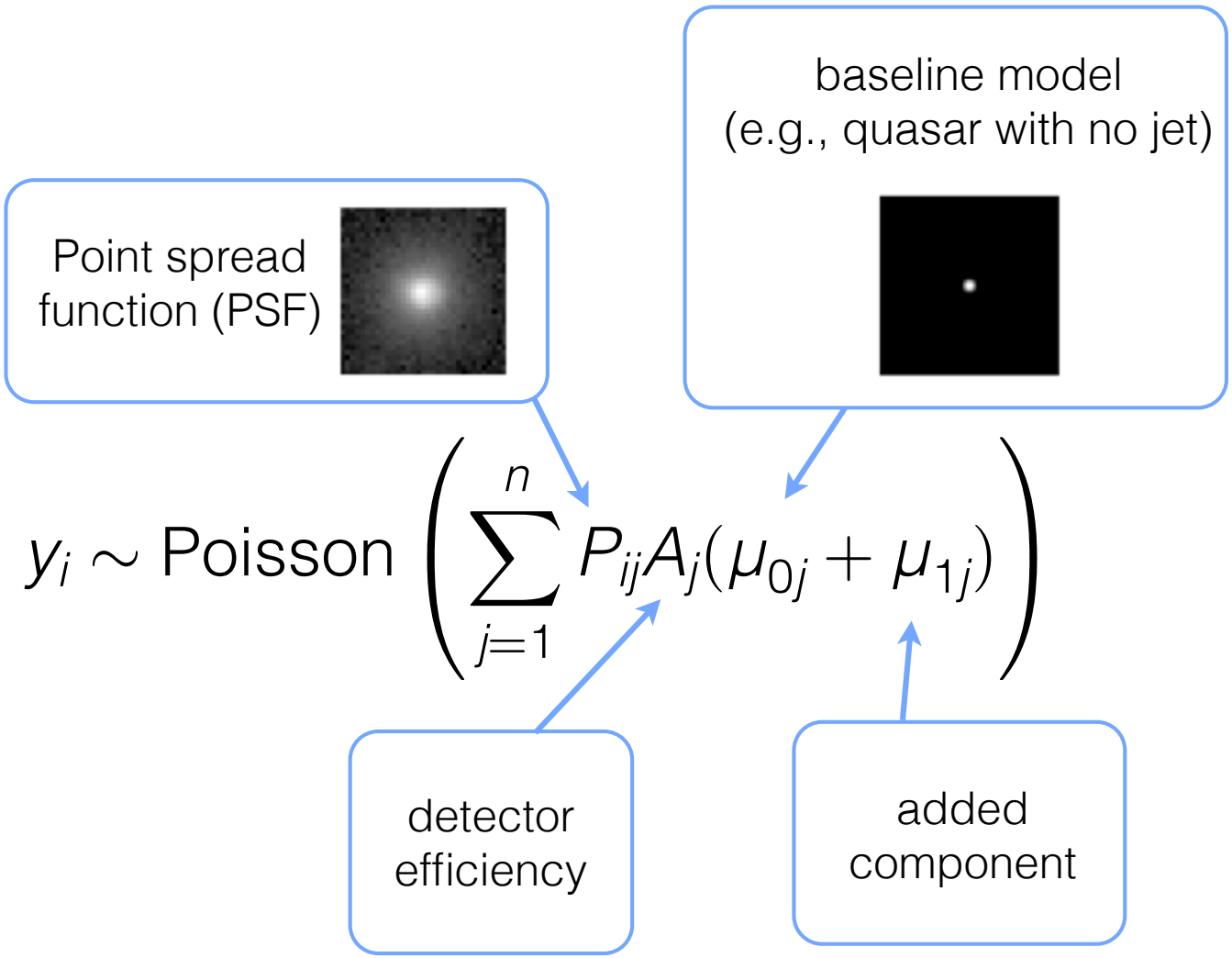
- Poisson noise
- Point spread function
- Detector inefficiencies
- Spatial correlations at multiple resolutions

- Jet structure not well specified a priori $\rightarrow$ difficult to parameterize
- The full hierarchical model includes a baseline model for quasar+background, and allows for nonparametric departures from this baseline
- Implemented in the R package `LIRA` (Low-count Image Reconstruction and Analysis), available at

`github.com/astrostat/LIRA`

Statistical model: Likelihood

Observe photon counts $(y_1, \ldots, y_n)$ in n pixels



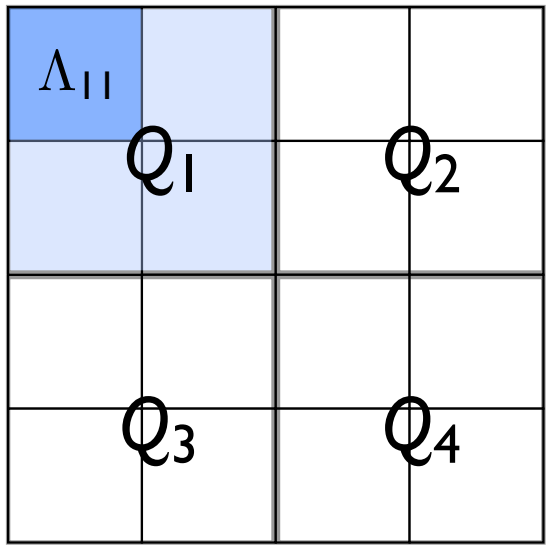
Statistical model: Multiscale smoothing prior on μ_1

Model:

$$y_i \sim \text{Poisson}\left(\sum_{j=1}^{n} P_{ij} A_j (\mu_{0j} + \mu_{1j})\right)$$

- Parameterize $\boldsymbol{\mu}_i = \tau_i \boldsymbol{\Lambda}_i$ for $i = 0$ or 1, where $\tau_i = \sum_{j=1}^{n} \mu_{ij}$
- $\boldsymbol{\Lambda}_0$ is specified according to baseline model
  - for example: quasar + constant background
- Wavelet-like smoothing prior on $\boldsymbol{\Lambda}_1$:

$$\Lambda_{1i} = \left(\sum_{j \in Q_{k(i)}} \Lambda_{1j}\right)\left(\frac{\Lambda_{1i}}{\sum_{j \in Q_{k(i)}} \Lambda_{1j}}\right)$$



Statistical model: Multiscale smoothing prior on μ_1

- We consider $2^D \times 2^D$ images and write

$$\text{pixel probability} = \prod_{k=1}^{D} \text{conditional probability at resolution } k$$

- We use a hierarchical model to allow different amounts of smoothing at different resolutions:

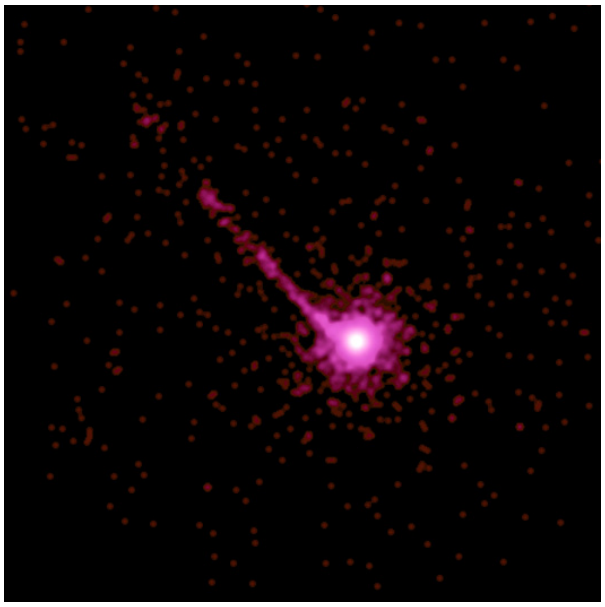
$$(\text{Conditional probability at resolution } k) \sim \text{Dirichlet}\{(\psi_k, \psi_k, \psi_k, \psi_k)\}$$
$$(\psi_1, \ldots, \psi_D) \sim \pi(\psi)$$

- This multiscale representation and the choice of prior $\pi(\psi)$ on the smoothing parameters are from Esch et al. (2004).[5]

[5] Esch, D. N., Connors, A., Karovska, M., and van Dyk, D. A. (2004). An Image Reconstruction Technique with Error Estimates. *The Astrophysical Journal*, **610**, 1213–1227.

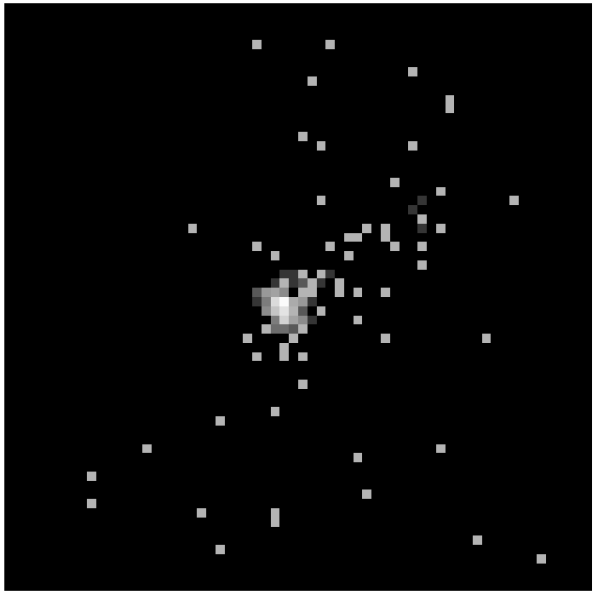
Some jets are obvious, and some are not ...

An obvious jet:



X-ray image of the quasar PKS 1127-145 with a ∼million light year jet
Credit: NASA/CXC/ A.Siemiginowska(CfA)/J.Bechtold(U.Arizona)

Not so obvious:



X-ray image of the quasar 0730+257
Source: Stein, van Dyk, Kashyap, Siemiginowska (2015+)

Quantifying Detection Significance[6]

Hypothesis testing framework

- Null hypothesis: point source + constant background
- Alternative hypothesis: null model + LIRA model

- **Procedure:**
  1. Fit full model to observed image (to account for PSF, etc.)
  2. Fit same model to images simulated from null model: no jet
  3. Compare scalar summaries of posterior distributions
- **Problem:** complicated model fitted via MCMC, *must limit N*

How to quantify the significance of jet detection ***under computational constraints****?*

[6] Stein, N. M., van Dyk, D. A., Kashyap, V. L., and Siemiginowska, A. (2015). Detecting Unspecified Structure in Low-Count Images. *The Astrophysical Journal*, tentatively accepted.
McKeough, K., Siemiginowska, A., Kashyap, V. L., Stein, N. M., van Dyk, D., et al. (2015). Chadra X-ray Imaging of the Highest-Redshift Quasar Jets *The Astrophysical Journal*, in preperation.

Naive Monte Carlo p-values

- Suppose we have a test statistic S
- S_{obs} is the test statistic computed from the observed image
- $S_1, \ldots, S_N$ are the statistics computed from N null images
- Two common Monte Carlo p-values:

$$\hat{p}_0 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(S_i \geq S_{\text{obs}}), \qquad \hat{p}_1 = \frac{1 + \sum_{i=1}^{N} \mathbb{I}(S_i \geq S_{\text{obs}})}{N+1}$$

Problems when N is not very large

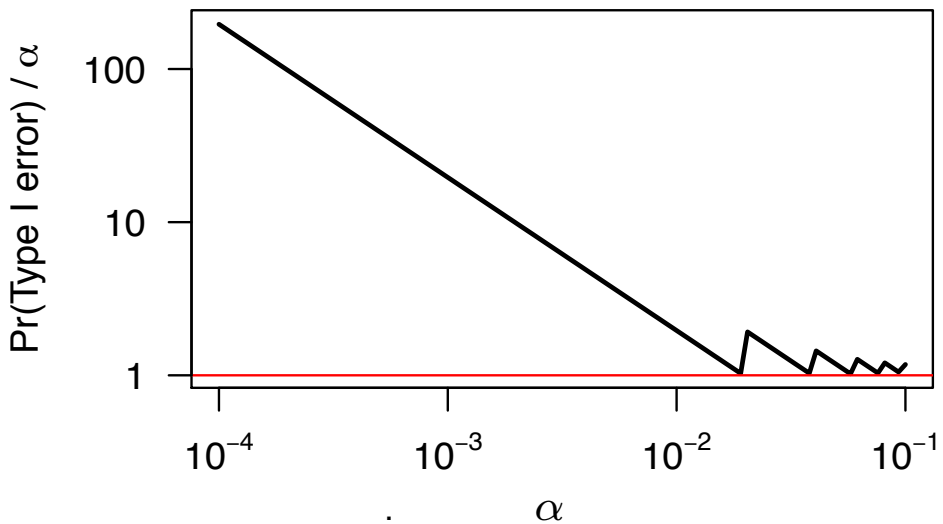
- $\hat{p}_0$ leads to too many false positives: e.g., $\Pr(\hat{p}_0 = 0 | H_0) = \frac{1}{N+1}$
- $\hat{p}_1 \geq 1/(N+1)$, so impossible to achieve high significance

Hypothesis Testing and Type I Error

Hypothesis Testing

- Choose level α and calculate p-value: p
- Reject the null hypothesis if $p \leq \alpha$

Suppose $N = 50$ and we reject if $\hat{p}_0 \leq \alpha$:



With the Monte Carlo p-value, $\hat{p}_0$,
Pr(Type I error) may be significantly larger than α

Our Approach

How to achieve high significance when N is not very large?

- Use a **posterior tail probability** as a test statistic
- Estimate an **upper bound** on a p-value

- Estimate an upper bound $\hat{u}$ on a p-value
- Reject null hypothesis if $\hat{u} \leq \alpha$
- We can test at **high significance levels** (small α) if we can obtain small $\hat{u}$ even when N is not large
- Rejecting when $(p \leq)$ $\hat{u} \leq \alpha$ guarantees Pr(Type I error) $\leq \alpha$

A novel test statistic

Model:

$$y_i \sim \text{Poisson}\left(\sum_{j=1}^{n} P_{ij} A_j (\tau_0 \boldsymbol{\Lambda}_{0j} + \tau_1 \boldsymbol{\Lambda}_{1j})\right)$$

- Choose a scalar parameter ξ, e.g., $\xi = \tau_1/(\tau_0 + \tau_1)$
- Test statistic: a posterior probability under the full model:

$$S_c(\boldsymbol{y}_{\text{obs}}) = \Pr(\xi \geq c \mid \boldsymbol{y}_{\text{obs}})$$

- Large $\hat{\xi}$ favor $H_A \longrightarrow$ large $S_c(\boldsymbol{y}_{\text{obs}})$ favor H_A.
- How large? Use the p-value

$$p = \Pr\{S_c(\boldsymbol{y}_0) \geq S_c(\boldsymbol{y}_{\text{obs}})\}, \quad \text{with} \quad \boldsymbol{y}_0 \sim H_0$$

Constructing an upper bound on p

By Markov's inequality:

$$p = \Pr\{S_c(\boldsymbol{y}_0) \geq S_c(\boldsymbol{y}_{\text{obs}})\} \leq \frac{E\{S_c(\boldsymbol{y}_0)\}}{S_c(\boldsymbol{y}_{\text{obs}})} = \frac{\Pr_0(\xi \geq c)}{\Pr(\xi \geq c \mid \boldsymbol{y}_{\text{obs}})} = u$$

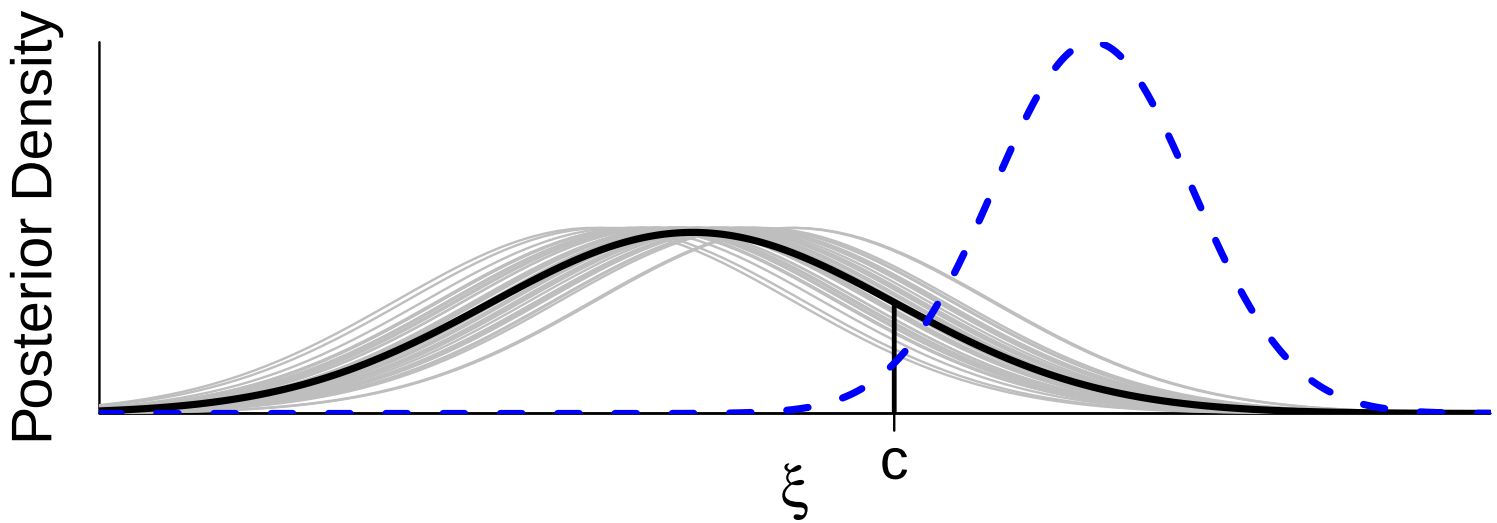
[$\Pr_0(\xi \geq c)$ is with respect to $g(\xi) \equiv E\{\pi(\xi \mid \boldsymbol{y}_0)\} = \int p(\xi|\boldsymbol{y}_0)p(\boldsymbol{y}_0|H_0)d\boldsymbol{y}_0$.]

The Procedure

- Sample $\boldsymbol{y}_0^{(1)}, \ldots, \boldsymbol{y}_0^{(N)}$ from the null model
- Obtain samples from each posterior: $\pi(\xi \mid \boldsymbol{y}_0^{(n)})$, $\pi(\xi \mid \boldsymbol{y}_{\text{obs}})$
- Fix $\gamma = \Pr_0(\xi \geq c)$ and Estimate
  1. $\hat{g}(\xi) = N^{-1} \sum_{i=1}^{N} \pi(\xi \mid \boldsymbol{y}_0^{(i)})$
  2. $\hat{c}$ as the $(1-\gamma)$ quantile of $\hat{g}(\xi)$
  3. $\hat{S}_{\hat{c}}(\boldsymbol{y}_{\text{obs}}) = \widehat{\Pr}(\xi \geq \hat{c} \mid \boldsymbol{y}_{\text{obs}})$
  4. $\hat{u} = \gamma / \hat{S}_{\hat{c}}(\boldsymbol{y}_{\text{obs}})$

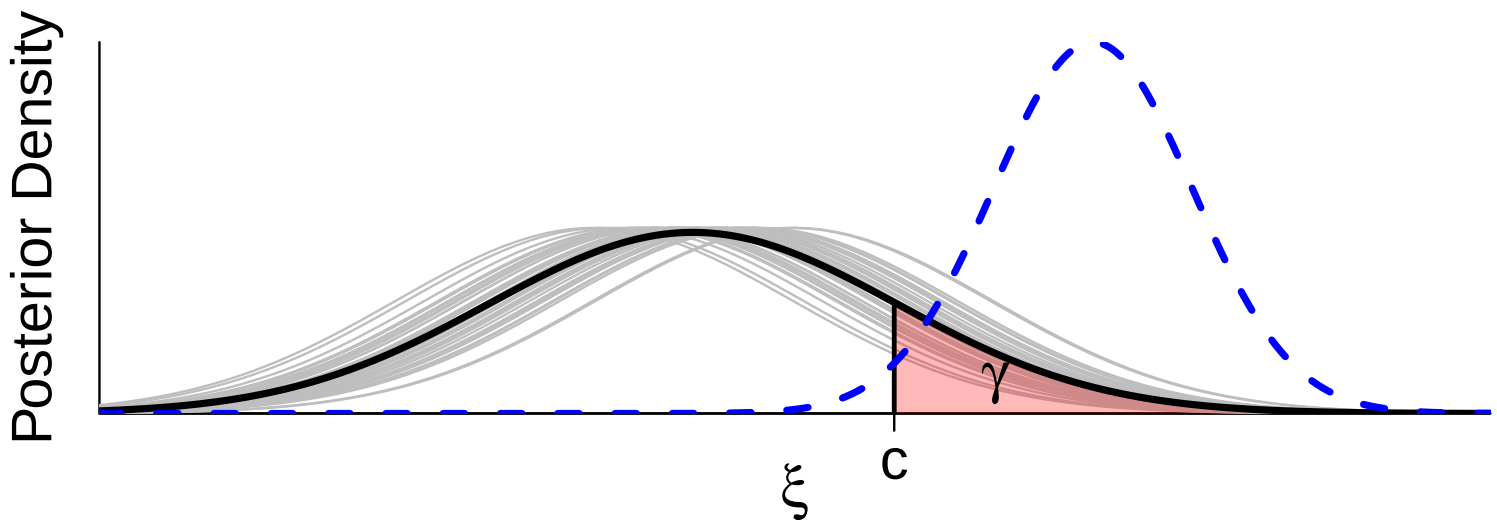
The estimated upper bound:

$$\hat{u} = \gamma / \hat{S}_{\hat{c}}(\boldsymbol{y}_{\text{obs}})$$



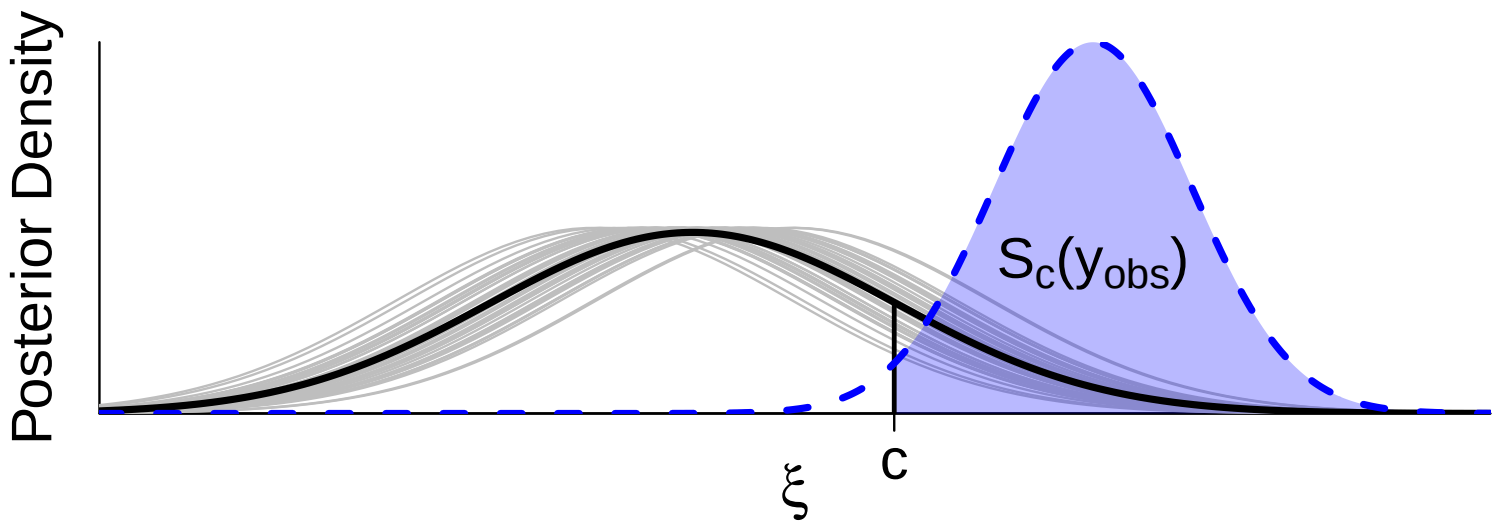
The estimated upper bound:

$$\hat{u} = \gamma / \hat{S}_{\hat{c}}(\boldsymbol{y}_{\text{obs}})$$

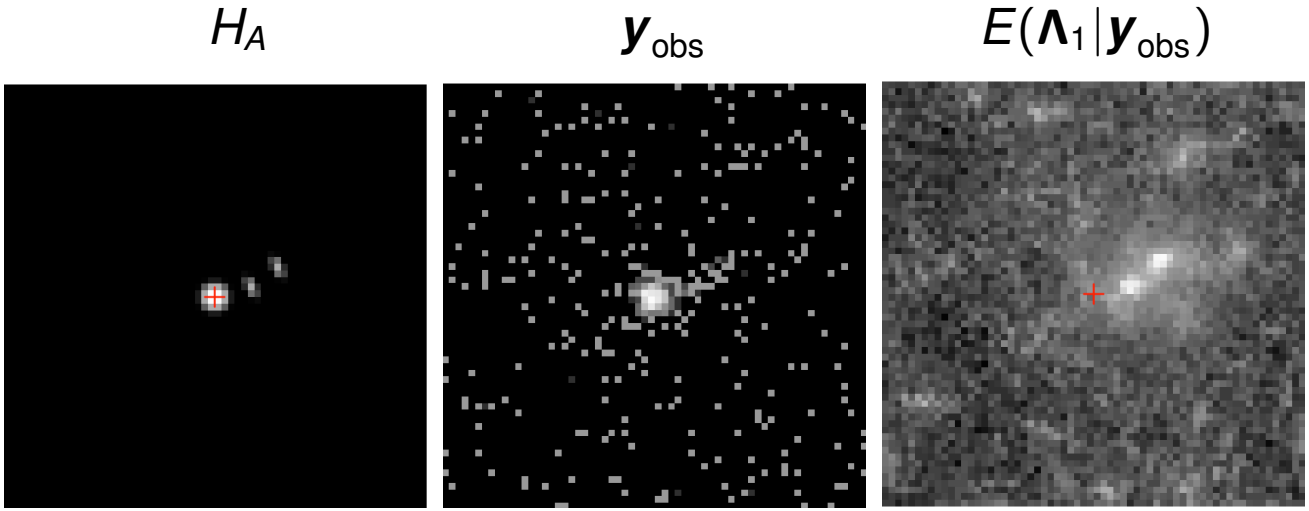


The estimated upper bound:

$$\hat{u} = \gamma / \hat{S}_{\hat{c}}(\boldsymbol{y}_{\text{obs}})$$



Simulation



- H_0 : point source + background
  H_A : point source + background + two elliptical sources
- $N = 50$ null images for each of 1000 simulation $\boldsymbol{y}_{\text{obs}}$
- Compared three methods, rejecting the null hypothesis if
  1. $\hat{u} \leq \alpha$
  2. $\hat{p}_1 \leq \alpha$
  3. $\hat{p}_0 \leq \alpha$, each for a variety of choices of α, γ

Simulation Results

- Expected count in simulated jet = 40

		Type I error rate (%)			Power (%)		
γ (%)	α (%)	$\hat{u}$	$\hat{p}_1$	$\hat{p}_0$	$\hat{u}$	$\hat{p}_1$	$\hat{p}_0$
1.0	2.0	0.1	2.0	3.9	**99.7**	**100.0**	100.0
0.5	2.0	0.3	2.0	3.9	**99.7**	**100.0**	100.0
	1.0	0.1	0	2.0	**97.6**	0	100.0
0.1	2.0	0.8	2.0	3.9	**99.6**	**99.8**	100.0
	1.0	0.2	0	2.0	**98.4**	0	99.8
	0.5	0.1	0	2.0	**96.2**	0	100.0

- Because $\hat{u} \geq \gamma$, it is impossible to reject if $\gamma > \alpha$.
- **Bold** indicates the best power (up to simulation precision) among methods with appropriately bounded Type I error rates.

Simulation Results

- Expected count in simulated jet = 20

		Type I error rate (%)			Power (%)		
γ (%)	α (%)	$\hat{u}$	$\hat{p}_1$	$\hat{p}_0$	$\hat{u}$	$\hat{p}_1$	$\hat{p}_0$
1.0	2.0	0.1	2.0	3.9	29.8	**74.0**	81.7
0.5	2.0	0.4	2.0	3.9	41.4	**70.9**	80.7
	1.0	0.0	0	2.0	**18.1**	0	72.0
0.1	2.0	0.8	2.0	3.9	48.8	**67.8**	78.6
	1.0	0.5	0	2.0	**33.2**	0	66.8
	0.5	0.0	0	2.0	**19.3**	0	66.8

Discussion: The Big Picture

- **Principled methods** offer statistical optimality and allow us to incorporate the complexities of real-world science
- But may be **computational expensive**: statistical optimality may not be affordable
- **Pragmatic perspective:** trade statistical for computational efficiency to achieve feasible nearly-optimal methods
  - Reduce data complexity: segmenting solar images
  - Let science-based understanding inform data reduction and preprocessing (e.g., cluster normalized spectra)
  - Trade statistical power for faster computing: for fixed computing time, $\hat{u}$ allows valid tests with smaller α than $\hat{p}_0$ or $\hat{p}_1$

Thanks...

Solar Thermal Structure:

- Nathan Stein
- Vinay Kashyap
- Xiao-Li Meng

Classifying Active Regions:

- David Stenning
- Vinay Kashyap
- Thomas Lee

X-ray Feature Detection:

- Nathan Stein
- Vinay Kashyap
- Aneta Siegminowska
- Kathryn McKeough

And

The CHASC International Astrostatistics Center

For Further Reading I


Esch, D. N., Connors, A., Karovska, M., and van Dyk, D. A.
An Image Reconstruction Technique with Error Estimates.
The Astrophysical Journal, **610**, 1213–1227, 2004.


Stenning, D., Lee, T., van Dyk, D., Kashyap, V., Sandell, J., and Young, C.
Morphological Feature Extraction for Statistical Learning [in Solar Images]
Statistical Analysis and Data Mining, **6**, 329–345, 2013.


Stein, N., Kashyap, V., Meng, X. L., and van Dyk, D. A.
H-Means Image Segmentation to Identify Solar Thermal Features.
Proceedings of the 19th IEEE Intern'l Conf. on Image Processing, ICIP 2012


Stein, N. M., van Dyk, D. A., and Kashyap, V. L.
Tuning the Preprocessing of Solar Images to Preserve their Latent Structure.
Statistics and Its Interface, submitted, 2015.


Stein, N. M., van Dyk, D. A., Kashyap, V. L., and Siemiginowska, A.
Detecting Unspecified Structure in Low-Count Images.
The Astrophysical Journal, tentatively accepted, 2015.


McKeough, K., Siemiginowska, A., Kashyap, V. L., Stein, N. M., van Dyk, D., et al.
Chadra X-ray Imaging of the Highest-Redshift Quasar Jets
The Astrophysical Journal, in preperation, 2015.

Imperial College London

Get Involved!

Association of Astrostatisticians

- New ASA Interestgroup:
  `http://community.amstat.org/astrostats/home`
- International Astrostatistics Association (New! Working Groups!)
- Astrostatistics and Astroinformatics Portal:
  `http://asaip.psu.edu`

Data Challenges Competitions

- Banff Challenge 1 & 2: Davison and Sartori (Stat Sci, 2008)
  `http://www.birs.ca/events/2010/5-day-workshops/10w5068`
- GREAT08 & GREAT10: `arXiv:0908.0945v1 arXiv:1202.5254v2`
- Strong Lens Time Delay: `http://timedelaychallenge.org`

Vast Public Data Resources

- The Virtual Observatory: `http://www.usvao.org`
- The Sun Today: `http://www.thesuntoday.org`

Mathematical Morphology for Solar Features

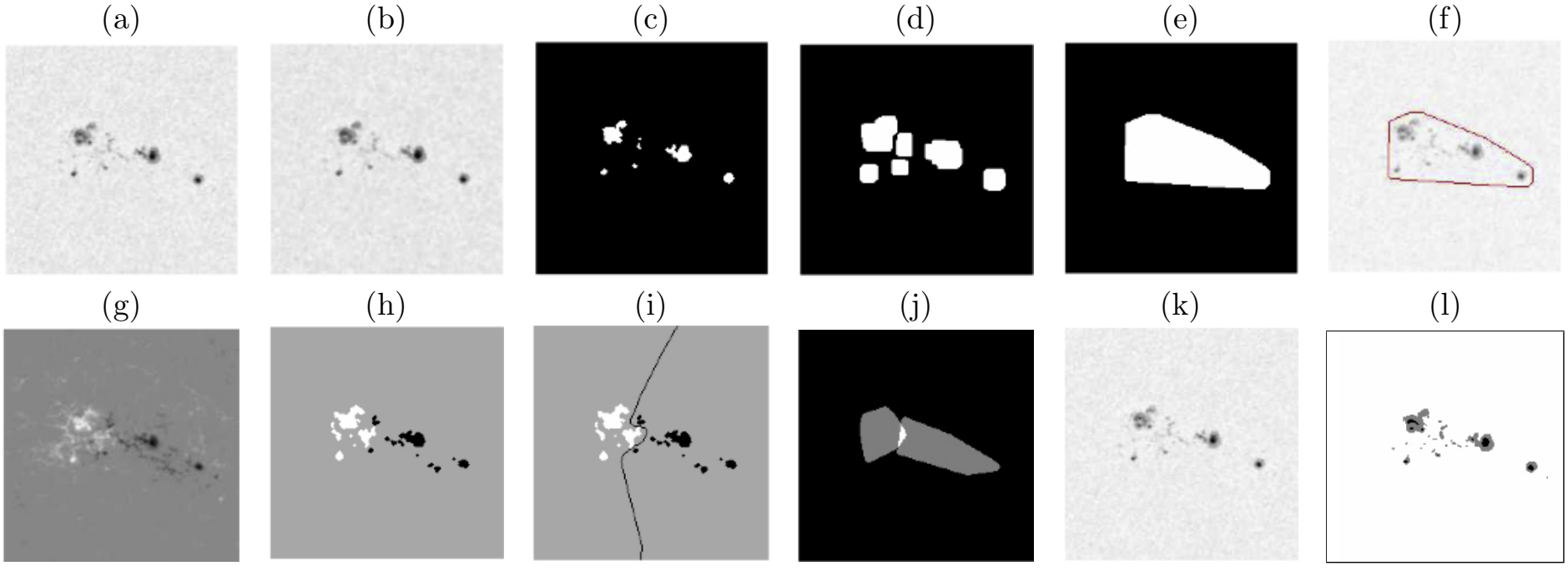


Figure 3 The original $\beta\gamma$ white light image (a) is cleaned (b) and thresholded to produce a binary representation of the sunspot group (c). This image is then dilated (d) and has a convex hull placed around the result (e) and the area inside the hull becomes the sunspot area (f) in the magnetogram. Then, in the $\beta\gamma$ magnetogram (g), morphological opening followed by thresholding on both the image and inverse image yields the trinary primal sketch of the active region in (h). Region growing gives the separating boundary in (i). Convex hulls are utilized to measure polarity mixture in (j). We smooth the white light image in (k) and apply thresholding iteratively in (l) to produce a representation of the umbrae and penumbrae that can be used to detect delta spots.