Sparse modeling: some unifying theory and "word-imaging"

Bin Yu

UC Berkeley
Departments of Statistics, and EECS

Based on joint work with:

Sahand Negahban (UC Berkeley)
Pradeep Ravikumar (UT Austin)
Martin Wainwright (UC Berkeley)

B. Gawalt, J. Jia, L. Miratrix, L. El Ghouai
S. Clavier (International Relations, SFSU)

IT Revolution → Data Revolution

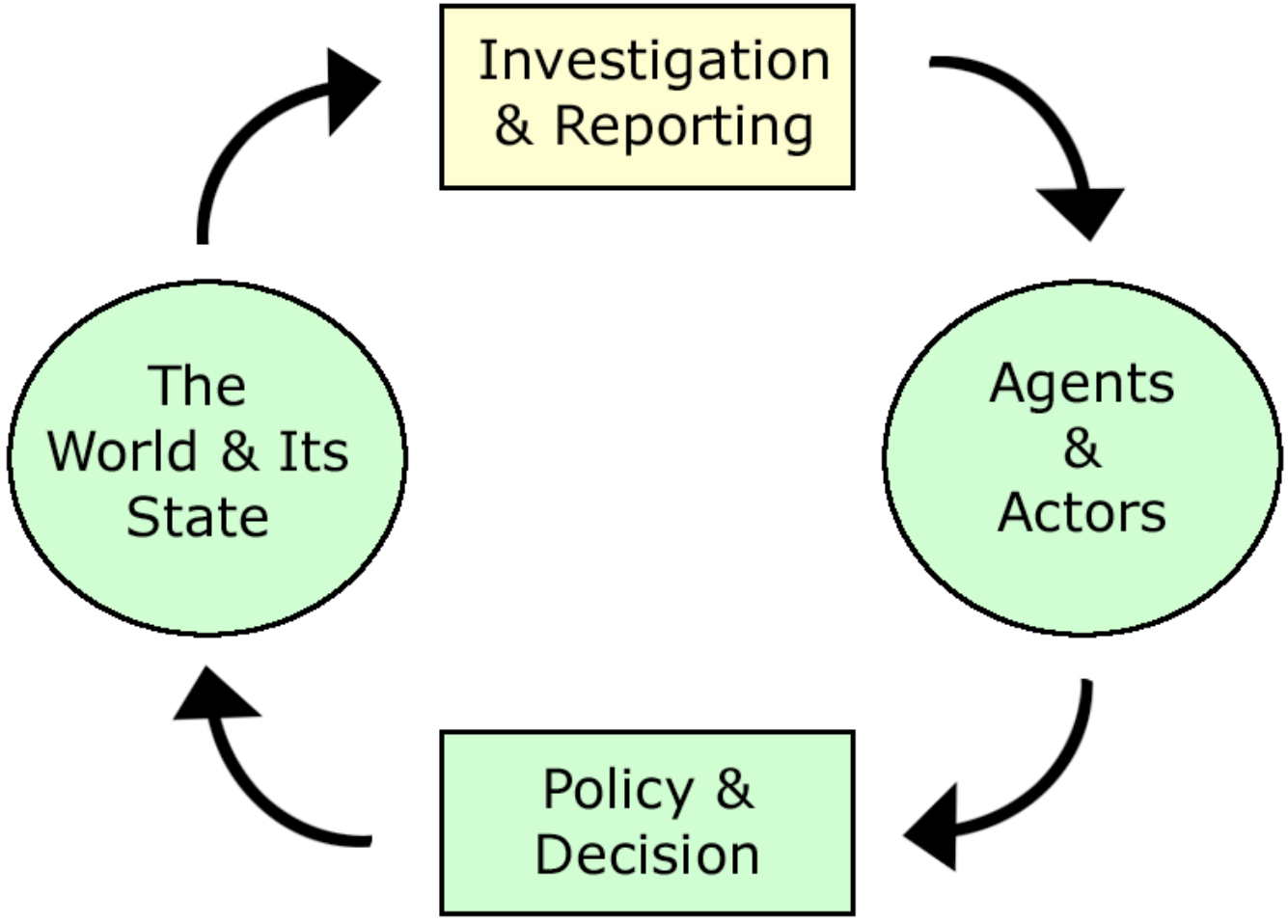
- Modern data problems in science and engineering (p and n large)

- Examples:
  - genomics
  - remote sensing
  - imaging science (e.g., medicine, astronomy etc.)
  - neuroscience

- Modern data problems in social sciences (p and n large)

- Examples:
  - document clustering
  - natural language processing
  - social networks (e.g FEC (Federal Election Campaign) data)
  - Word-imaging: how is "China" portrayed in NYT international section?

A Tidy Model of How the World Works



Standards and Practices

- To be persuasive we must be believable; to be believable we must be credible; to be credible we must be truthful.
  - *Edward R. Murrow*

... versus...

- You supply the photographs, and I'll supply the war.
  - *William Randoph Hearst*

Improve News Media Analysis, Improve News Media, Improve How the World Works

- Holes in current approach
  - Time and labor constraints
  - Case study approach too prone to bias
- Statistical machine learning techniques
  - Fast, scale well
  - Reproducible results
  - Designed around predictive tasks
- Harness machine learning to power media studies
  - New predictive framework needed for media study
  - New design guidelines and metrics needed for machine learning

Our application: word image in the New York Times

- Word Image: a small set of words describing/distinguishing a topic
- As a predictive problem:
  - Predict appearance of a query word q in a document from the document's use of other words
- Predictive model must be interpretable
  - Predictor weights must directly and simply drive label
  - No. of predictors used must be few: sparse model **approximation**
  - The faster predictors can be computed, the better
- Chosen predictor words form a set known as the Word Image for q
- Word image must be evaluated two ways:
  - Can labels (appearance indicator for q) be effectively predicted?
  - Are the chosen words meaningful w.r.t. q?

Solving a modern data problem (word-imaging)

- Data processing (much work, done by EE collaborators)
- Subject knowledge (political scientist, international relation expert)
- Methodology or algorithms (ℓ_1 logistic regression)
- Theory (understanding sparse methods in idealized situations to build intuition in high-dim space)
- Validation (human experiments)

Recent lessons from high-dim (p large) theory

- Low dimensional structures are needed for meaningful information extraction on parameters ("consistency" and "rates")
  - when $p \gg n$, require additional constraints on structure/complexity
    - ⋆ sparsity
    - ⋆ low-rank matrices
    - ⋆ manifold

- Sparse structures are well suited for interpretation, visualization, and computation/transmission/storage.
  - Pursuit of "simplicity" – modern model selection
  - Sparsity holds well for some problems and necessary for others.
  - Much research on sparse modeling lately...

Remainder of the talk

- Unified analysis on ℓ_2 estimation error for M-estimation with decomposable regularizers in high-dim (p large)

- "Word-imaging" through sparse predictive modeling and human validation

Loss functions and regularization

- **Model class:** parameter space $\Omega \subset \mathbb{R}^p$, and set of probability distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$

- **Data:** samples $\mathcal{X}_1^n = (x_i, y_i)$, $i = 1, \ldots, n$ are drawn from unknown $\mathbb{P}_{\theta^*}$

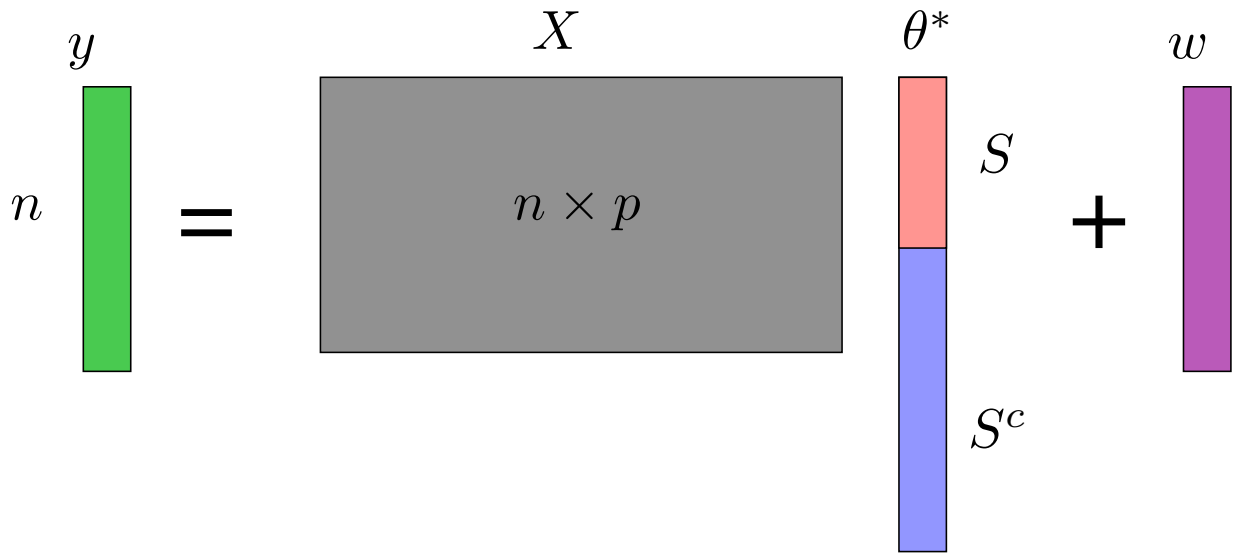
- **Estimation:** Minimize loss function plus regularization term:

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}_n(\theta; \mathcal{X}_1^n) + \lambda_n \, r(\theta) \right\}.$$

Estimate — Loss function — Regularizer

- **Analysis:** Given some loss $d(\cdot)$ (e.g., ℓ_2 norm), bound $d(\widehat{\theta}_{\lambda_n} - \theta^*)$ under high-dimensional scaling $(n, p) \to +\infty$.

Example: Sparse linear regression



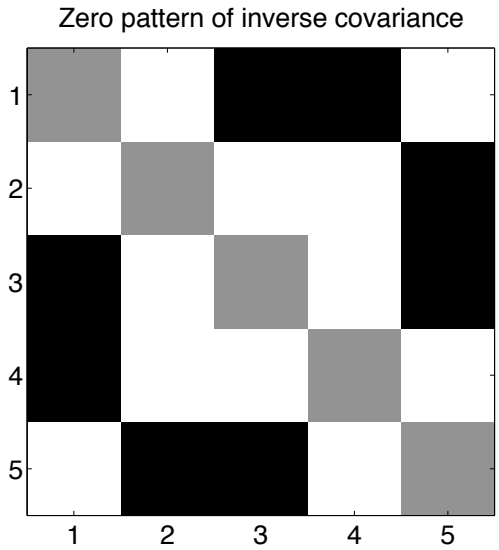
Set-up: noisy observations $y = X\theta^* + w$ with sparse θ^*

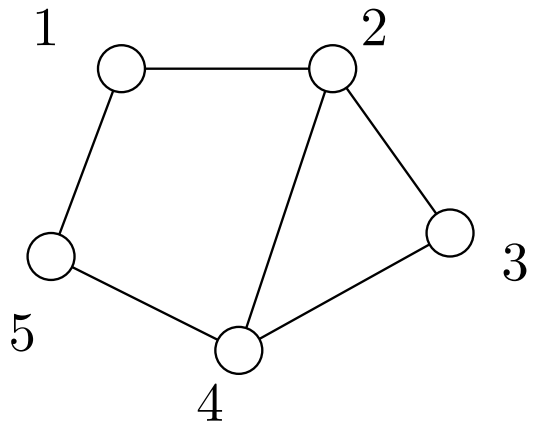
Estimator: Lasso program

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|$$

Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Huo, 2001; Tropp, 2004; Fuchs, 2004; Efron et al., 2004; Meinshausen/Buhlmann, 2005; Candes/Tao, 2005; Donoho, 2005; Zhao & Yu, 2006; Zou, 2006; Wainwright, 2006; Koltchinskii, 2007; Tsybakov et al., 2007; van de Geer, 2007; Bickel et al., 2008 ....

Example: Structured (inverse) covariance matrices





Set-up: Samples from random vector with sparse covariance Σ or sparse inverse covariance $\Theta^* \in \mathbb{R}^{p \times p}$.

Estimator:

$$\widehat{\Theta} \in \arg\min_{\Theta} \langle\langle \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T, \Theta \rangle\rangle - \log\det(\Theta) + \lambda_n \sum_{i \neq j} |\Theta_{ij}|_1$$

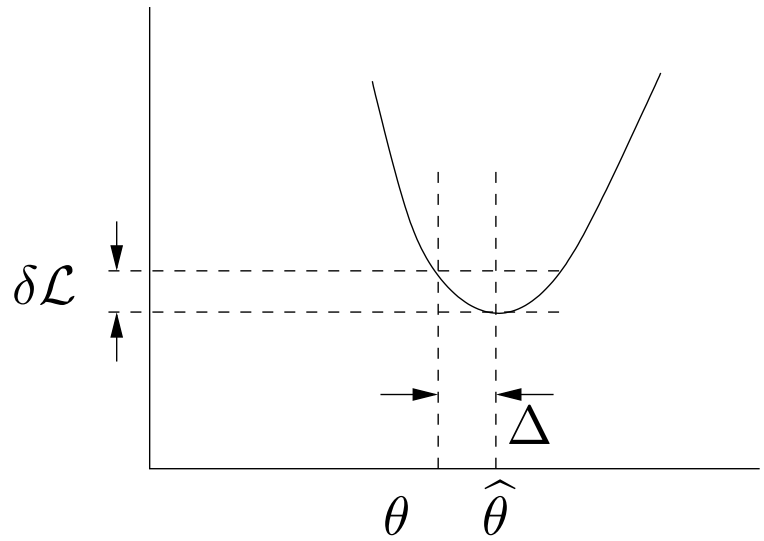
Some past work: Yuan & Lin, 2006; d'Asprémont et al., 2007; Bickel & Levina, 2007; El Karoui, 2007; Rothman et al., 2007; Zhou et al., 2007; Friedman et al., 2008; Ravikumar et al., 2008, Lam and Fan, 2009, Cai and Zhou, 2009

Unified analysis

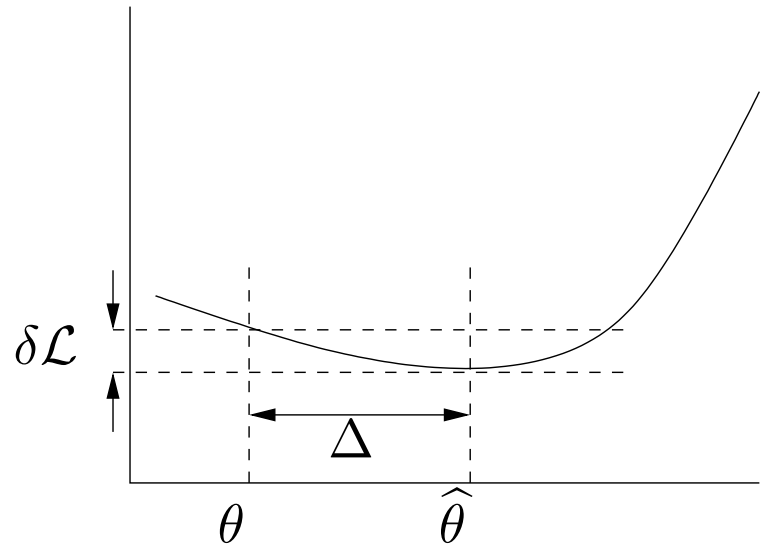
- many high-dimensional models and associated results case by case
- is there a core set of ideas that underlie these analyses?
- Two key properties
  - decomposability of regularizer r
  - restricted strong convexity of loss function $\mathcal{L}$
- Main theorem
- Some consequences

Important properties of regularizer/loss

Strong convexity of cost (curvature captured in Fisher Info when p fixed):



(a) High curvature: easy to estimate

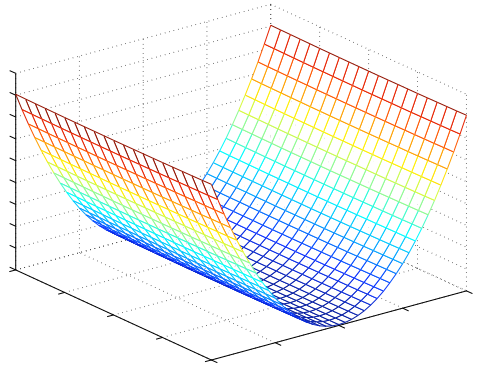


(b) Low curvature: harder

Important properties of regularizer/loss

1 Restricted strong convexity (RSC) (courtesy of high-dim):

- loss functions are often flat in many directions in high dim
- "curvature" needed only for directions $\Delta \in \mathcal{C}$ in high dim



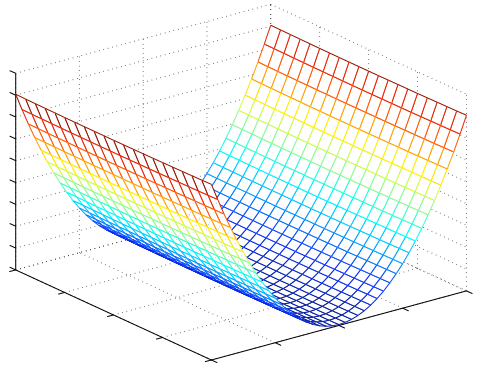
- loss function $\mathcal{L}_n(\theta) := \mathcal{L}_n(\theta; \mathcal{X}_1^n)$ satisfies

$$\underbrace{\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*)}_{\text{Excess loss}} - \underbrace{\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle}_{\substack{\text{score} \\ \text{function}}} \quad \geq \quad \gamma(\mathcal{L}) \underbrace{d^2(\Delta)}_{\substack{\text{squared} \\ \text{error}}} \qquad \text{for all } \Delta \in \mathcal{C}.$$

Important properties of regularizer/loss

1 Restricted strong convexity (RSC) (courtesy of high-dim):

- loss functions are often flat in many directions in high dim
- "curvature" needed only for directions $\Delta \in \mathcal{C}$ in high dim



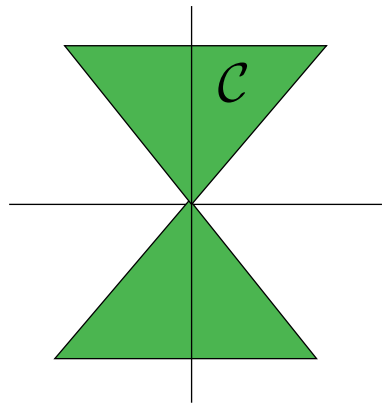
- loss function $\mathcal{L}_n(\theta) := \mathcal{L}_n(\theta; \mathcal{X}_1^n)$ satisfies

$$\underbrace{\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*)}_{\text{Excess loss}} - \underbrace{\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle}_{\substack{\text{score} \\ \text{function}}} \quad \geq \quad \gamma(\mathcal{L}) \underbrace{d^2(\Delta)}_{\substack{\text{squared} \\ \text{error}}} \qquad \text{for all } \Delta \in \mathcal{C}.$$

2 Decomposability of regularizers makes $\mathcal{C}$ small in high-dim:

- for subspace pairs $(A, B^\perp)$ where A represents model constraints:

$$r(u+v) = r(u)+r(v) \quad \text{for all } u \in A \text{ and } v \in B^\perp$$



- forces error $\Delta = \widehat{\theta}_{\lambda_n} - \theta^*$ to $\mathcal{C}$

Main theorem

Estimator

$$\widehat{\theta} \quad \in \quad \arg\min_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}_n(\theta; \mathcal{X}_1^n) + \lambda_n r(\theta) \right\}.$$

Subspace pair: $(A, B^\perp)$, where A represents model constraints.

Theorem (Negahban, Ravikumar, W., & Yu, 2009)

Say θ^* belongs to the subspace A, and regularizer decomposes across pair $(A, B^\perp)$ with $A \subseteq B$. With regularization constant chosen such that $\lambda_n \geq 2r^*(\nabla L(\theta^*; \mathcal{X}_1^n))$, then any solution $\widehat{\theta}$ satisfies

$$d(\widehat{\theta} - \theta^*) \quad \leq \quad \frac{1}{\gamma(\mathcal{L})} \left[\Psi(B)\, \lambda_n \right].$$

Quantities that control rates:

- restricted strong convexity parameter: $\gamma(\mathcal{L})$
- dual norm of regularizer: $r^*(v) := \sup_{r(u)=1} \langle v, u \rangle$.
- optimal subspace const.: $\Psi(B) = \sup_{\theta \in B \setminus \{0\}} r(\theta)/d(\theta)$

Theorem (Negahban, Ravikumar, Wainwright & Y. 2009)

Say regularizer decomposes across pair $(A, B^{\perp})$ with $A \subseteq B$, and restricted strong convexity holds for $(A, B^{\perp})$ and over $\mathcal{C}$. With regularization constant chosen such that $\lambda_n \geq 2r^*(\nabla L(\theta^*; \mathcal{X}_1^n))$, then any solution $\widehat{\theta}_{\lambda_n}$ satisfies

$$d(\widehat{\theta}_{\lambda_n} - \theta^*) \leq \underbrace{\frac{1}{\gamma(\mathcal{L})}\left[\Psi(B)\,\lambda_n\right]}_{\text{Estimation error}} + \underbrace{\sqrt{2\frac{\lambda_n}{\gamma(\mathcal{L})}\, r(\pi_{A^{\perp}}(\theta^*))}}_{\text{Approximation error}}$$

Quantities that control rates:

- restricted strong convexity parameter: $\gamma(\mathcal{L})$
- dual norm of regularizer: $r^*(v) := \sup_{r(u)=1} \langle v, u \rangle$.
- optimal subspace const.: $\Psi(B) = \sup_{\theta \in B \setminus \{0\}} r(\theta)/d(\theta)$

Summary of unified analysis

- decomposability of regularizer r leads to "small" constraint set in high-dim
- restricted strong convexity (RSC) of loss functions needed on small set

- actual rates determined by:
  - noise measured in dual function r^*
  - subspace constant Ψ in moving from r to error norm d
  - restricted strong convexity constant

- recovered some known results as corollaries:
  - sparse linear regression with Lasso
  - multivariate group Lasso
  - inverse covariance matrix estimation

- derived some new results on:
  - low-rank matrix estimation
  - weak sparsity and generalized linear models
  - other models?

Our application: word image in the New York Times

- Word Image: a small set of words describing/distinguishing a topic
- As a predictive problem:
  - Predict appearance of a query word q in a document from the document's use of other words
- Predictive model must be interpretable
  - Predictor weights must directly and simply drive label
  - No. of predictors used must be few: sparse model **approximation**
  - The faster predictors can be computed, the better
- Chosen predictor words form a set known as the Word Image for q
- Word image must be evaluated two ways:
  - Can labels (appearance indicator for q) be effectively predicted?
  - Are the chosen words meaningful w.r.t. q?

l_1 Regularized Logistic Regression (L1LR)

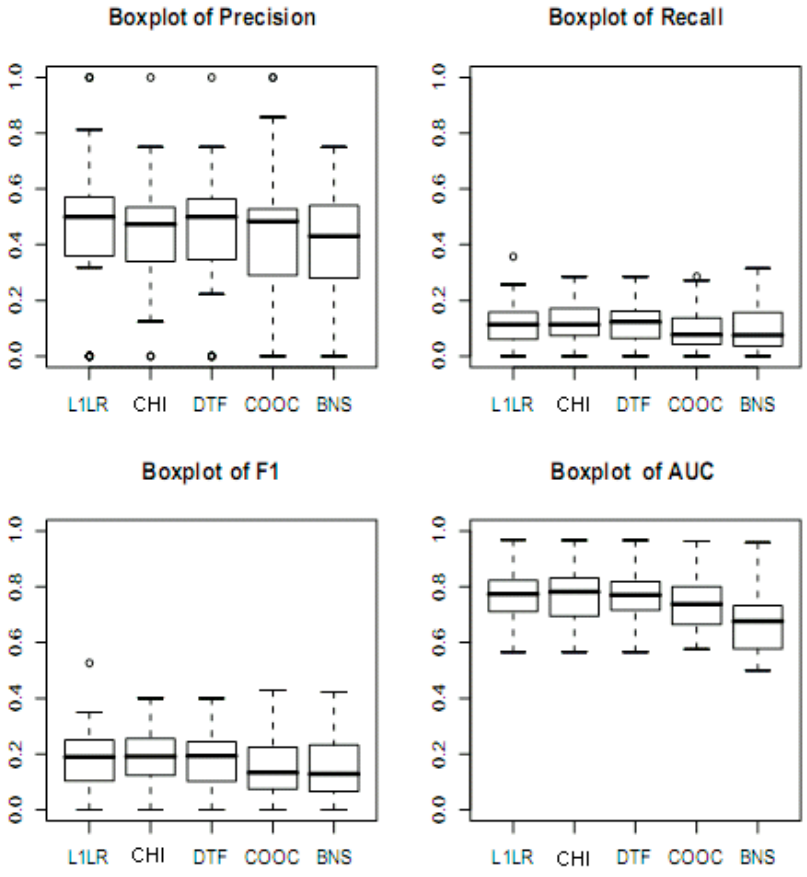
$$\mathcal{L}_{L1LR}(\beta) = -\sum_{i=1}^{m} \log\left(1 + \exp(-y_i(\beta_0 + x_i^T\beta))\right) + \lambda \sum_{j=1}^{n} |\beta_j| \qquad (1)$$

- L1LR loss function encourages fitting to the data, discourages non-zero values of β
- As $\lambda \to \infty$, $\beta_j \to 0 \;\; \forall j = 1, \ldots, n$
- By binary search, isolate value of λ which leaves ~ 15 nonzero predictors
- Greater computational complexity than previous four methods, but still solved efficiently

Selected features: "CHINA"

COOC	DTF	BNS	CHI	L1LR
year	killing	[not] recur	recurring	korea
chinas	institutions	[not] recurring	purified	united
north	view	[not] stalins	[not] nazis	north
beijing	larger	[not] kenya	marches	global
government	history	[not] marches	[not] holocaust	countries
states	outside	[not] eradicate	[not] perpetrators	russia
mr	place	[not] victims	eradicate	states
united	death	[not] goldhagen	[not] kenya	chinas
chinese	russia	[not] holocaust	stalins	beijing
said	world	[not] killing	goldhagen	chinese

Predictive Performance Results



- L1LR, CHI, and DTF do not have significant differences from each other
- L1LR, CHI, and DTF all perform significantly better than both COOC and BNS

Human Reader Survey

Please read the following paragraphs:

Paragraph 1:

After nearly two decades of independence, Moldova's citizens are still at odds over the basic question of who they are. That division boiled over last week, when a huge anti-Communist demonstration turned violent. Its participants, in their teens and 20s, say they are desperate to escape a Soviet time warp and enter Europe. But many of their elders feel more affinity with Russia, and see the protests as a plot by their western neighbor Romania to snatch away Moldova's sovereignty.

Paragraph 2:

Mr. Lukyanov pointed out that the United States and Russia approach Iran from sharply different perspectives. Russia and Iran are neighbors, and the Kremlin has for many years had positive dealings with Iran on regional issues, including unrest in Chechnya and in Central Asia.

Paragraph 3:

Last week the government's point man on the economic crisis, the deputy prime minister Igor I. Shuvalov, seemed to underline that policy. He told an economic forum in Moscow that the government would withhold support from industry and cut the budget, allowing Russia to husband reserves to support the ruble.

Q8) Which of the following word lists is the most useful summary of the above paragraphs as described in the instruction sheet? (You may select two, one, or none as desired.)

List A	List B	List C	List D
NOT pakistans	baghdad	NOT enriched	georgia
NOT boldest	iraqi	NOT officials	moscow
NOT unfolded	war	NOT slated	ukraine
NOT consult	afghanistan	NOT stockpile	russian
NOT islamabad	american	NOT lightly	putin
NOT offensive	troops	NOT vienna	russias
NOT oversees	bush	NOT accord	europe
NOT arrived	oil	NOT reactor	china
NOT capital	military	NOT geneva	united
NOT head	invasion	NOT research	gas

Continue to Q8B)...

The paragraphs on the previous page are best described as focusing on the topic(s) of _____

If at least two of the three paragraphs focus on a topic, then consider them to be focusing on the topic overall.

(a) russia
(b) iraq
(c) both of the above
(d) neither of the above

Continue to Q9A)...

(Few questions were misidentified in part B)

Human Survey Results

Scheme a	Scheme b	% a	n	p_1	$\%\Delta_{a,b}$	p_2	$-\log B$	p-value
L1LR	COOC	70	23	0.115	9	0.300	3.4	0.151
L1LR	DTF	60	22	0.503	31	0.000	8.3	**0.002**
L1LR	CHI	95	26	0.000	46	0.000	17.1	**0.000**
L1LR	BNS	94	21	0.000	50	0.000	15.8	**0.000**
COOC	DTF	62	25	0.383	11	0.257	2.3	0.327
COOC	CHI	95	27	0.000	26	0.001	17.0	**0.000**
COOC	BNS	75	24	0.077	43	0.000	10.2	**0.000**
DTF	CHI	67	26	0.302	25	0.001	8.1	**0.003**
DTF	BNS	79	26	0.057	26	0.001	9.8	**0.001**
CHI	BNS	80	28	0.109	-2	0.785	2.5	0.297

- L1LR significantly bests all but COOC
- COOC not significantly preferred over cousin DTF
- CHI and BNS roundly rejected, except between each other

Summary and future directions

- L1LR success indicates effectiveness of sophistication in ML approaches
- Traditional ML practices wouldn't yield these images – new design criteria were applied
- Scale and complexity can be easily accomodated (used BBR from Madigan group)
- Posing journalism analysis problems in a predictive framework in a way that takes advantage of these and future tools should be encouraged

- Dynamic word-imaging (was "China" portrayed differently last year?)
- Evidence for increased political polarization based on parties' platforms?
- Website up and running soon for collaborators?
- Building semantics into penalty for better word-images?