Optimal, Robust Information Fusion in Uncertain Environments

MURI Review Meeting

Integrated Fusion, Performance Prediction, and Sensor Management for Automatic Target Exploitation

Alan S. Willsky

November 3, 2008

What is needed: An expressive, flexible, and powerful framework

- **Capable of capturing uncertain and complex sensor-target relationships**
  - Among a multitude of different observables and objects being sensed
- **Capable of incorporating complex relationships about the objects being sensed**
  - Context, behavior patterns
- **Admitting scalable, distributed fusion algorithms**
- **Admitting effective approaches to learning or discovering key relationships**
- **Providing the "glue" from front-end processing to sensor management**

Our choice*: Graphical Models

- Extremely flexible and expressive framework
  - Allows the possibility of capturing (or learning) relationships among features, object parts, objects, object behavior, and context
    - E.g., constraints or relationships among parts, spatial and spatio-temporal relationships among objects, etc.
  - Natural framework to consider distributed fusion
- While we can't beat the dealer (NP-Hard is NP-Hard),
  - The flexibility and structure of graphical models provides the potential for developing scalable, approximate algorithms

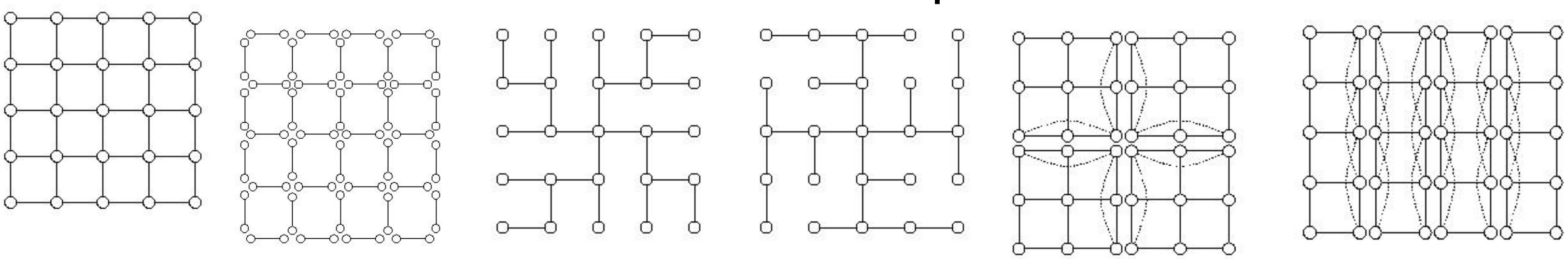
What did we say at last year? What have we done recently? - I

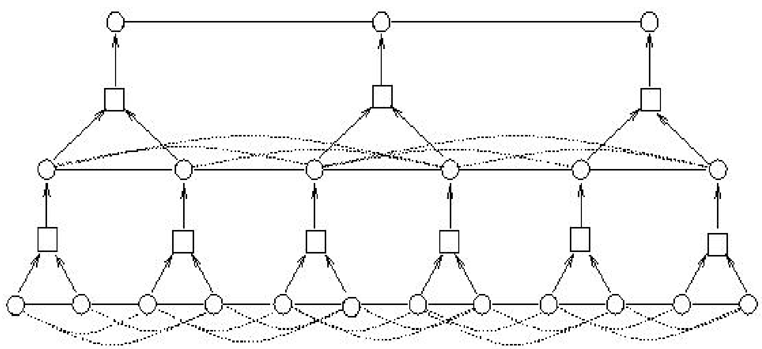
- Scalable, broadly applicable inference algorithms
  - Build on the foundation we have
  - Provide performance bounds/guarantees

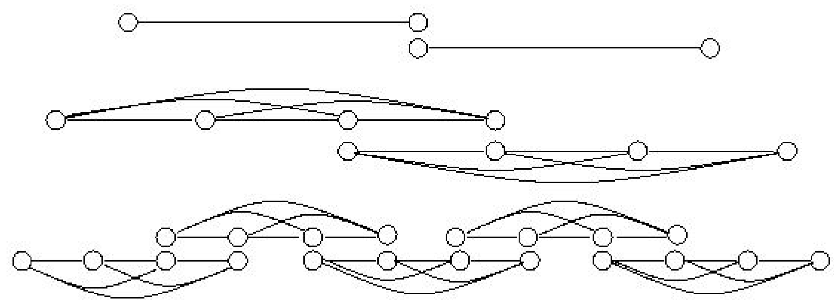
- Some of the accomplishments this year
  - ***Lagrangian relaxation methods for tractable inference***
  - Multiresolution models with "multipole" structure, allowing near optimal, very efficient inference

Lagrangian Relaxation Methods for Optimization/Estimation in Graphical Models

- Break an intractable graph into tractable pieces
  - There will be overlaps (nodes, edges) in these pieces
  - There may even be additional edges and maybe even some additional nodes in some of these pieces





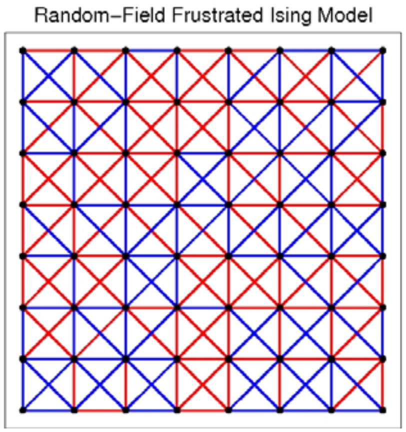


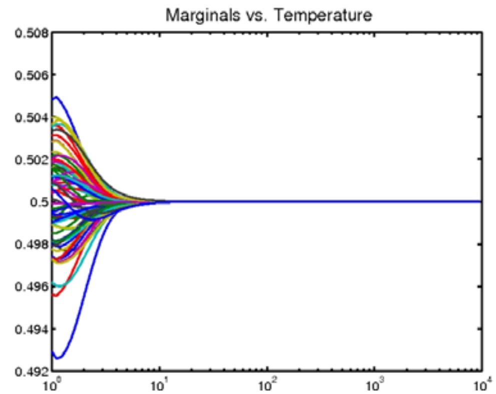
Constrained MAP estimation on the set of tractable subgraphs

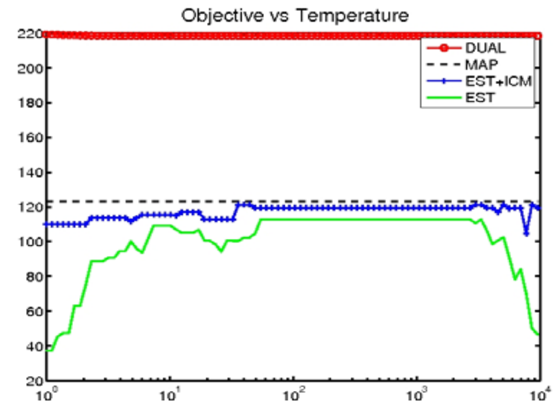
- Define graphical models on these subgraphs so that when replicated node/edge values agree we match the original graphical model
- Solve MAP with these agreement constraints
- Duality: Adjoin constraints with Lagrange multipliers, optimize w.r.t. replicated subgraphs and then optimize w.r.t. Lagrange multipliers
  - Algorithms to do this have appealing structure, alternating between tractable inference on the individual subgraphs, and moving toward or forcing local consistency
  - Generalizes previous work on "tree-agreement," although new algorithms using smooth (log-sum-exp) approximation of max
    - Leads to sequence of successively "cooled" approximations
    - Each involves iterative scaling methods that are adaptations of methods used in the ***learning*** of graphical models
  - There may or may not be a duality gap
  - If there is, the solution generated isn't feasible for the original problem (fractional assignments)
  - Can often identify the inconsistencies and overcome them through the inclusion of additional tractable subgraphs

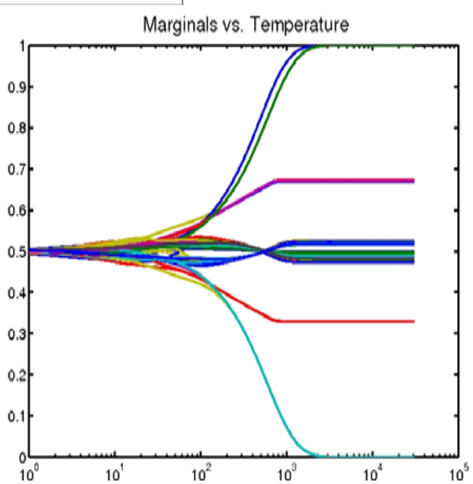
Example – Frustrated Ising - I

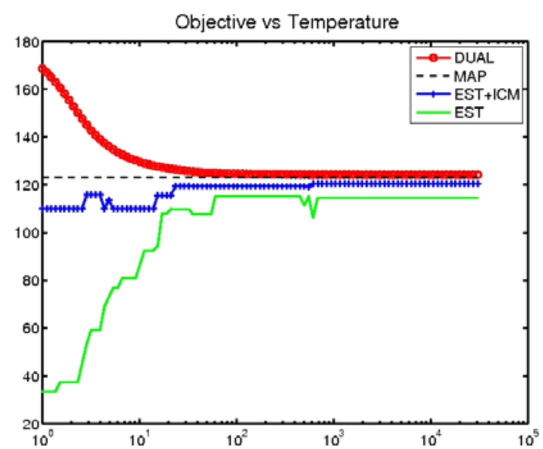
Models of this and closely related types arise in multi-target data assocation



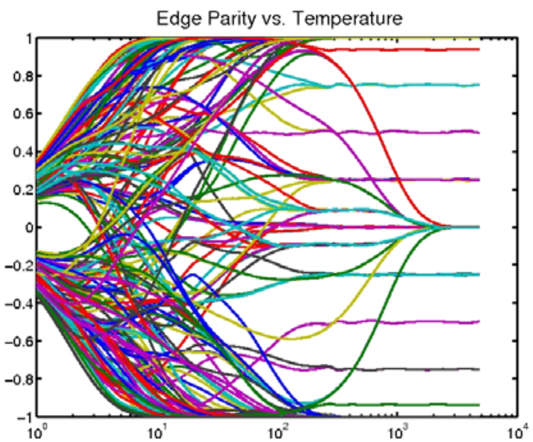


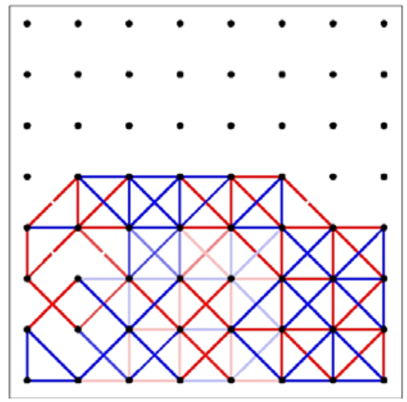


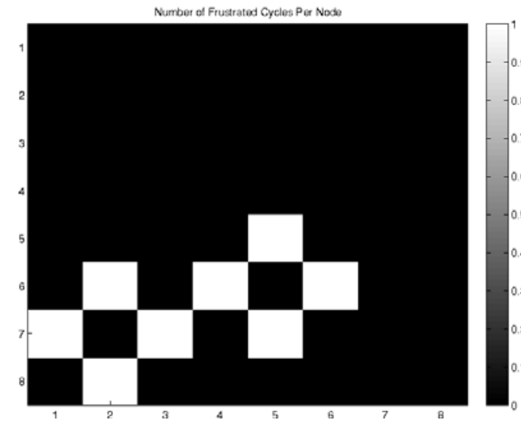


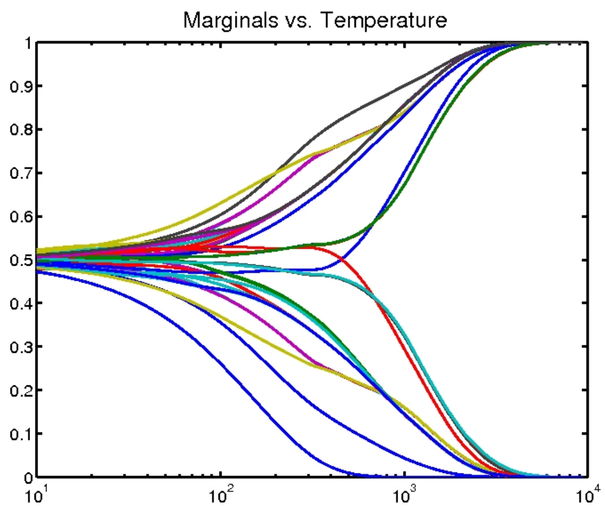


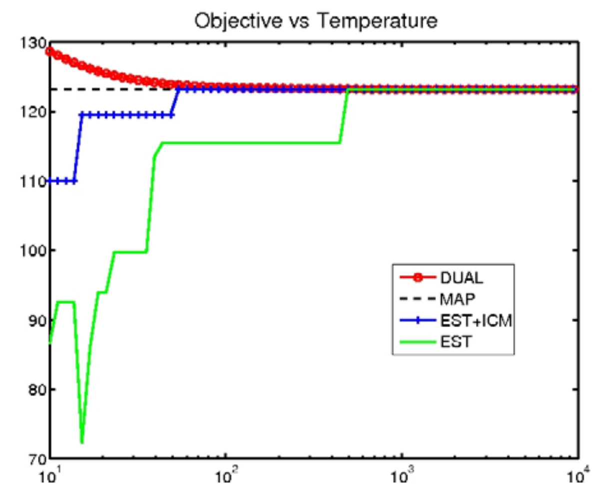
Example – Frustrated Ising - II



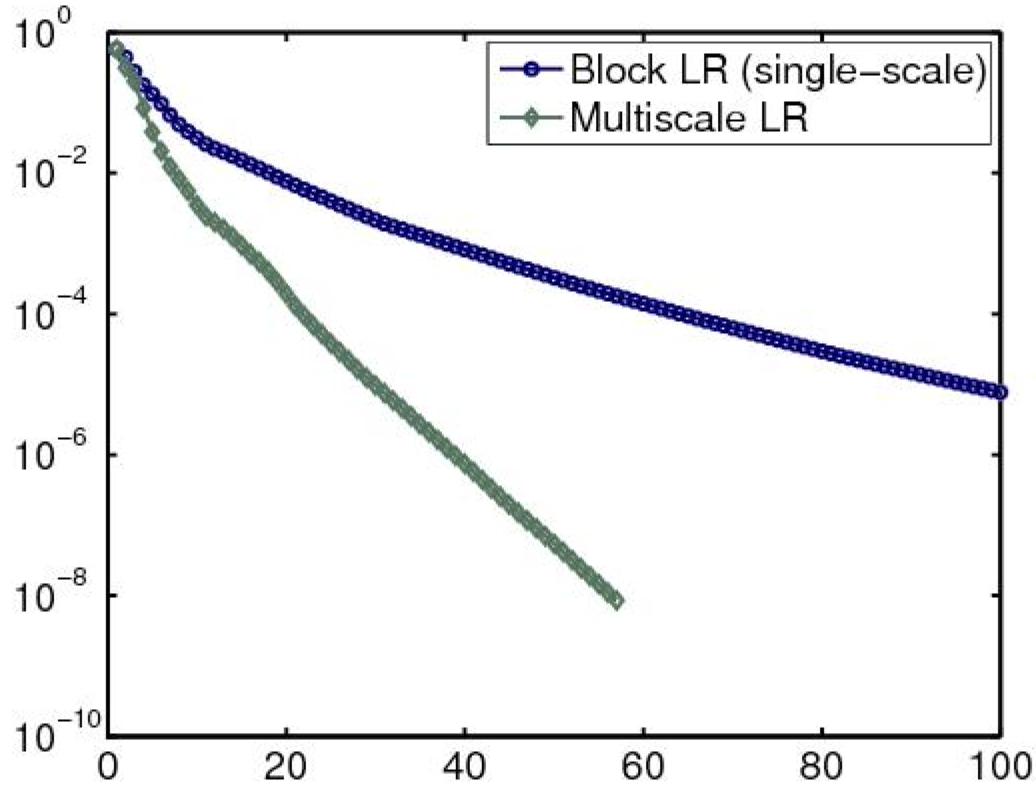








Example – Multiscale for 2-D MRFs



What did we say last year?
What have we recently? - II

- Graphical-model-based methods for sensor fusion for tracking, and identification
  - Graphical models to learn motion patterns and behavior (preliminary)
  - Graphical models to capture relationships among features-parts-objects
- Some of the accomplishments this year
  - ***Hierarchical Dirichlet Processes to learn motion patterns and behavior – much more***
  - New graphical model-based algorithms for multi-target, multi-sensor tracking

HDPs for Learning/tracking motion patterns (and other things!)

- Objective – learn motion patterns of targets of interest
  - Having such models can assist tracking algorithms
  - Detecting such coherent behavior may be useful for higher-level activity analysis
- Last year
  - Learning additive jump-linear system models
- This year
  - Learning switching autoregressive models of behavior and detecting such changes
  - Extracting and de-mixing structure in complex signals

Reminder from last year: Jump-mean processes

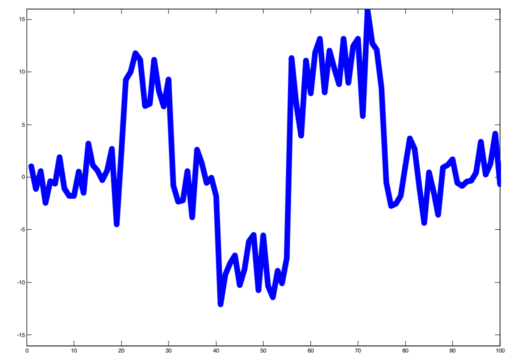
- Markov jump-mean process
  - System “jumps” between finite set of acceleration means
  - Hybrid continuous-discrete state:

$$\bar{x}_t = \begin{bmatrix} x_t \\ z_t \end{bmatrix}$$

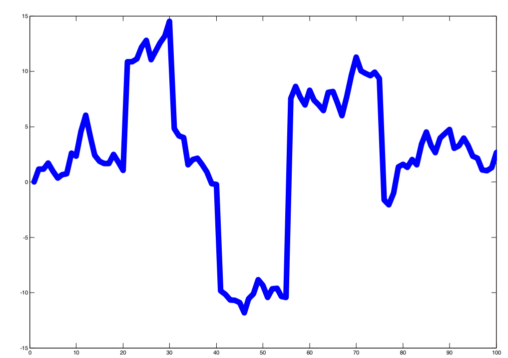
  - Dynamics described by:

$$\begin{aligned} x_t &= Ax_{t-1} + Bu_t(z_t) + v_t \\ &= Ax_{t-1} + \tilde{u}_t(z_t) \\ u_t|z_t &\sim \mathcal{N}(\mu_{z_t}, \Sigma_{z_t}) \end{aligned}$$

  - System is non-linear due to mode uncertainty



Constant Velocity (CV)



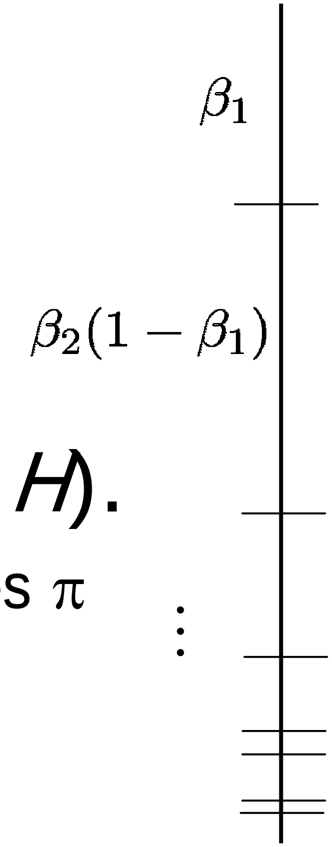
Constant Acceleration (CA)

Some questions

- How many possible maneuver modes are there?
- What are their individual statistics?
- What is the probabilistic structure of transitions among these modes?
- Can we learn these
  - Without placing an *a priori* constraint on the number of modes
  - Without having *everything* declared to be a different "mode"
- The key to doing this: Dirichlet processes

Dirichlet Process via Stick Breaking

$$\beta_i \sim \mathrm{Beta}(1,\alpha) \quad i = 1,2,\ldots$$
$$\pi_1 = \beta_1$$
$$\pi_i = \beta_i \prod_{j=1}^{i-1}(1-\beta_j) \quad i = 2,3,\ldots$$



- Corresponds to a draw from DP(α, *H*).
  - Mixture components drawn with probabilities π and with parameters drawn from *H*

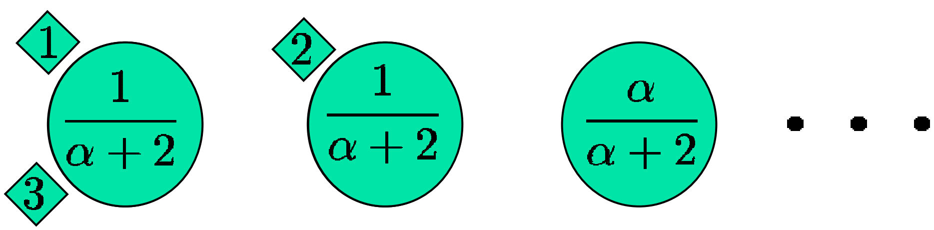
Chinese Restaurant Process

- Predictive distribution:

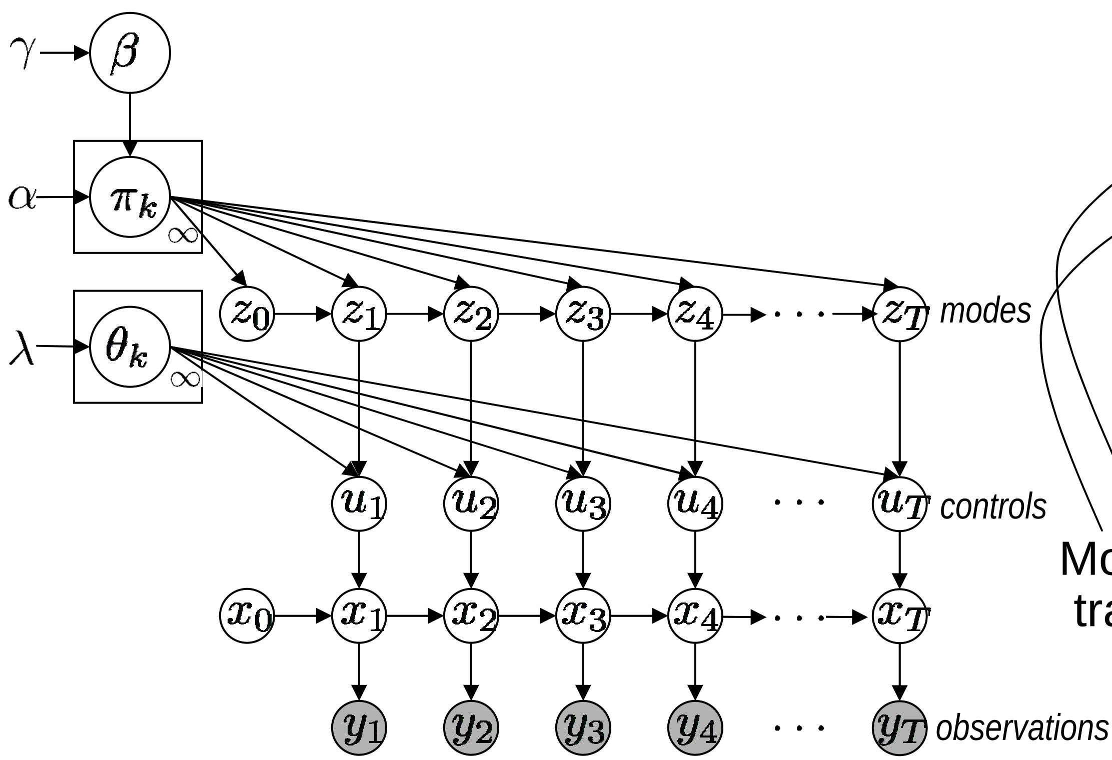
$$p(z_t = z | z_{\backslash t}, \alpha, H) = \frac{\alpha}{\alpha + T}\delta(z, K+1) + \frac{1}{\alpha + T}\sum_{k=1}^{K} T_k \delta(z, k)$$

Number of current assignments to mode k

- Chinese restaurant process:



Graphical Model of HDP-HMM-KF



$$
\begin{aligned}
\beta &\sim \mathrm{Stick}(\alpha) \\
\pi_k &\sim \mathrm{DP}(\alpha, \beta) \\
\theta_k &\sim \mathrm{NIW}(\lambda)
\end{aligned}
$$

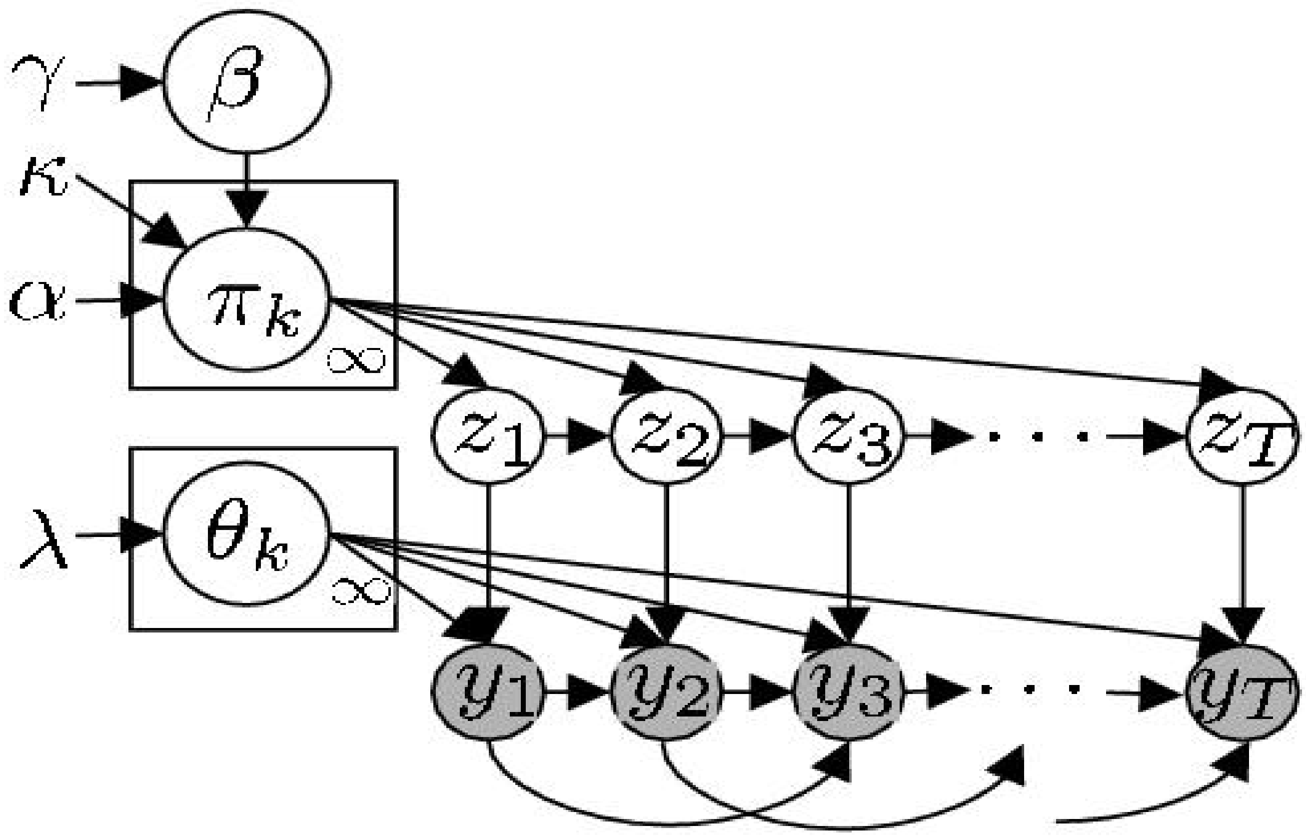
"Average" transition density which encourages states to transition back to a finite subset of the infinite state space

Mode-specific transition density

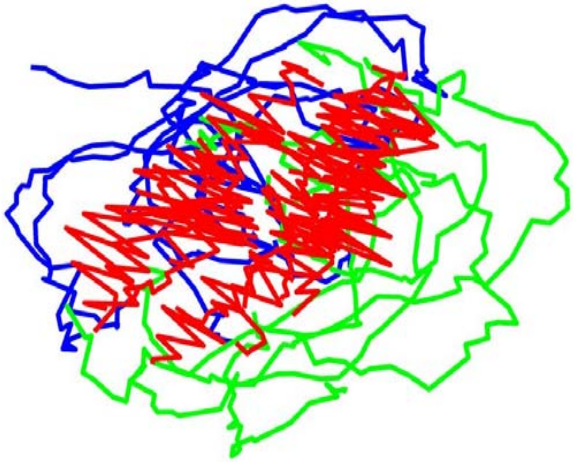
Learning and using HDP-based models

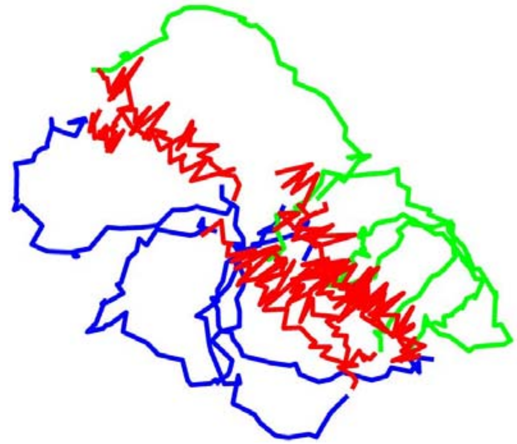
- Learning models from training data
  - Gibbs sampling-based methods
  - Exploit conjugate priors to marginalize out intermediate variables
  - Computations involve both forward filtering and reverse smoothing computations on target tracks

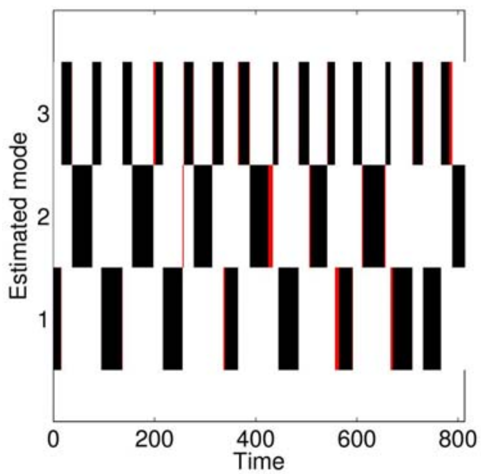
New models/results this year – I: Learning switching LDS and AR models

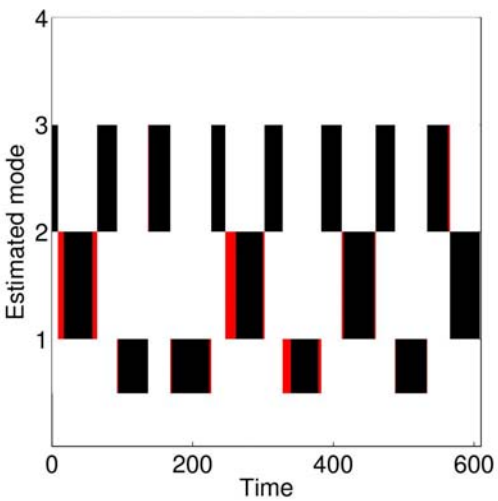


Learning switching AR models – II: Behavior extraction of bee dances





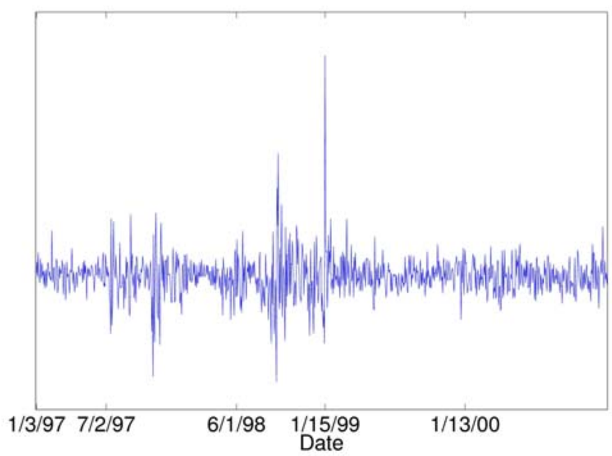


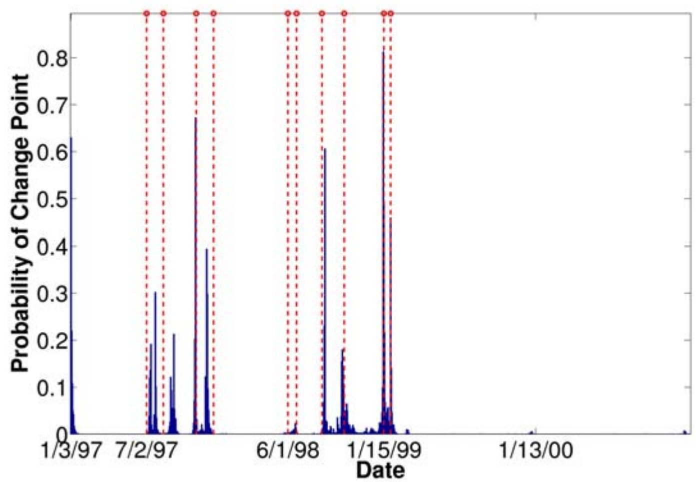


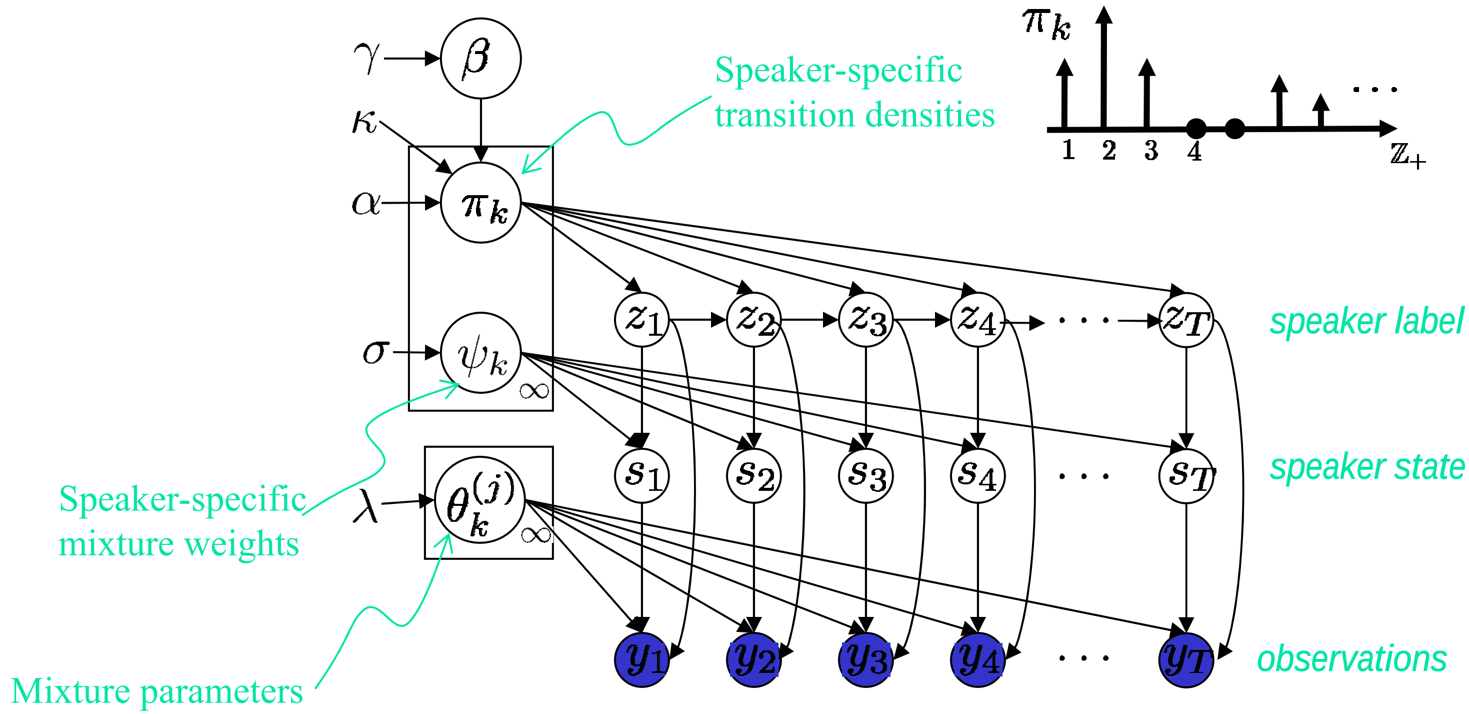
Learning switching AR models – III: Extracting major world events from Sao Paulo stock data

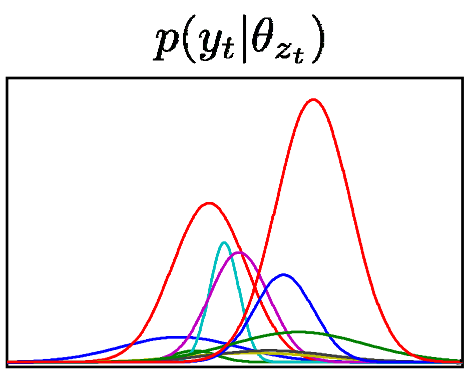
Using the ***same*** HDP model and parameters as for bee dances

- Identifies events and mode changes in volatility with comparable accuracy to that achieved by in-detail economic analysis
- Identifies three distinct modes of behavior (economic analysis did not use or provide this level of detail)

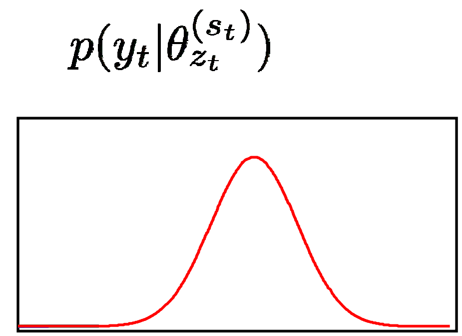








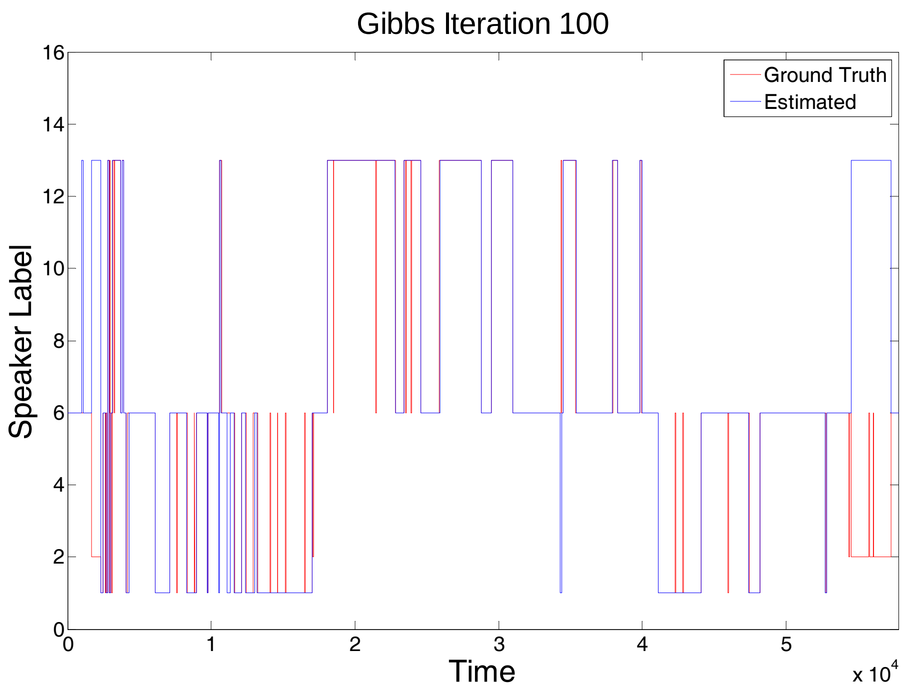
Speaker-specific emission distribution – infinite Gaussian mixture

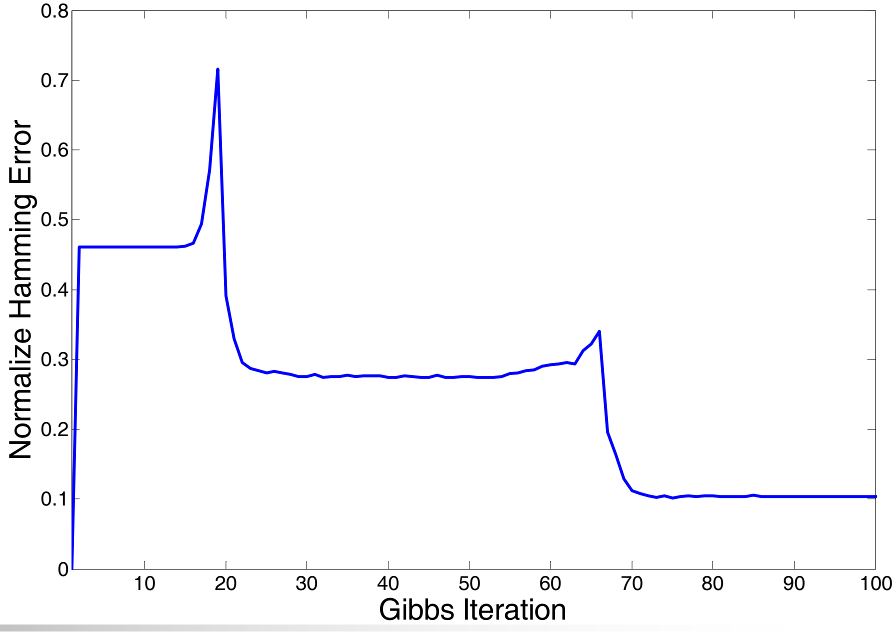


Emission distribution conditioned on speaker state

New this year – II: HMM-like model for determining the number of speakers, characterizing each, and segmenting an audio signal ***without any training***

Performance: Surprisingly good without ***any*** training





What did we say last year? What have we done recently? - III

- Learning model structure
  - Exploiting and extending advances in learning (e.g., information-theoretic and manifold-learning methods) to build robust models for fusion
  - Direct ties to integrating signal processing products ***and*** to directing both signal processing and search
- Some of the accomplishments this year
  - Learning graphical models directly for discrimination (much more than last year – some in John Fisher's talk)
  - ***Learning from experts: Combining dimensionality reduction and level set methods***
  - Combining manifold learning and graphical modeling

Learning graphical models directly for discrimination - I

- If the ultimate objective of model construction is to use models for discrimination, why don't we ***design*** these models to optimize discrimination performance?
  - If there is an abundance of data, this really doesn't matter
  - However, for high-dimensional data and relatively sparse sets of data, there can be a substantial difference between learning a model for its own sake and learning one to optimize discrimination
  - The latter objective focuses more on ***saliency***
  - In addition, we can try to do this in a manner that makes discrimination as easy as possible

Learning graphical models directly for discrimination - II

- Learning ***generative*** tree models from data
  - Criterion: Minimizing KL Divergence, $D(p_e||p)$ between tree model, p, and empirical distribution, p_e
  - Chow-Liu: Reduces to a max-weight spanning tree problem
    - Efficient solution methods exist, including Kruskal's (greedy) algorithm
- Learning tree models to discriminate two classes
  - Criterion: Minimize expected divergence between tree models (averaging over empirical distributions; extension of J-divergence)
  - Can be reduced to ***two*** spanning tree problems, one for each model

$$\min_{\widehat{p}} D(p_e||\widehat{p}) - D(q_e||\widehat{p}) \qquad \min_{\widehat{q}} D(q_e||\widehat{q}) - D(p_e||\widehat{q})$$

- ***Extend this to discriminative forests***
  - ***Greedy algorithm: At each stage, either***
    - ***Add edge to one forest, to the other, to both, or stop***
    - ***Puts maximal weight on salient relationships***

J - Divergence

- Let *p, q* denote empirical distributions.
- Let p_A, q_B denote ***information projections*** of these empirical distributions to graphs $\mathcal{G}_A$ and $\mathcal{G}_B$
  - Projections match marginals associated with vertices and edges of the graphs
- J-Divergence:

$$\hat{J}(p,q) = \int (p-q)\log\left(\frac{p_A}{q_B}\right)$$

J – Divergence for Tree Models

- If $\mathcal{G}_A$ and $\mathcal{G}_B$ are trees

$$\hat{J}(p,q) = \sum_{s \in \mathcal{V}} J(p_s, q_s) + \sum_{(s,t) \in \mathcal{E}_p \cup \mathcal{E}_q} w_{st}$$

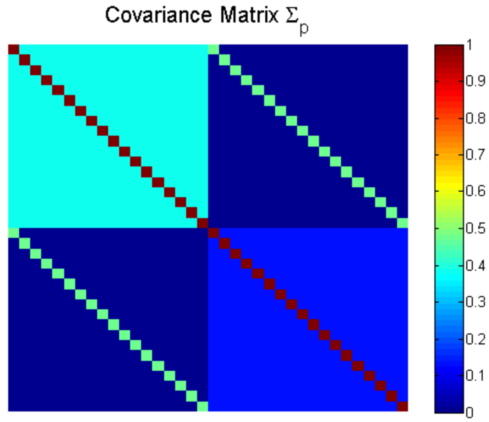
- where

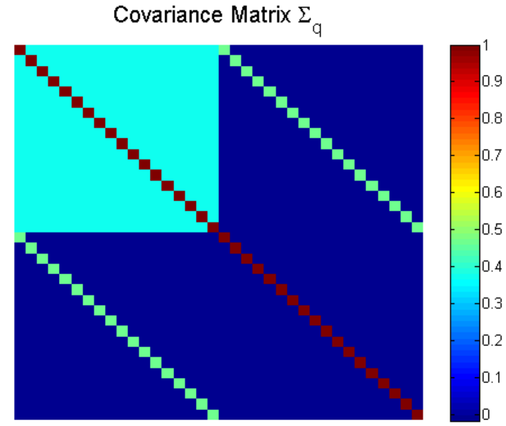
$$w_{st} = \begin{cases} I_p(x_s; x_t) - I_q(x_s; x_t) \\ \quad + D\left(q_{s,t} || p_{s,t}\right) - D\left(q_s q_t || p_s p_t\right) & (s,t) \in \mathcal{E}_p \setminus \mathcal{E}_{pq} \\ \\ I_q(x_s; x_t) - I_p(x_s; x_t) \\ \quad + D\left(p_{s,t} || q_{s,t}\right) - D\left(p_s p_t || q_s q_t\right) & (s,t) \in \mathcal{E}_q \setminus \mathcal{E}_{pq} \\ \\ J(p_{st}, q_{st}) - J\left(p_s p_t, q_s q_t\right) & (s,t) \in \mathcal{E}_{pq} \end{cases}$$

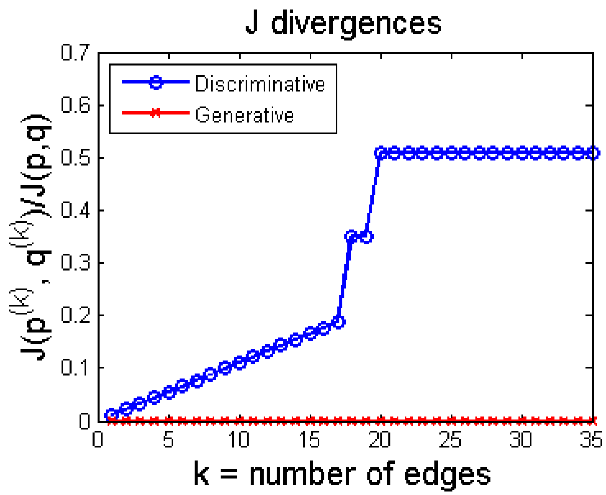
Optimal (but greedy) algorithm

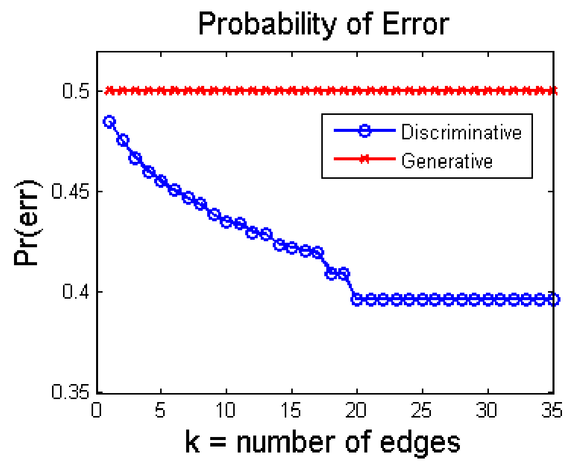
- If at any stage in the construction of $\mathcal{G}_A$ and $\mathcal{G}_B$ all remaining w_{st} are negative, **STOP**
- Otherwise: at any stage
  - Edges already included in one or both trees are no longer available
  - For other edges, addition to one or both trees may no longer be possible (as loops will be formed)
- For those edges that remain (and the set of possibilities still active – i.e., inclusion in one or both trees still feasible)
  - Choose the largest of the weights and associated edges (in one or both trees)

Emphasizing saliency: A simple example







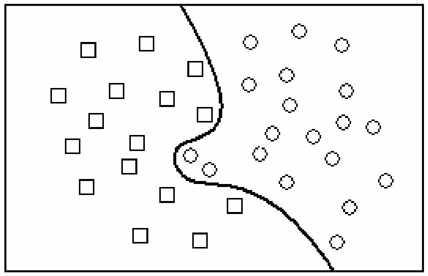


Learning from experts: Combining Dimensionality Reduction and Curve Evolution

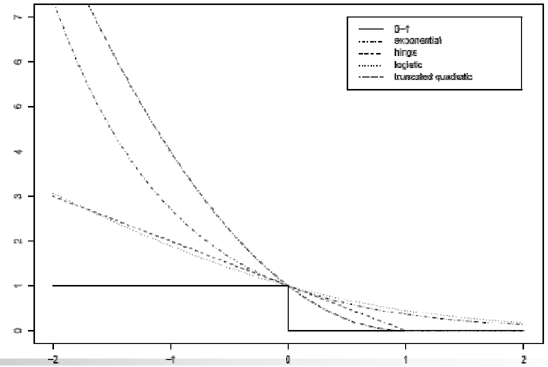
- How do we learn from expert analysts
  - Probably can't explain what they are doing in terms that directly translate into statistical problem formulations
    - Critical features
    - Criteria (are they really Bayesians?)
  - Need help because of huge data overload
- Can we learn from examples of analyses
  - Identify lower dimension that contains "actionable statistics"
  - Determine decision regions

The basic idea of learning regions

- Hypothesis testing partitions feature space



- We don't just want to separate classes
  - We'd like to get as much "margin" as possible
- Use a margin-based loss function on the **signed distance function** of the boundary curve



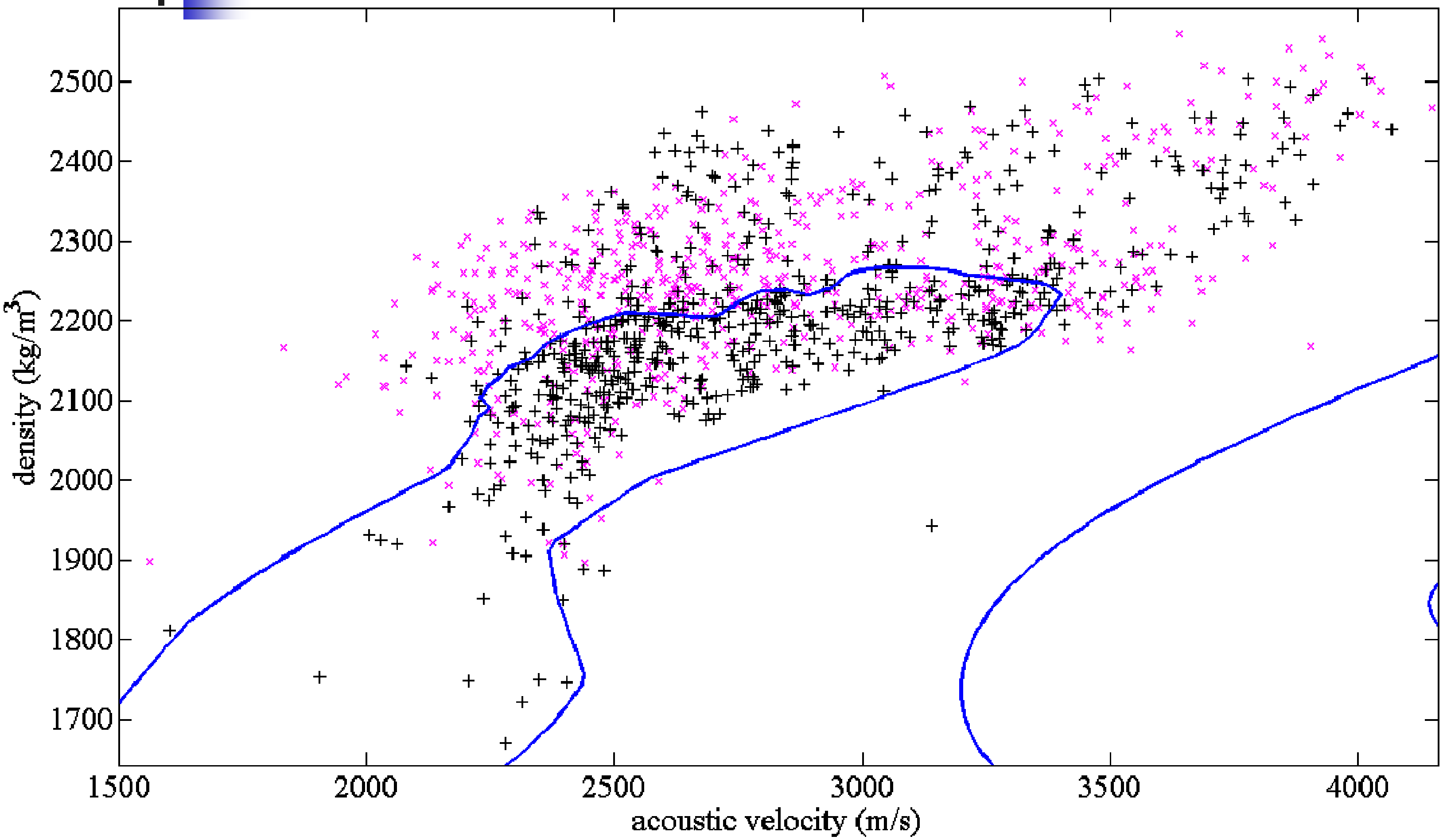
Curve Evolution Approach to Classification

- Signed distance function $\varphi(\mathbf{x})$
- Margin-based loss function $\mathrm{L}(z)$
- Training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$
  - $\mathbf{x}_n$ real-valued features in D dimensional feature space
  - y_n binary labels, either $+1$ or -1
- Minimize energy functional with respect to $\varphi(\cdot)$

$$\mathrm{E}(\varphi) = \sum_{n=1}^{N} \mathrm{L}\left(y_n \varphi(\mathbf{x}_n)\right)$$

  - Use curve evolution techniques

Example



2500
2400
2300
2200
2100
2000
1900
1800
1700
density (kg/m^3)
1500
2000
2500
3000
3500
4000
acoustic velocity (m/s)

Add in dimensionality reduction

- $D \times d$ matrix **A** lying on Stiefel manifold ($d < D$)
- Linear dimensionality reduction by $\mathbf{A}^T\mathbf{x}$

$$\mathrm{E}(\mathbf{A},\varphi)=\sum_{n=1}^{N}\mathrm{L}\left(y_n \varphi\left(\mathbf{A}^T\mathbf{x}_n\right)\right)$$

- Nonlinear mapping χ **=** A(**x**)
  - χ is d-dimensional
- Nonlinear dimensionality reduction plus manifold learning

$$\sum_{n=1}^{N}\mathrm{L}\left(y_n \varphi(\boldsymbol{\chi}_n)\right)+\lambda\,\mathrm{E}_{\text{manifold}}\left(\boldsymbol{\chi}_1,\ldots,\boldsymbol{\chi}_N;\mathbf{x}_1,\ldots,\mathbf{x}_N\right)$$

What else is there and what's next -I

- New graphical model-based algorithms for multi-target, multi-sensor tracking
  - Potential for significant savings in complexity
  - Allows seamless handling of late data and track-stitching over longer gaps
- Multipole models and efficient algorithms
- Complexity reduction: blending manifold learning and graphical modeling

What else is there and what's next -II

- Performance Evaluation/Prediction/Guarantees
  - Guarantees/Learning Rates for Dimensionality Reduction/Curve Evolution for Decision Boundaries
  - Guarantees and Error Exponents for Learning of Discriminative Graphical Models (see John Fisher's talk)
  - Guarantees/Learning Rates for HDP-Based Behavioral Learning
  - Complexity Assessment
    - For matching/data association (e.g., how complex are the subgraphs that need to be included to find the best associations)
    - For tracking (e.g., how many "particles" are needed for accurate tracking/data association)
  - Harder questions: How good are the optimal answers
    - Just because it's optimal doesn't mean it's good

Some (partial) answers to key questions - I

- Synergy
  - The whole being more than the sum of the parts
    - E.g., results/methods that would not have even existed without the collaboration of the MURI
  - Learning of discriminative graphical models from low-level features
    - Cuts across low-level SP, learning, graphical models, and resource management
  - Blending of complementary approaches to complexity reduction/focusing of information
    - Manifold learning meets graphical models
  - Blending of learning, discrimination, and curve evolution
    - Cuts across low-level SP, feature extraction, learning, and extraction of geometry
  - Graphical models as a unifying framework for fusion across all levels
    - Incorporating different levels of abstraction from features to objects to tracks to behaviors

Some (partial) answers to key questions - II

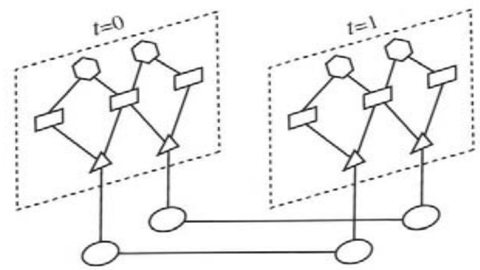
- Addressing higher levels of fusion
  - One of the major objectives of using graphical models is to make that a natural part of the formulation
  - See previous slide on synergy for some examples
  - The work presented today on automatic extraction of dynamic behavior patterns addresses this directly
    - Other work (with John Fisher) also
- Transitions/transition avenues
  - The Lagrangian Relaxation method presented today has led directly to a module in BAE-AIT's ATIF (All-Source Track and ID Fusion) System
    - ATIF originally developed under a DARPA program run by AFRL and is now an emerging system of record and widely employed multi-source fusion system
  - Discussions ongoing with BAE-AIT on our new approach to multi-target tracking and its potential for next generation tracking capabilties
    - E.g., for applications in which other "tracking services" beyond targeting are needed

Some (partial) answers to key questions - III

- Thoughts on "End States"
  - More than a set of research results and "point" transitions
    - The intention is to ***move the dial***
  - Foundation for new (very likely ***radically*** new) and integrated methods for very hard fusion, surveillance, and intelligence tasks
    - Approaches that could not possibly be developed under the constraints of 6-2 or higher funding because of programmatic constraints – but that are dearly needed
    - Thus, while we do and will continue to have point transitions, the most profound impact of our MURI will be approaches that have major impact down the road
    - Plus the new generation of young engineers trained under this program
  - Some examples
    - New methods for building graphical models that are both tractable and useful for crucial militarily relevant problems of fusion across all levels
    - New graphical models for tracking and extraction of salient behavior
    - Learning from experts: learning discriminative models and extracting saliency from complex, high-dimensional data
      - What is it that that image analyst sees in those data?

Multi-target, multi-sensor tracking

- A new graphical model, making explicit data associations within each frame and stitching across time using target dynamics (modeled here as independent).



- This is a complete representation of the overall probabilistic model
  - The question is: What informational ***queries*** do we want to make
- E.g., to compute marginals (rather than most likely MHT tracks)
  - Exponential explosion is embedded in the ***messages***
  - The key: rather than pruning hypotheses across time, we ***approximate*** messages from one time to another, ***both forward and backward in time***

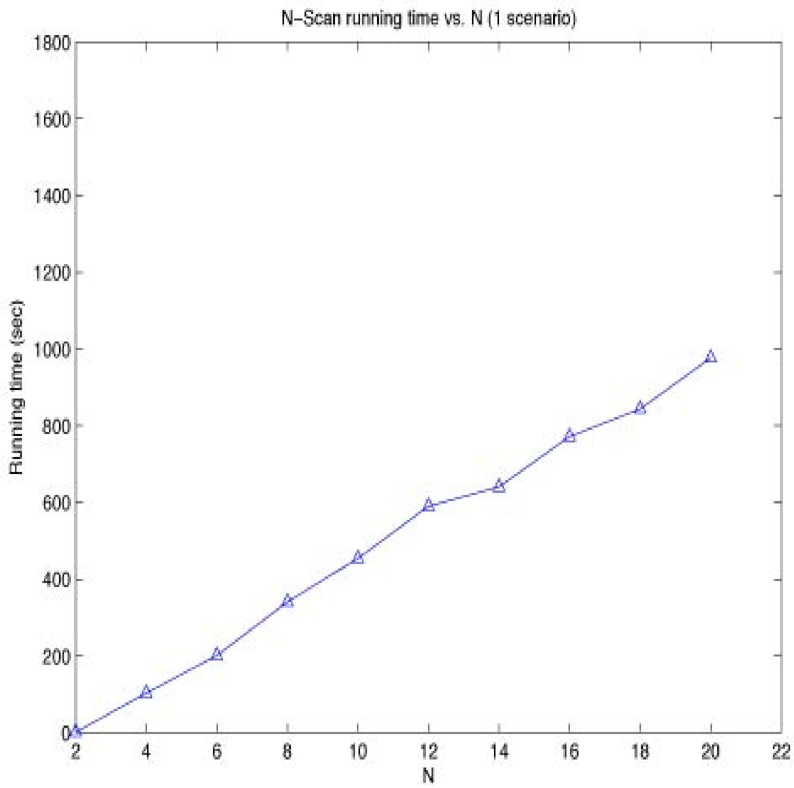
Key points

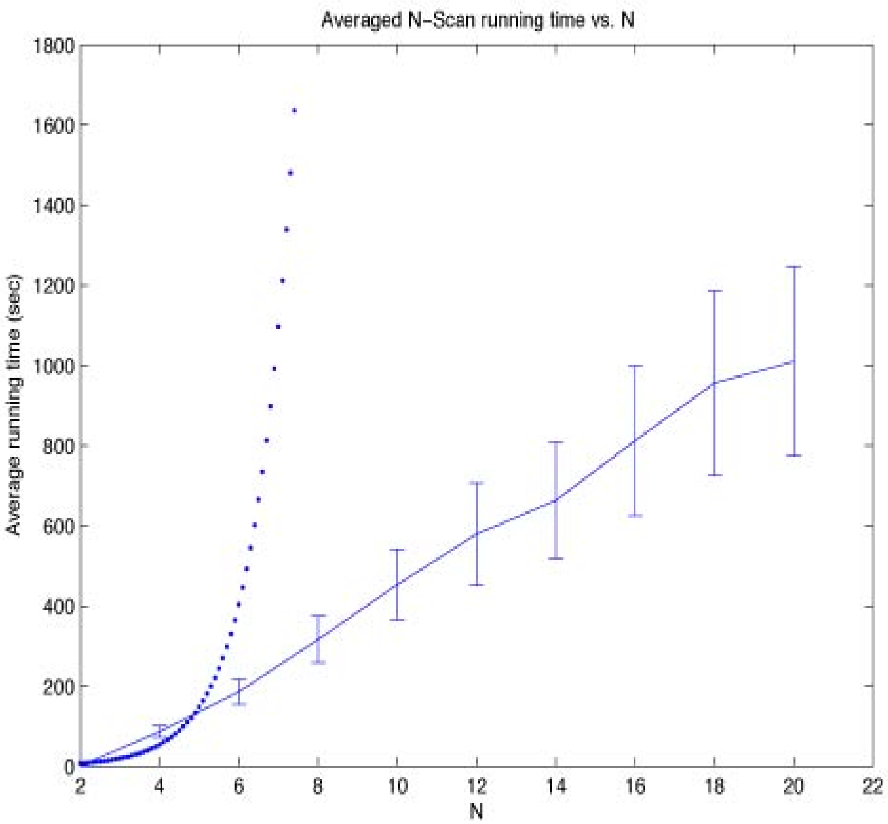
- Very different than other tracking methods
  - Rather than bringing old data association hypotheses forward toward new data, ***we bring the data back to the older association hypotheses***
  - Messages from one time frame back in time to another are important primarily ***to resolve association hypotheses***
- Method for approximating frame-to-frame messages
  - Basically a problem in mixture density approximation
    - "Particles" represent track hypotheses propagated backward or forward in time ***or*** aggregates of such hypotheses

Previously ***completely*** (and now only mostly) unsubstantiated claims

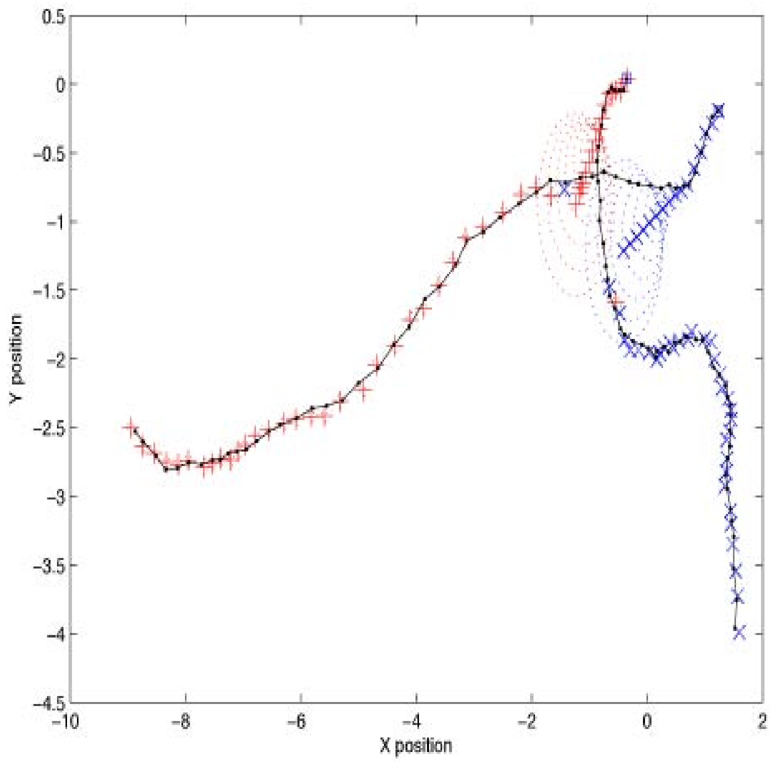
- The structure of this graphical representation makes it seamless to incorporate out-of-time or latent data
  - As long as the data are within the time window over which hypotheses are maintained
- As opposed to exponential growth in hypotheses for state-of-the-art algorithms
  - Our method offers the possibility of ***linear*** growth with time window
  - If we can control the number of particles in message generation without compromising accuracy
  - Note that we are approximating messages, ***not*** pruning hypotheses
- If true, we not only get seamless incorporation of latent data
  - But also greatly enhanced capabilities for ***track-stitching*** (e.g., when distinguishing data or human intel provides key information)

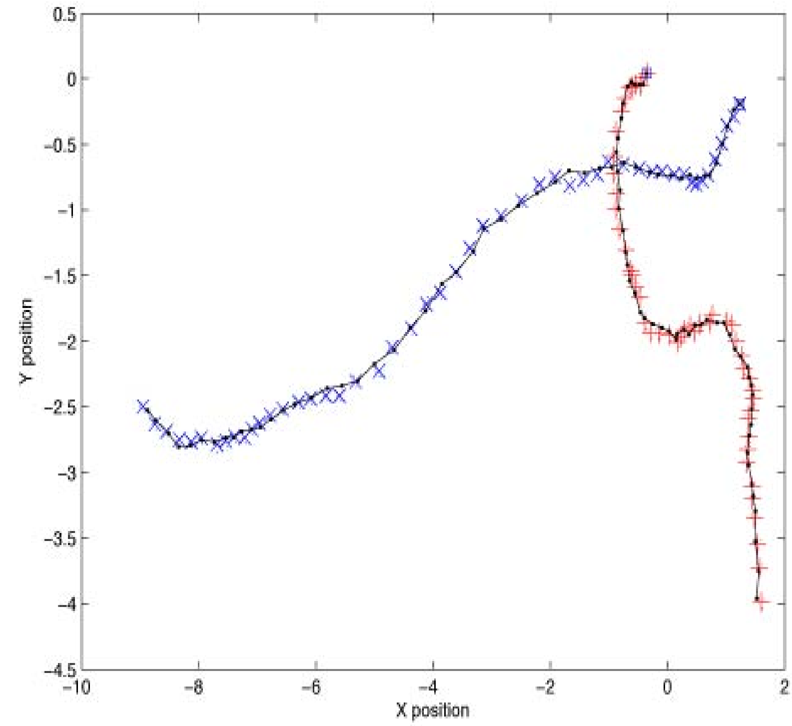
Linearity of complexity





Incorporating latent data





Track Stitching