Bayesian evidence kernel for linear dynamical systems

Yingbo Song

CS 6998

Spring 2007

- General theme of project:
  - Semi-parametric learning for time series datasets with kernels between latent graphical models.

- PRESENTATION OUTLINE

- Part I: hidden Markov models *(goals: application oriented)*
  - Modeling human travel patterns with HMMs
  - Visualizing similar behavior patterns by embedding HMMs

- Part II: Linear dynamical systems *(goals: deriving new algorithms)*
  - Quick overview of linear dynamical systems, Bayesian evidence
  - Bayesian evidence kernel for linear dynamical systems
  - Derivations

Overview: kernels for graphical models

- MOTIVATION
  - Kernel based learning is well understood. Many algorithms exist but mostly for fixed sized vector data.

  - Time series data: A sequence of fixed sized samples drawn from some stochastic process that has some sort of internal state transitions.

  - Models:
    - HMM – Markov property, multiple discrete hidden states models
    - LDS – Markov property, single continuous hidden state model

  - We can extend kernel learning into the time series domain. Some applications include:
    - spectral clustering of graphical models (HMMs, LDS, Bayes nets…)
    - embedding HMMs, LDS, etc …
    - support HMM machine, …
    - and so on…

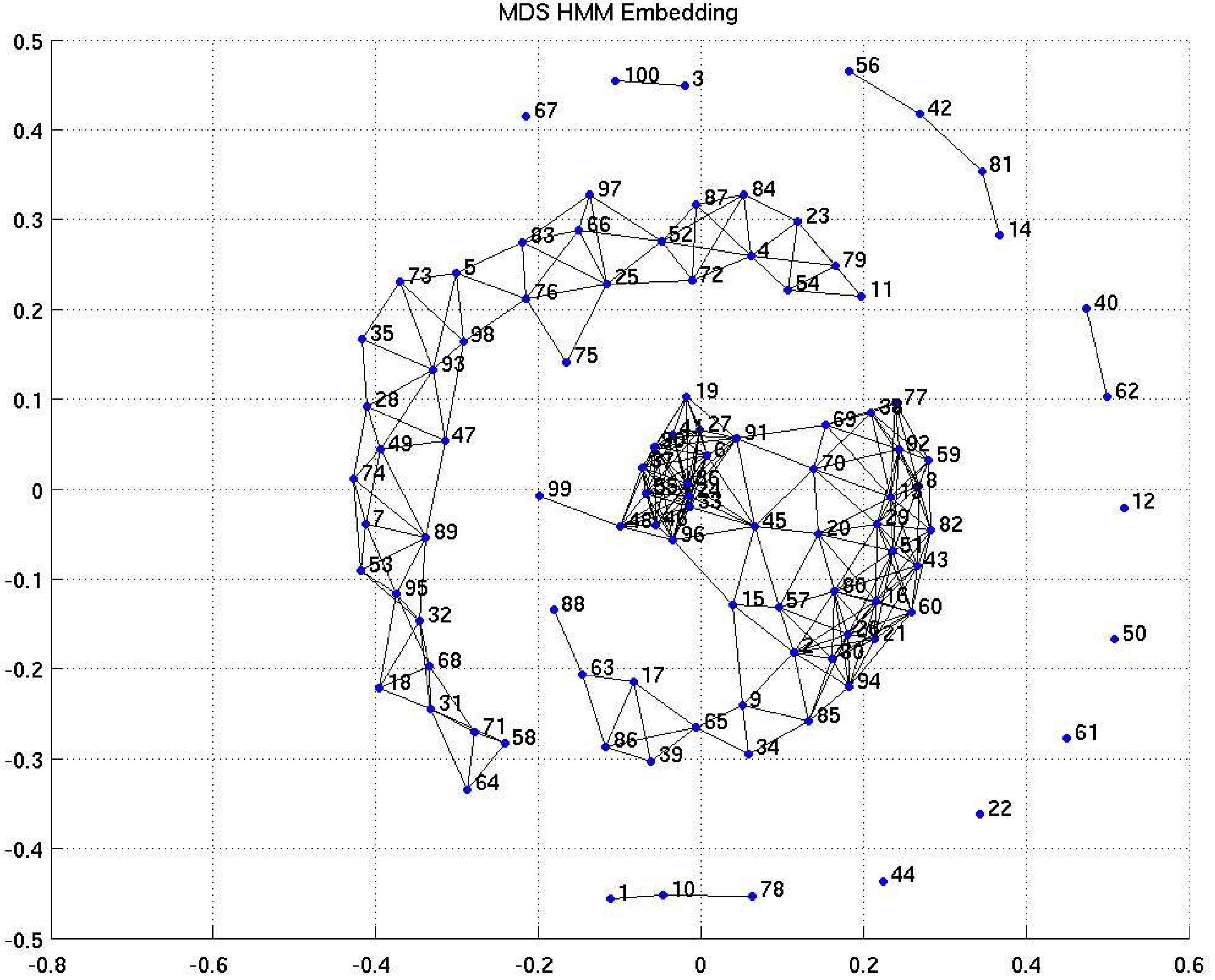
  - Also: Variational methods for large margin detection/classification of graphical models
    - *“Large Margin Latent Graphical Models”, T. Jebara*.

Embedding hidden Markov models

- General idea of semi-parametric learning for time series datasets
  - Time series data → HMMs, LDS, latent graphical models
  - Kernel between graphical models → Gram matrix
  - Gram matrix → spectral clustering, embedding, etc

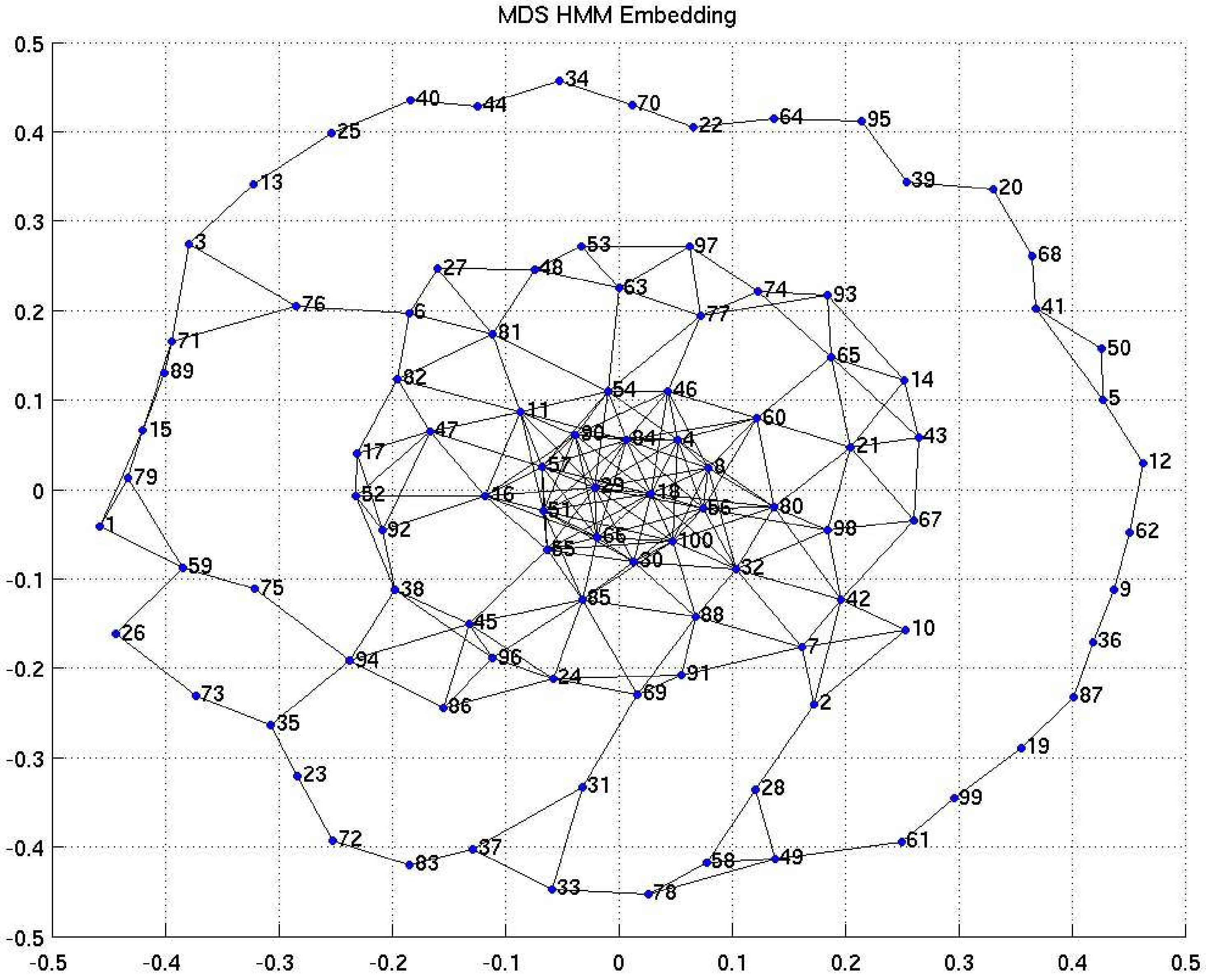
- Project: Part I (Behavior visualization)
  - Visualizing similarities in people's traveling patterns
  - Given a large set of data, 3.5 million time/position samples for ~4300 people, find the people with similar movement patterns.
  - Approach: (very straight forward)
    1. Isolate the time step samples for each person.
    2. Feature extraction – PCA for lat-longitude, sine/cosine for day/hour
    3. Train an HMM for each individual. Junction tree algorithm
    4. Use the probability product kernel to generate the gram matrix
    5. Use the gram matrix to calculate an embedding for the HMM parameters. multi-dimensional scaling (MDS), semi-definite embedding

MAO Dataset – Embedding of top 100 users – most active ranked



A 2-D look at HMM parameter manifold for movement behavior!

MAO Dataset – Embedding of top 100 users – day/hour entropy ranked



Bayesian evidence kernel for linear dynamical systems

- Part II: Linear dynamical systems
  - a type of latent graphical model
  - similar to HMMs but has a single continuous hidden state
  - transition between states is defined by linear projections
  - emissions also defined by linear projection
  - applications: tracking features across video frames
    - Target tracking, computer vision,
  - Completely specified by the following parameters
    - $\theta = \{A, C, Q, R, V, \pi\}$

$$\begin{aligned} x_t &= Ax_{t-1} + w_t \qquad w_t \sim \mathcal{N}(\vec{0}, Q) \\ y_t &= Cx_t + v_t \qquad v_t \sim \mathcal{N}(\vec{0}, R) \end{aligned}$$

    - x - state, y – emission
  - Regular LDS: Use EM to recover parameters: θ^*

Bayesian evidence kernel for linear dynamical systems

- Estimating the parameters of a LDS: EM
  - Initialize LDS parameters: A,C ← random projection matrices. Sample from multivariate normal *N*(0,*I*) for everything else.
  - E-STEP: Get the likelihood (filter and smooth)

    ```
    // Kalman filter
    for each observation t from 1:T
        1. predict state based on of the all previous observations 1:t
        2. update state prediction based on “Kalman gain”
    end
    // Shumway-Stoffer smoother
    for each observation t  from T-1:1
        1. calculate information gain for t given t +1
        2. update state t with information
    end
    ```

  - M-STEP: gradient ascent
  - Repeat until convergence
  - Recovers A, C, Q, R, V, π LDS parameters
    - minimizes the mean squared regression error
    - Can’t guarantee optimal solution – local max, etc…

Bayesian evidence kernel for linear dynamical systems

- Probability of observing a sequence, just assume gaussian:

$$\begin{aligned} p(\{x\},\{y\}) &= p(x_0)p(y_0|x_0)\prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t) \\ &= \mathcal{N}(x_0|\pi,V)\mathcal{N}(y_0|Cx_0,R)\prod_{t=1}^{T}\mathcal{N}(x_t|Ax_{t-1},Q)\mathcal{N}(y_t|Cx_t,R) \end{aligned}$$

- To be Bayesian, introduce a distribution on the prior and integrate.

$$p(\theta) = p(A,C,Q,R,V,\pi)$$

- For detection, want to calculate:

$$\mathcal{Z}(\{y\}) = \int p(\{x\},\{y\})p(\theta)d\theta$$

- For classification, need the kernel:

$$\mathcal{K}(\{y\},\{\bar{y}\}) = \int p(\{x\},\{y\})p(\{\bar{x}\},\{\bar{y}\})p(\theta)d\theta$$

Bayesian evidence kernel for linear dynamical systems

- Distribution over the priors:
  - Previous work by Matt Beal: used more complex priors but showed that the solving for the integral was intractable, used variational approximation methods

$$\begin{aligned} p(\{x\},\{y\}) &= p(x_0)p(y_0|x_0)\prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|x_t) \\ &= \mathcal{N}(x_0|\pi,V)\mathcal{N}(y_0|Cx_0,R)\prod_{t=1}^{T}\mathcal{N}(x_t|Ax_{t-1},Q)\mathcal{N}(y_t|Cx_t,R) \end{aligned}$$

- To make things tractable, we can introduce some constraints:
  - assume identity matrix for V,Q,R; white noise
  - also constraint A and C to be square matrices
  - zero mean identity variance for initial state prior
  - normal distribution over *each row* of transition and emission matrices
    - Lyapunov conditions for steady state covariance keeps the transition matrix near the origin. Like random projections. Not so much for C though…
- Alternatively
  - Wishart distributions (distribution where random variables are matrices) are probably better but will likely make the integrals harder to solve.

M. Beal (2003). Variational algorithms for approximate bayesian inference. Doctoral dissertation, Gatsby Computational Neuroscience Unit. University college London.

Bayesian evidence kernel for linear dynamical systems

- Distributions for the priors, a gaussian per row. A and C are projection matrices so assume rows are independent:

$$
\begin{aligned}
p(A|\vec{\alpha}) &= \prod_{i=1}^{M} N(A'_{(i)}|\vec{0}, \alpha_i I) \\
p(C|\vec{\gamma}) &= \prod_{i=1}^{M} N(C'_{(i)}|\vec{0}, \gamma_i I)
\end{aligned}
$$

- $A_{(i)}$ is the ith row of the matrix *A*

- Now the distribution over the priors is given as:

$$
\begin{aligned}
p(\pi, A, C|\vec{\alpha}, \vec{\gamma}) &= N(\pi|\vec{0}, I) \prod_{i=1}^{M} N(A'_{(i)}|\vec{0}, \alpha_i I) N(C'_{(i)}|\vec{0}, \gamma_i I) \\
&= \frac{1}{(2\pi)^{MD+D/2} (\prod_{i=1}^{M} \alpha_i \gamma_i)^{D/2}} \exp\left(-\frac{D}{2} \sum_{i=1}^{M} \left(\frac{A_{(i)} A'_{(i)}}{\alpha_i} + \frac{C_{(i)} C'_{(i)}}{\gamma_i} \right) \right)
\end{aligned}
$$

Bayesian evidence kernel for linear dynamical systems

- Expand the LDS equation to get:

$$\begin{aligned} p(\{x\},\{y\}) &= \frac{1}{(2\pi)^{(T+1)D/2}} \exp\left(-\frac{1}{2}(x_0-\pi)'(x_0-\pi)\right) \exp\left(-\frac{1}{2}(y_0-Cx_0)'(y_0-Cx_0)\right) \\ & \prod_{t=1}^{T} \exp\left(-\frac{1}{2}(x_t-Ax_{t-1})'(x_t-Ax_{t-1})\right) \exp\left(-\frac{1}{2}(y_t-Cx_t)'(y_t-Cx_t)\right) \end{aligned}$$

- Need to solve the following equation to get the normalizer:

$$\begin{aligned} \mathcal{Z}(\{y\}) &= \int p(\{x\},\{y\})p(\theta)d\theta \\ &= \int p(\{x\},\{y\})N(x_0|\pi,I)\prod_{i=1}^{M} N(A'_{(i)}|\vec{0},\alpha_i I)N(C'_{(i)}|\vec{0},\gamma_i I)d\pi\, dA\, dC \\ &= \frac{1}{(2\pi)^{(T+1)D/2}} \int\int\int \exp\left(-\frac{1}{2}(x_0-\pi)'(x_0-\pi)\right) \exp\left(-\frac{1}{2}(y_0-Cx_0)'(y_0-Cx_0)\right) \\ & \prod_{t=1}^{T} \exp\left(-\frac{1}{2}(x_t-Ax_{t-1})'(x_t-Ax_{t-1})\right) \exp\left(-\frac{1}{2}(y_t-Cx_t)'(y_t-Cx_t)\right) \\ & \frac{1}{(2\pi)^{MD+D/2}(\prod_{i=1}^{M}\alpha_i\gamma_i)^{D/2}} \exp\left(-\frac{D}{2}\sum_{i=1}^{M}\left(\frac{A_{(i)}A'_{(i)}}{\alpha_i}+\frac{C_{(i)}C'_{(i)}}{\gamma_i}\right)\right)d\pi\, dA\, dC \end{aligned}$$

Bayesian evidence kernel for linear dynamical systems

- Solve the following to get the kernel:

$$
\begin{aligned}
\mathcal{K}(\{y\},\{\bar{y}\}) &= \int p(\{x\},\{y\}|\theta)p(\{\bar{x}\},\{\bar{y}\}|\theta)p(\theta)d\theta \\
&= \frac{1}{(2\pi)^{(T+\mathcal{T}+4)D/2}} \int\int\int \exp\left(-\frac{1}{2}\left[(x_0-\pi)'(x_0-\pi)+(\bar{x}_0-\pi)'(\bar{x}_0-\pi)\right]\right) \\
&\quad \exp\left(-\frac{1}{2}\left[(y_0-Cx_0)'(y_0-Cx_0)+(\bar{y}_0-C\bar{x}_0)'(\bar{y}_0-C\bar{x}_0)\right]\right) \\
&\quad \prod_{t=1}^{T} \exp\left(-\frac{1}{2}(x_t-Ax_{t-1})'(x_t-Ax_{t-1})\right) \exp\left(-\frac{1}{2}(y_t-Cx_t)'(y_t-Cx_t)\right) \\
&\quad \prod_{t=1}^{\mathcal{T}} \exp\left(-\frac{1}{2}(\bar{x}_t-A\bar{x}_{t-1})'(\bar{x}_t-A\bar{x}_{t-1})\right) \exp\left(-\frac{1}{2}(\bar{y}_t-C\bar{x}_t)'(\bar{y}_t-C\bar{x}_t)\right) \\
&\quad \frac{1}{(2\pi)^{MD+D/2}(\prod_{i=1}^{M}\alpha_i\gamma_i)^{D/2}} \exp\left(-\frac{D}{2}\sum_{i=1}^{M}\left(\frac{A_{(i)}A'_{(i)}}{\alpha_i}+\frac{C_{(i)}C'_{(i)}}{\gamma_i}\right)\right) d\pi\, dA\, dC
\end{aligned}
$$

Bayesian evidence kernel for linear dynamical systems

- Derivations for the normalizer:
- Move the normalization constants outside and integrate over pi
- Assume probability of 1st emission is 1 to make the equation a little simpler

$$
\begin{aligned}
\mathcal{Z}(\{y\}) \; = \; & \frac{1}{(2\pi)^{TD/2}} \cdot \frac{1}{(2\pi)^{MD+D/2}(\prod_{i=1}^{M} \alpha_i \gamma_i)^{D/2}} \\
& \prod_{t=1}^{T} \int \int \exp\left(-\frac{1}{2}(x_t - Ax_{t-1})'(x_t - Ax_{t-1})\right) \exp\left(-\frac{D}{2}\sum_{i=1}^{M}\left(\frac{A_{(i)}A'_{(i)}}{\alpha_i}\right)\right) \\
& \exp\left(-\frac{1}{2}(y_t - Cx_t)'(y_t - Cx_t)\right) \exp\left(-\frac{D}{2}\sum_{i=1}^{M}\left(\frac{C_{(i)}C'_{(i)}}{\gamma_i}\right)\right) dA\, dC
\end{aligned}
$$

- now focus on solving the inner integral over A and C

$$
\begin{aligned}
\mathcal{I} \; = \; & \int_{A_{(M)}} \cdots \int_{A_{(1)}} \exp\left(-\frac{1}{2}(x_t - Ax_{t-1})'(x_t - Ax_{t-1})\right) \exp\left(-\frac{D}{2}\sum_{i=1}^{M}\left(\frac{A_{(i)}A'_{(i)}}{\alpha_i}\right)\right) dA_{(i)} ... dA_{(M)} \\
& \int_{A_{(D)}} \cdots \int_{C_{(1)}} \exp\left(-\frac{1}{2}(y_t - Cx_t)'(y_t - Cx_t)\right) \exp\left(-\frac{D}{2}\sum_{i=1}^{M}\left(\frac{C_{(i)}C'_{(i)}}{\gamma_i}\right)\right) dC_{(i)} ... dC_{(D)}
\end{aligned}
$$

To be completed…

Bayesian evidence kernel for linear dynamical systems

- Currently working on:
  - Fixing the derivations
  - Trying out different distributions on the priors
  - Clustering and embedding LDS models