Dynamical Factor Graphs (DFG) for Time Series Modeling

Piotr Mirowski
Advisor: Prof. Yann LeCun
Courant Institute of Mathematical Sciences, New York University

mirowski@cs.nyu.edu
http://cs.nyu.edu/~mirowski
Motivation for DFG

- **State-space model**
  - **Unknown latent states**
  - Potentially high-dimensional continuous latent states
  - Highly nonlinear dynamics or observation/control models (convolutional net)
  - Handle long sequences in linear time

- **Human MoCap**
  - Few visible markers
  - Many (hidden) joint angles

- **Chaotic time series**
  - Unobserved data
  - Complex, deterministic dynamics

- **Gene regulation networks**
  - Missing micro-array data

---

![Diagram of state-space model](Image)

\[ Y(t-1) \rightarrow Y(t) \rightarrow Y(t+1) \]
\[ Z(t-1) \rightarrow Z(t) \rightarrow Z(t+1) \]
\[ X(t-1) \rightarrow X(t) \rightarrow X(t+1) \]

\[ g \] observation model
\[ f \] dynamical model
\[ h \] control model

---

\[ Y(t+n) \] Predicted behavior

\[ Y(t) \] Observation model

\[ Z(t) \] Dynamical model

\[ X(t) \] Control model

---

\[ K\text{schischang et al, 2001; Taylor et al, 2006; Lorenz, 1963; Krouk et al, PNAS, submitted} \]
Dynamical Factor Graph (DFG)

- $n$-dimensional observed variable $Y(t)$
  - $Y(t) = g(Z(t)) + \omega(t)$ (Gaussian noise)

- $m$-dimensional latent variable $Z(t)$
  - $Z(t) = f(Z_{t-p}^{t-1}, Y(t-1)) + \epsilon(t)$ (Gaussian noise)

- $n$-dimensional observed variable
- $m$-dimensional latent variable
- Time-embedded sequence of $p$ latent variables

- Observation model $g$
- Dynamical model $f$
Highly nonlinear factors: convolutional networks

- Higher-order nonlinearity than:
  - radial basis functions
  - single hidden-layer Perceptrons
- No closed-form optimization, but gradient-based

\[ \text{n-dimensional input} \]
\[ \text{with time embedding } p=11: \]
\[ n \times p \]

Layer 1
- 12 filters:
  - \( n \times 5 \)
  - \( 1 \times 3 \) convolution (across time); time-step of 2

Layer 2
- 12 filters:
  - \( 1 \times 3 \)

Layer 3
- Full connection:
  - \( n \)-dimensional vector

\[ \text{12} \times 3 \text{ full connection} \]

\[ \text{[LeCun et al, 1998; Mirowski et al, AAAI 2007]} \]
Energy-based graph of a DFG

\[ E_o(t-1) \]
\[ ||Y^*(t-1) - Y(t-1)||_2^2 \]
\[ Y^*(t-1) \]
\[ g(Z(t-1), W_o) \]
\[ Y(t) \]
\[ ||Y^*(t) - Y(t)||_2^2 \]
\[ g(Z(t), W_o) \]
\[ Z(t-1) \]
\[ ||Z^*(t) - Z(t)||_2^2 \]
\[ f(Z_{t-p}, Y_{t-p}, W_d) \]
\[ Z(t) \]
\[ observation \]
\[ energy \]
\[ E_o(t) \]
\[ dynamical \]
\[ energy \]
\[ E_d(t) \]
\[ observation \]
\[ parameters \]
\[ W_o \]
\[ dynamical \]
\[ parameters \]
\[ W_d \]

Learning and inference: deterministic gradient-based EM

[Ghahramani & Roweis, 1999; Mirowski & LeCun, ECML 2009]
Smoothness penalty on latent variables

- Underconstrained latent variable inference
  \[ R_z(Z) = \sum_t \| Z(t) - Z(t + 1) \|^2 = \sum_t \sum_{k=1}^m (Z_k(t) - Z_k(t + 1))^2 \]
  \[ \rightarrow \text{L}_1\text{-norm smoothness penalty to tolerate “news”} \]
  (or random walk assumption with likely “shocks”)

- Smoothness penalty can add to
  or replace the dynamical model
Results\textsuperscript{1}: Inferring the Lorenz chaotic attractor

**Data**

Lorenz dynamical model

\[
\begin{align*}
\frac{\partial y_1}{\partial t} &= -16y_1 + 16y_2 \\
\frac{\partial y_2}{\partial t} &= 45.92y_1 - y_2 - y_1 \times y_3 \\
\frac{\partial y_3}{\partial t} &= y_1 \times y_2 - 2y_3
\end{align*}
\]

**Problem**

Partial observation \( y(t) = y_1(t) + y_2(t) + y_3(t) \)

Learn the DFG on train data with latent variables of dimension \( m=3 \)

**Results**

 Latent state attractor inferred on test data is similar to Lorenz attractor

1-step prediction error of -46.2dB smaller than in SVR (-41.6dB)

[\text{Lorenz, 1963; Mattera et al, 1999; Mirowski & LeCun, ECML 2009}]
Results\textsuperscript{2}: CATS time series

Data and problem

CATS time series prediction competition
Noisy chaotic time series (5000 points) with missing data (100 points)

Results

Predictions of 5 segments of missing data beat the CATS benchmark

[Data and problem description]

[Graph showing CATS target and inferred time series]

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Kalman Smoother</th>
<th>DFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1 (5 segments)</td>
<td>408</td>
<td>390</td>
</tr>
<tr>
<td>E2 (4 segments)</td>
<td>346</td>
<td>288</td>
</tr>
</tbody>
</table>

[Lendasse et al, 2004; Mirowski & LeCun, ECML 2009]
Results:\ Missing MoCap markers

Data
- Observations $Y$: 49-dimensional Motion Capture markers

Problem
- Model missing data (e.g. occlusions...)
  - Test sequence: 260 frames
  - 2 subsequences of 65 frames with missing data:
    - Left leg
    - Entire upper body

Approach
- Infer latent variables (E-step) on test sequence, (without gradient from missing $Y_i(t)$), generate $Y$ from $Z$

[Taylor et al, 2006; Mirowski & LeCun, ECML 2009]
Results: Missing MoCap markers

Original data

Reconstruction of missing upper body

Reconstruction of missing left leg

Lower NMSE than nearest neighbors; Inferred smooth, realistic motion

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Nearest Neighb.</th>
<th>DFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Leg 1</td>
<td>0.77</td>
<td>0.59</td>
</tr>
<tr>
<td>Missing Leg 2</td>
<td>0.47</td>
<td>0.39</td>
</tr>
<tr>
<td>Missing Upper Body 1</td>
<td>1.24</td>
<td>0.9</td>
</tr>
<tr>
<td>Missing Upper Body 2</td>
<td>0.8</td>
<td>0.48</td>
</tr>
</tbody>
</table>

[Taylor et al, 2006; Mirowski & LeCun, ECML 2009]
Results: Learning Genetic Regulatory Networks (GRN)

Data and assumptions

Micro-arrays of protein expression levels for Arabidopsis Thaliana (plant), in reaction to NO$_3$- sampled in time

$$Y(t) = f(Y(t-1)) + e(t)?$$

>22k genes, only 7 time points and 2 “replicates”

Clustering restricts to subset of 76 genes

Problem (and results)

Infer the GRN from dynamics on gene expressions

[Krouk et al, submitted to PNAS]
Results: Learning Genetic Regulatory Networks (GRN)

Approach
- Linear dynamics $f +$ Gaussian noise
- GRN sparsity through $L_1$-norm regularization of $f$
- Latent variables $Z = \text{“smoothed” gene expression } Y$
i.e. identity observation model $h +$ Gaussian noise
- Cross-validate on last or 2 last time points

Results
- Leave-out-last and leave-out-2-last predictions:
  70% +/- 3% correct direction
  (vs. 51% or 64% naive extrapolation)
- Incorporate prior knowledge about impossible
gene interactions: set sparse connections in $f$
- Impute micro-array data (infer missing values)

[Krouk et al, submitted to PNAS]
Results\(^5\): Epileptic seizure prediction from EEG

**Problem**

Epilepsy: a *chronic* illness  
Affects 1% of world population  
Seizures are *harrowing*  
40% of patients medication *refractory*  
Avoid resective surgery treatment

Discriminate  
(binary classification)

- Pre-ictal  
- Interictal  
- Ictal  
- Post-ictal

2h  
1h  
>1h

**Data**

Depth electrodes  
Grid electrodes  
Strip electrodes

**Approach**

Extraction of features from EEG, pattern recognition + classification convolutional networks  
Seizure onset  
intracranial EEG

Classify patterns of synchronization of EEG

Approach

Current results

21-patient public dataset: predicted 71% seizures, no false positives, >30min ahead of seizure
Best results ever achieved

Revised approach

Results: Epileptic seizure prediction from EEG

Examples of patterns of cross-correlation

1min of interictal EEG

1min of preictal EEG

Brownian motion

Latent patient state Z

Time-to-seizure Y

Y(t-1)

Y(t)

Z(t-1)

Z(t)

X(t-1)

X(t)

Thank you

• Further references:
  
  
  
  
  
  
Additional material
Inference of latent variables $Z$

\[
E_d(Z_{t-p}^{t-1}, Z(t), W_d) = \frac{1}{2} \| f(Z_{t-p}^{t-1}; W_d) - Z(t) \|^2_2
\]

\[
E_o(Z(t), Y(t), W_o) = \frac{1}{2} \| g(Z(t), W_o) - Y(t) \|^2_2
\]

Inference of latent variables $Z$

\[=\]

minimization w.r.t. $Z$ of

\[
E_d(Y_{t-p}^t; W_d) = \min_Z E_d(Z_{t-p}^{t-1}, Z(t), W_d)
\]

\[
E_o(Y(t), W_o) = \min_Z E_o(Z(t), Y(t), W_o)
\]

Total energy of the DFG given a sequence $Y$ and model parameters $W_o, W_d$:

\[
E(Y_{t_a}^{t_b}; W_d, W_o) = \alpha \sum_{t=t_a+1}^{t_b} E_d(Y_{t-p}^t; W_d) + \beta \sum_{t=t_a}^{t_b} E_o(Y(t), W_o)
\]

[Ghahramani & Roweis, 1999; Ranzato et al, 2007]
Learning of DFG model parameters

Loss function to minimize

\[ L(Y, Z; W) = \sum_t (\alpha E_d(Z_{t-p}^t, Z(t), W_d) + \beta E_o(Z(t), Y(t), W_o) + R(W) + R_Z(Z) \]

e.g. L_1 regularization of model parameters

Sparsity or smoothness penalties during state inference

Deterministic gradient-based version of Expectation-Maximization

E-step (latent variable inference)

annealed gradient descent on minibatch of Z until convergence

M-step (parameter learning)

1 step of stochastic gradient descent (diagonal Levenberg-Marquard)

\[ \tilde{Z} = \arg \min_Z L(\tilde{W}, Y, Z) \]

\[ \tilde{W} = \arg \min_W L(W, Y, \tilde{Z}) \]

[LeCun et al, 1998b; Ghahramani & Roweis, 1999; Ranzato et al, 2007]