Physiological Acuity Modelling with (Ugly) Temporal Clinical Data

Marzyeh Ghassemi
CSAIL PhD Candidate
Agenda

• Techniques
  • Topic Models (LDA)
  • Gaussian Processes (GP)

• Applications
  • KDD 2014 - Unfolding Physiological State: Mortality Modeling in Intensive Care Units
  • AAAI 2015 - A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data
Agenda

• Techniques
  • Topic Models (LDA)
  • Gaussian Processes (GP)

• Applications
  • KDD 2014 - Unfolding Physiological State: Mortality Modeling in Intensive Care Units
  • AAAI 2015 - A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data
Topic Model Tutorial

• Content is from:
  • Steyvers & Griffiths 2006 paper
  • Blei ICML 2012 Tutorial
Topic Models – Popularity is great

• All the right cliques:
  • Directed graphical models
  • Conjugate priors and nonconjugate priors
  • Time series modeling
  • Modeling with graphs
  • Hierarchical Bayesian methods
  • Approximate posterior inference (MCMC, variational methods)
  • Exploratory and descriptive data analysis
  • Model selection and Bayesian nonparametric methods
  • Mixed membership models
  • Prediction from sparse and noisy inputs
Data/Discovery Process

Make assumptions

Collect data

Infer the posterior

Check

Predict

Explore
How to Get the “Latent”?

- Graphical Models ~ Matrix Decomp ~ Tensor Decomp
Intuition: Documents are made of Topics

• Every document is a mixture of topics
• Every topic is a distribution over words
• Every word is a draw from a topic

**Seeking Life’s Bare (Genetic) Necessities**

*COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s bacteria can be sustained with just 250 genes and that the earliest life forms required as mere 125 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, these predictions show not that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Srir Anandalog, a genetics professor at the University of California. But coming up with a constraintless answer may be more than just a scientific number puzzle, particularly as more and more genomes are rapidly mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arccy Muthugan, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing..."
Circles & Boxes

- Observe: N words over D documents $W_{d,n}$

- Infer:
  - Per-word topic assignment $Z_{d,n}$
  - Per-doc topic proportion $\theta_d, \beta_k$
  - Corpus topic distribution $\alpha, \eta$

- Dirichlet Priors Give:
  - Sparsity $\alpha$
  - Exclusivity $\eta$
LDA – Latent Dirichlet Allocation

- We observe words, we infer everything else, with our assumed structure

\[
\prod_{i=1}^{K} p(\beta_i | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \left( \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)
\]
“Dirich-let” It On Too Thick?

• What are $\alpha$ & $\eta$?
• Each hyperparameter is a prior “observation count”:
  • $\alpha$ is the number of times a topic is sampled in a document before having observed anything from the document.
  • $\eta$ is the number of times words are sampled from a topic before any words are observed from the corpus.

![Diagram](image)

**Figure 3.** Illustrating the symmetric Dirichlet distribution for three topics on a two-dimensional simplex. Darker colors indicate higher probability. Left: $\alpha = 4$. Right: $\alpha = 2$. 
Why Do We Need Inference?

• Want the posterior distribution $p(z|w)$ - assignment of word to topics
• We could estimate $\theta_d, \beta_k$ using EM, or marginalize out with approx. inf.

$$\prod_{i=1}^{K} p(\beta_i | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \left( \prod_{n=1}^{N} p(z_{d,n} | \theta_d)p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

• Many Approximate Methods
  • Sampling – *randomly* resample a *specific* tagging for each word, given specific taggings of all other words, and a specific value for $\theta$.
  • Variational Inference - *deterministically* update the *distribution* over taggings for each word, given *distributions* over the taggings for other words and a *distribution* over $\theta$. 

LDA
Agenda

• Techniques
  • Topic Models (LDA)
  • Gaussian Processes (GP)

• Applications
  • **KDD 2014** - Unfolding Physiological State: Mortality Modeling in Intensive Care Units
  • **AAAI 2015** - A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data
GP Tutorial

• Content is from:
  • Phillip Henning MLSS 2013 Tutorial
  • Murphy’s Machine Learning Book (and code!)
• GPs define a prior over functions, which can be converted into a posterior over functions once we’ve seen some data.

• Assumes \( p(f(x_1), \ldots, f(x_n)) \) is jointly Gaussian, with some mean and covariance given by

• Computation is \( O(N^3) \).

• GPs can be thought of as a Bayesian alternative to sparser/faster kernel methods (SVM), with probabilistic outputs.
Multivariate Gaussian

\[ N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{N/2}\lvert \Sigma \rvert^{1/2}} \exp\left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \]

\[ \bullet \quad x, \mu \in \mathbb{R}^N, \Sigma \in \mathbb{R}^{N \times N} \]
\[ \bullet \quad \Sigma \text{ is positive semidefinite} \]
Why do we like them?

- Closure under multiplication
- Closure under linear maps
- Closure under marginalization

GPs

\[
\int \mathcal{N} \left[ \left( \begin{array}{c} x \\ y \end{array} \right) ; \left( \begin{array}{c} \mu_x \\ \mu_y \end{array} \right), \left( \begin{array}{cc} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{array} \right) \right] dy = \mathcal{N}(x; \mu_x, \Sigma_{xx})
\]

\[
p(x | y) = \frac{p(x, y)}{p(y)} = \mathcal{N} \left( x; \mu_x + \Sigma_{xy} \Sigma_{yy}^{-1} (y - \mu_y), \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \right)
\]

\[
\begin{cases}
\mathcal{N}(x; a, A) \mathcal{N}(x; b, B) = \mathcal{N}(x; c, C) \mathcal{N}(a; b, A + B) \\
C := (A^{-1} + B^{-1})^{-1} & c := C(A^{-1}a + B^{-1}b) \\
p(z) = \mathcal{N}(z; \mu, \Sigma) \\
p(Az) = \mathcal{N}(Az, A\mu, A\Sigma A^T)
\end{cases}
\]
What can we do?

given $y \in \mathbb{R}^N$, $p(y|f)$, what's $f$?
Linear Regression

\[ f(x) = w_1 + w_2 x = \phi_x^T w \]
\[ \phi_x = \begin{pmatrix} 1 \\ x \end{pmatrix} \]
\[ p(w) = \mathcal{N}(w; \mu, \Sigma) \]
\[ p(f) = \mathcal{N}(f; \phi_x^T \mu, \phi_x^T \Sigma \phi_x) \]

Prior over linear functions

\[
p(y | w, \phi_X) = \mathcal{N}(y; \phi_X^T w, \sigma^2 I)
\]
\[
p(f_x | y, \phi_X) = \mathcal{N}(f_x; \phi_x^T \mu + \phi_x^T \Sigma \phi_X (\phi_X^T \Sigma \phi_X + \sigma^2 I)^{-1} (y - \phi_X^T \mu),
\phi_x^T \Sigma \phi_x - \phi_x^T \Sigma \phi_X (\phi_X^T \Sigma \phi_X + \sigma^2 I)^{-1} \phi_X^T \Sigma \phi_x
\]

Posterior over linear functions
Is this hard??

```matlab
% prior on w
F = 2; % number of features
phi = @(a)(bsxfun(@power,a,0:F-1)); % phi(a) = [1;a]
mu = zeros(F,1); % p(w) = N(μ,Σ)
Sigma = eye(F); % features of x

% prior on f(x)
n = 100; % ‘test’ points
x = linspace(-6,6,n); % samples from prior
phix = phi(x); 
m = phix * mu; % marginal stddev, for plotting
kxx = phix * Sigma * phix'; % p(f_x) = N(m,k_xx)
s = bsxfun(@plus,m,chol(kxx + 1.0e-8 * eye(n))' * randn(n,3)); % features of data
stdpi = sqrt(diag(kxx)); % gives Y,X,sigma
load('data.mat'); N = length(Y); % gives Y,X,sigma

% prior on Y = f_X + ε
phiX = phi(X); % features of data
M = phiX * mu;
kXX = phiX * Sigma * phiX'; % p(f_X) = N(M,kXX)
G = kXX + sigma^2 * eye(N); % p(Y) = N(M,kXX + σ^2 I)
R = chol(G); % most expensive step: O(N^3)
kXX = phix * Sigma * phiX';
A = kXX / R; % cov(f_x,f_X) = k_xx
mpost = m + A * (R' \\ (Y-M)); % pre-compute for re-use
vpost = kxx - A * A'; % p(f_x | Y) = N(m + k_xx X (kXX + σ^2 I)^{-1} (Y - M),
spost = bsxfun(@plus,mpost,chol(vpost + 1.0e-8 * eye(n))' * randn(n,3)); % p(Y | X)
stdpo = sqrt(diag(vpost)); % marginal stddev, for plotting
```
More realistic data

\[ f(x) = \phi_x^T w \]

Graph showing a plot of data points with the equation \( f(x) = \phi_x^T w \) and axes labeled as \( x \) and \( f(x) \).
Cubic Regression

\[
f(x) = \phi(x)^\top w \quad \phi(x) = \begin{pmatrix} 1 & x & x^2 & x^3 \end{pmatrix}^\top
\]

Prior

Posterior

\[
f(x) = \phi(x)^\top w \quad \phi(x) = \begin{pmatrix} 1 & x & x^2 & x^3 \end{pmatrix}^\top
\]
Not any harder.

\[
F = 4
\]

\[
\begin{align*}
\text{load('data.mat'); } N &= \text{length}(Y); \\
\text{mpost} &= m + A \times (R' \setminus (Y-M)); \\
\text{vpost} &= kxx - A \times A'; \\
\text{spost} &= \text{bsxfun(@plus, mpost, chol(vpost + 1.0e-8 * eye(n))' * randn(n,3))}; \\
\text{stdpo} &= \text{sqrt(diag(vpost))};
\end{align*}
\]
Septic Regression

\[ f(x) = \phi(x)^\top w \quad \phi(x) = \begin{pmatrix} 1 & x & x^2 & \cdots & x^7 \end{pmatrix}^\top \]
Fourier Regression

$$\phi(x) = (\cos(x) \cos(2x) \cos(3x) \ldots \sin(x) \sin(2x) \ldots)^T$$
Step Regression

\[ \phi(x) = -1 + 2 \begin{pmatrix} \theta(x-8) & \theta(8-x) & \theta(x-7) & \theta(7-x) & \ldots \end{pmatrix}^\top \]

Prior

Posterior
How many features should we use?

\[ p(f_x | y, \phi_X) = \mathcal{N}(f_x; \phi_x^T \mu + \phi_x^T \Sigma \phi_X (\phi_X^T \Sigma \phi_X + \sigma^2 I)^{-1}(y - \phi_X^T \mu), \phi_x^T \Sigma \phi_x) \]

all objects involving \( \phi \) are of the form
- \( \phi^T \mu \) — the mean function
- \( \phi^T \Sigma \phi \) — the kernel

once these are known, cost is independent of the number of features

remember the code:

\[
\begin{align*}
M &= \text{phiX} \ast \text{mu}; \\
m &= \text{phiX} \ast \text{mu}; \\
kXX &= \text{phiX} \ast \text{Sigma} \ast \text{phiX'}; \\
kxx &= \text{phiX} \ast \text{Sigma} \ast \text{phiX'}; \\
kX &= \text{phiX} \ast \text{Sigma} \ast \text{phiX'};
\end{align*}
\]

\% \( p(f_X) = \mathcal{N}(M, kXX) \)
\% \( p(f_x) = \mathcal{N}(m, kxx) \)
\% \( \text{cov}(f_x, f_X) = kX \)
Pre-compute the kernel

```matlab
% prior
F = 2; % number of features
phi = @(a)(bsxfun(@power,a,0:F)); % \phi(a) = [1;a]
k = @(a,b)(phi(a)' * phi(b)); % kernel
mu = @(a)(zeros(size(a,1))); % mean function

% belief on f(x)
n = 100; x = linspace(-6,6,n)'; % 'test' points
m = mu(x);
kxx = k(x,x); % \textbf{p}(f_x) = \mathcal{N}(m,k_{xx})
s = bsxfun(@plus,m,chol(kxx + 1.0e-8 * eye(n))' * randn(n,3)); % samples from prior
stdpi = sqrt(diag(kxx)); % marginal stddev, for plotting
load('data.mat'); N = length(Y); % gives Y,X,sigma

% prior on Y = f_X + \epsilon
M = mu(X); % \textbf{p}(f_X) = \mathcal{N}(M,k_{XX})
kXX = k(X,X);
G = kXX + sigma^2 * eye(N); % \textbf{p}(Y) = \mathcal{N}(M,k_{XX} + \sigma^2 I)
R = chol(G); % most expensive step: \mathcal{O}(N^3)

kXX = k(x,X);
A = kXX / R; % \textbf{pre-compute for re-use}

mpost = m + A * (R' \backslash (Y-M)); % \textbf{p}(f_x | Y) = \mathcal{N}(m + k_{xx} (k_{XX} + \sigma^2 I)^{-1} (Y - M), k_{xx} - k_{XX} (k_{XX} + \sigma^2 I)^{-1} k_{xx})
vpost = kxx - A * A';
spost = bsxfun(@plus,mpost,chol(vpost + 1.0e-8 * eye(n))' * randn(n,3)); % samples
stdpo = sqrt(diag(vpost)); % marginal stddev, for plotting
```
Definition

A function \( k : X \times X \to \mathbb{R} \) is a Mercer kernel if, for any finite collection \( X = [x_1, \ldots, x_N] \), the matrix \( k_{XX} \in \mathbb{R}^{N \times N} \) with elements \( k_{XX,(i,j)} = k(x_i, x_j) \) is positive semidefinite.

Lemma

Any kernel that can be written as

\[
    k(x, x') = \int \phi_\ell(x) \phi_\ell(x') \, dl
\]

is a Mercer kernel. (assuming integral over positive set)

Proof: \( \forall X \in X^N, v \in \mathbb{R}^N \)

\[
    v^T k_{XX} v = \int \sum_i v_i \phi_\ell(x_i) \sum_j v_j \phi_\ell(x_j) \, dl = \int \left[ \sum_i v_i \phi_\ell(x_i) \right]^2 \, dl \geq 0 \]
GPs

Gaussian Process Priors

Definition

A function \( k : X \times X \rightarrow \mathbb{R} \) is a Mercer kernel if, for any finite collection \( X = [x_1, \ldots, x_N] \), the matrix \( k_{XX} \in \mathbb{R}^{N \times N} \) with elements \( k_{XX,(i,j)} = k(x_i, x_j) \) is positive semidefinite.

Definition

Let \( \mu : X \rightarrow \mathbb{R} \) be any function, \( k : X \times X \rightarrow \mathbb{R} \) be a Mercer kernel.
A Gaussian process \( p(f) = \mathcal{GP}(f; \mu, k) \) is a probability distribution over the function \( f : X \rightarrow \mathbb{R} \), such that every finite restriction to function values \( f_X := [f_{x_1}, \ldots, f_{x_N}] \) is a Gaussian distribution \( p(f_X) = \mathcal{N}(f_X; \mu_X, k_{XX}) \).
E.g. Kernelization of Step Functions

\[ \phi = @(a)(\text{bsxfun}(\gt, a, \text{linspace}(-8, 8, 5))./\text{sqrt}(5)); \]

\[ \phi = @(a)(\text{bsxfun}(\gt, a, \text{linspace}(-8, 8, 20))./\text{sqrt}(20)); \]

\[ \phi = @(a)(\text{bsxfun}(\gt, a, \text{linspace}(-8, 8, 100))./\text{sqrt}(100)); \]

\[ k = @(a,b)(\theta.^2 \times \text{bsxfun}(\text{min}, a+8, b'+8)/16); \]

\[ \text{cov}(f_x, f_x) = \int_{c_{\min}}^2 \theta(x_i - c)\theta(x_j - c) \, dc = \min(x_i, x_j) - c_{\min} \]

aka. the Wiener process
Applying It

\[ k = @(a,b)(\theta^2 \ast \text{bsxfun}(@\text{min},a+8,b'+8)/16); \]
Gaussians are closed under
  - linear projection / marginalization / sum rule
  - linear restriction / conditioning / product rule
they provide the linear algebra of inference
combine with nonlinear features $\phi$, get nonlinear regression
in fact, number of features can be infinite
(nonparametric) Gaussian process regression
Agenda

• Techniques
  • Topic Models (LDA)
  • Gaussian Processes (GP)

• Applications
  • **KDD 2014** - Unfolding Physiological State: Mortality Modeling in Intensive Care Units
  • **AAAI 2015** - A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data
We’ve Got A Really Big Problem

• ICUs are busy, and carestaff are often inundated with information.

• Which patient needs attention?

How sick is he? Tests? Treatment?

Signals

Numerical

Narrative

Snapshot

00:00 12:00 24:00 36:00 48:00

Age Gender SAPS I

ICD9 EH CoMor

Discharge Note
Path Note
Doc Note
Doc Note
Nurse Note
What Do We Already Know?

• In 2009, 118 validated mortality prediction tools published.**
  • Modest accuracy
  • Large variability
  • Models based on numeric, waveform, or snapshot data
  • Snapshot data (e.g. ICD9) is not “realtime” or actionable

• A good predictive rule must be*:
  • Accurate in a wide variety of clinical settings
  • Easy to incorporate into routine clinical practice
  • Improves prognostic accuracy

* Grady, Deborah, and Seth A. Berkowitz. "Why is a good clinical prediction rule so hard to find?." Archives of internal medicine 171.19 (2011): 1701-1702.
Unfolding Physiological State: Mortality Modeling in Intensive Care Units

• KDD 2014

• Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, Peter Szolovits
Lots of Data Sources

Signals

Numerical

Narrative

Snapshot

00:00 12:00 24:00 36:00 48:00

ICD9 EH CoMor

Age Gender SAPS I

Nurse Note Doc Note Doc Note Path Note Discharge Note
Every Cat Needs a Plan

- Create forward-facing models every 12 hours that only use data what would have actually been available, or “realtime” data.

- Incorporate clinical text with snapshot data.

- Measure performance on mortality prediction in-hospital, at 30-days and 1-year post-discharge.

Hypothesis: Text information decomposed into topic features adds value to snapshot data.
Model Setup: Overview

Data

Un-supervised LDA Model

Aggregated Feature Matrix

Structured SVM Model

12 Hours 24 Hours 36 Hours

Patient 1

Patient N

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>SAPS I</th>
<th>max{SAPS I}</th>
<th>...</th>
<th>EH_Comor_{50}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

T_1 ... T_{50} Age Sex SAPS I max{SAPS I} ... EH_Comor_{50}

...
Model Setup: Data

- Use 19,308 adult patient records
- Gather per-patient snapshot information
- Collect 473,764 notes
  - Use only first admissions
  - Ignore discharge summaries

![Normalized Distribution of Note Offsets from First Note Time](image-url)
Model Setup: Latent Topic Features

<table>
<thead>
<tr>
<th>Topic #</th>
<th>Top Ten Words</th>
<th>Possible Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Hospital Mortality</td>
<td>27  name family neuro care noted status plan stitle dr remains</td>
<td>Discussion of end-of-life care</td>
</tr>
<tr>
<td></td>
<td>15  intubated vent ett secretions propofol abg respiratory resp care sedated</td>
<td>Respiratory failure</td>
</tr>
<tr>
<td></td>
<td>7   thick secretions vent trach resp tf tube coarse cont suctioned</td>
<td>Respiratory infection</td>
</tr>
<tr>
<td></td>
<td>5   liver renal hepatic ascites dialysis failure flow transplant portal ultrasound</td>
<td>Renal failure</td>
</tr>
<tr>
<td>Hospital Survival</td>
<td>1  cabg pain ct artery coronary valve post wires chest sp neo</td>
<td>Cardiovascular Surgery</td>
</tr>
<tr>
<td></td>
<td>40  left fracture ap views reason clip hip distal lat report joint</td>
<td>Fracture</td>
</tr>
<tr>
<td></td>
<td>16  gtt insulin bs lasix endo monitor mg am plan iv</td>
<td>Chronic diabetes</td>
</tr>
<tr>
<td>1 Year Mortality</td>
<td>3  picc line name procedure catheter vein tip placement clip access</td>
<td>PICC line insertion</td>
</tr>
<tr>
<td></td>
<td>4   biliary mass duct metastatic bile cancer left ca tumor clip</td>
<td>Cancer treatment</td>
</tr>
<tr>
<td></td>
<td>45  catheter name procedure contrast wire french placed needle advanced clip</td>
<td>Coronary catheterization</td>
</tr>
</tbody>
</table>
Model Setup: Time-varying Topics

- Time-varying Topic Model:
  - Normalized topic distribution (50 features)
Model Setup: Admission Baseline

- Admission Baseline Model:
  - Age, gender, admitting SAPS I score (3 features)
• **Combined Time-varying Model:**
  • Admission and topic features (53 features)
Model Setup: Retrospective Topics

- Retrospective Topic Model:
  - Retrospective note features from entire patient stay (50 features).
Model Setup: Retrospective Topics + Admission

- **Retrospective Topic + Admission Model:**
  - Combined topic and admission feature (53 features).
Model Setup: Retrospective Derived

- **Retrospective Derived Features Model:**
  - Age, gender, admitting/min/max/final\{SAPS I\} and Elixhauser co-morbidity scores (36 features).
Model Setup: Retrospective Topics + Derived

- Retrospective Topic + Derived Features Model:
  - Combine all retrospective (86 features).
Mortality Prediction Results

In-Hospital Mortality

1-Year Mortality

30-Day Mortality

- Admissions Baseline Model
- Time-varying Topic Model
- Combined Time-Varying Model
- Retrospective Derived Feature Model
- Retrospective Topic Model
- Retrospective Topic + Admission Model
- Retrospective Topic + Derived Feature Model
We Solved A Problem, Everything Is Awesome

• Text Data Is Valuable
  • A combination of latent topic features and snapshot features worked best

• Long-term Predictions Are Harder
  • Combinations of features were best able to perform over first 24 hours.

• “Realtime” Models Are More Valuable
  • Retrospective models out-performed continuous actionable.
A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data

• AAAI 2015

• Marzyeh Ghassemi, Marco A. F. Pimentel, Tristan Naumann, Thomas Brennan, David A. Clifton, Peter Szolovits, Mengling Feng
Noisy, Sparse, Irregularly Sampled Data

- We use MTGPs to model the movements between and within multiple signals. This transforms a variety of irregularly-sampled clinical data into a new latent space using the MTGP hyperparameter.

A sample function with 4 tasks. Tasks 1 and 2 were correlated; 4 was anti-correlated with 1 and 2; and 3 was uncorrelated.

STGP predictions on all tasks. Mean absolute prediction error (over the 4 tasks) doesn’t use signal interaction.

MTGP predictions on all tasks. Predictions improved by taking into account the correlation between the different tasks.
Projection to Latent Space

θ provides a new latent search space to examine and evaluate the similarity of any two given multi-dimensional functions.
But What Are These Hyperparameters?

• We use the squared exponential kernel, so each signal’s $\theta$ govern the function input/output scale, and the inter-signal $\theta$ correspond to the correlation between the different outputs.

• These parameters are:
  1. a means of representing the functional behavior
  2. a set of observations learned directly from data; and
  3. generalizable to any type of longitudinal data, including categorical and numerical types

\[
K_{MT}(X_n, l, \theta_c, \theta_t) = K_c(l, \theta_c) \otimes K_t(X_n, \theta_t)
\]

\[
K_t = \theta_A^2 \exp \left\{ -\frac{\| x - x' \|^2}{2\theta_L^2} \right\}
\]

\[
K_c = LL^T, \quad L = \begin{bmatrix}
\theta_{c,1} & 0 & \cdots & 0 \\
\theta_{c,2} & \theta_{c,3} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\theta_{c,k-m+2} & \theta_{c,k-m+3} & \cdots & \theta_{c,k}
\end{bmatrix}
\]
Incorporating Text Data

- Perform a “pre-projection step”
- Clinical notes are transformed into timeseries using LDA
- New set of topic proportion timeseries are fitted using the MTGPs
- Inferred hyperparameters $\theta$ are derived, projecting into the new latent space.
Case Studies

- Estimating Signal in Traumatic Brain Injury Patients
- Mortality Prediction Using Clinical Progress Notes
Estimating Signal in Traumatic Brain Injury

• ICP and ABP data were collected from 35 TBI patients who were monitored for more than 24-hours in a Neuro-ICU.

• Our goal was to forecast the MAP and ICP signals as well as estimate cerebrovascular pressure reactivity (PRx)

<table>
<thead>
<tr>
<th>Signal</th>
<th>Measure</th>
<th>STGP</th>
<th>MTGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICP</td>
<td>RMSE</td>
<td>0.91</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>MSSL</td>
<td>0.6</td>
<td>0.45</td>
</tr>
<tr>
<td>ABP</td>
<td>RMSE</td>
<td>2.77</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>MSSL</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>PRx-PRx</td>
<td>RMSE</td>
<td>-</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Mortality Prediction Using Clinical Notes

• 10,202 patients with 313,461 notes.

• Chose the 9 topics with a posterior likelihood above or below 5% of the population baseline likelihood across topics.

• Using MTGP hyperparameters as additional classification features also gave us improved results for mortality prediction (0.812 vs 0.788 AUC).
Acknowledgements

Thanks to:
Intel Science and Technology Center for Big Data
NIH NLM Biomedical Institute Research Training