Fast Kronecker Inference in Gaussian Processes with non-Gaussian Likelihoods

**Seth Flaxman**¹, Andrew Gordon Wilson¹, Daniel B. Neill¹, Hannes Nickisch², Alexander J. Smola¹

1. Carnegie Mellon University
2. Philips Research Hamburg

ICML - 7 July 2015 - Lille
Overview

- GP models require $O(n^3)$ computations and $O(n^2)$ storage.
- New scalable Kronecker methods for non-Gaussian likelihoods.
- Laplace approximation with linear conjugate gradients for inference.
- A lower bound on the GP marginal likelihood for kernel learning.
- Small area spatiotemporal forecasting with log-Gaussian Cox Processes.
Gaussian Processes

- Observations \((x_1, y_1), \ldots, (x_n, y_n)\)
Gaussian Processes

- Observations \((x_1, y_1), \ldots, (x_n, y_n)\)
- Regression: learn \(f\) for \(y = f(x) + \epsilon\).
Gaussian Processes

- Observations \((x_1, y_1), \ldots, (x_n, y_n)\)
- Regression: learn \(f\) for \(y = f(x) + \epsilon\).
- GPs specify a prior over functions:

\[
f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))
\]
Gaussian Processes

- Observations \((x_1, y_1), \ldots, (x_n, y_n)\)
- Regression: learn \(f\) for \(y = f(x) + \epsilon\).
- GPs specify a prior over functions:
  \[
  f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))
  \]
- Likelihood ("observation model"):
  \[
  y_i | f(x_i) = \mathcal{N}(y_i; f(x_i), \sigma^2)
  \]
Gaussian Processes

- Observations \((x_1, y_1), \ldots, (x_n, y_n)\)
- Regression: learn \(f\) for \(y = f(x) + \epsilon\).
- GPs specify a prior over functions:
  \[
  f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))
  \]
- Likelihood ("observation model"):
  \[
  y_i|f(x_i) = \mathcal{N}(y_i; f(x_i), \sigma^2)
  \]
- Combine prior and likelihood to get posterior
Log-Gaussian Cox Process (LGCP)

- LGCP [Moller et al. 1998, Diggle et al. 2013] is an inhomogeneous Poisson process with stochastic intensity $\lambda$:

$$\log \lambda(s) \sim \mathcal{GP}(\mu(s), k_\theta(\cdot, \cdot)) .$$
Log-Gaussian Cox Process (LGCP)

- LGCP [Moller et al. 1998, Diggle et al. 2013] is an inhomogeneous Poisson process with stochastic intensity $\lambda$:

$$\log \lambda(s) \sim \mathcal{GP}(\mu(s), k_\theta(\cdot, \cdot)) .$$

- Conditional on a realization $\lambda$, the number of points in a given space-time region $S$ is:

$$y_S | \lambda(s) \sim \text{Poisson} \left( \int_{s \in S} \lambda(s) \, ds \right) .$$
Log-Gaussian Cox Process (LGCP)

- Following Diggle et al. [2013], we discretize to a grid, with $y_i$ points in cell $i$. Thus GP prior with Poisson observation model:

$$f \sim \text{GP}(\mu(s), k_\theta(\cdot, \cdot)).$$

$$y_i | f(s_i) \sim \text{Poisson}(\exp[f(s_i)])$$
Log-Gaussian Cox Process (LGCP)

- Following Diggle et al. [2013], we discretize to a grid, with $y_i$ points in cell $i$. Thus GP prior with Poisson observation model:

$$f \sim \text{GP}(\mu(s), k_\theta(\cdot, \cdot)).$$

$$y_i | f(s_i) \sim \text{Poisson}(\exp[f(s_i)])$$
Kronecker methods

• Reduce computation time, memory usage by exploiting:
  – Grid structure in inputs
  – Separable kernel: \( k((s, t), (s', t')) = k(s, s')k(t, t') \)

• LGCP w/ computational grid is naturally suited to Kronecker methods

• Kronecker methods in GPs: Bonilla et al. [2007], Finley et al. [2009], Stegle et al. [2011], Saati [2011], Gilboa et al. [2013], Riihimki and Vehtari [2014], Wilson et al. [2014], Groot et al. [2014]
Kronecker methods

Multivariate Gaussian distribution:

\[(2\pi)^{-n/2} |K|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T K^{-1}(x-\mu)}\]

Costly terms:

\[|K| \text{ and } K^{-1}\]

Assume observations on a grid (by construction) and separable covariance:

\[K = K_s \otimes K_t\]

\[k((s, t), (s', t')) = k(s, s')k(t, t')\]

Then:

\[\text{det}(K) = \prod_i \text{det}(K_s)^m \text{det}(K_t)^n\]

\[K^{-1}v = (K_s^{-1} \otimes K_t^{-1})v\]
Non-Gaussian likelihoods: Inference

\[ p(f|y, X) \approx \mathcal{N}(f|\hat{f}, (K^{-1} + W)^{-1}) \]

for \( W = -\nabla \nabla \log p(y|f) \).

- The problem: covariance in Laplace approximation \((K^{-1} + W)^{-1}\) is not Kronecker
- Matrix inverse with LCG: matrix-vector multiplications are still fast
- Small number of evaluations required, each efficient:

\[
(K^{-1} + W)v \\
= K^{-1}v + Wv \\
= (K_1^{-1} \otimes K_2^{-1})v + Wv
\]
Non-Gaussian likelihoods: Learning

Laplace approximate marginal likelihood:

$$\log p(y|X, \theta) = \log \int \exp[\Psi(f)] df$$

$$\approx \log p(y|\hat{f}) - \frac{1}{2} \alpha^\top K^{-1} \alpha - \frac{1}{2} \log |I + KW|,$$

Tricky term: $\log |I + KW|$. For psd matrices $U$ and $V$, Fiedler [1971]:

$$\prod_i (u_i + v_i) \leq |U + V| \leq \prod_i (u_i + v_{n-i+1})$$

where $u_1 \leq u_2 \leq \ldots \leq u_n$ and $v_1 \leq \ldots \leq v_n$ are the eigenvalues of $U$ and $V$. 
Fiedler bound

$K$ has eigenvalues $e_1 \leq e_2 \leq \ldots \leq e_n$.
$W$ has eigenvalues $w_1 \leq w_2 \leq \ldots \leq w_n$.

\[
\log |I + KW| = \log(|K + W^{-1}| |W|) \\
\leq \log \prod_i (e_i + w_i^{-1}) \prod_i w_i \\
= \sum_i \log(1 + e_i w_i)
\]

Final bound on log-marginal likelihood:

\[
\log p(y|X, \theta) \geq \log p(y|\hat{f}) - \frac{1}{2} \hat{\alpha}^\top K^{-1} \hat{\alpha} - \frac{1}{2} \sum_i \log(1 + e_i w_i)
\]
Experiments: synthetic data

Accuracy of our marginal likelihood approximation

Approximation ratio

# of observations

0

10

20

30

0.05

1.005

1.010

1.015

1.020

1.025

10000

20000

30000
Experiments: synthetic data

Accuracy of our log-determinant approximation

<table>
<thead>
<tr>
<th>Method</th>
<th>Approximation ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiedler</td>
<td></td>
</tr>
<tr>
<td>low rank, r = 5</td>
<td></td>
</tr>
<tr>
<td>low rank, r = 10</td>
<td></td>
</tr>
<tr>
<td>low rank, r = 15</td>
<td></td>
</tr>
<tr>
<td>low rank, r = 20</td>
<td></td>
</tr>
<tr>
<td>low rank, r = 30</td>
<td></td>
</tr>
</tbody>
</table>

- # of observations (x-axis)
- Approximation ratio (y-axis)
Experiments: synthetic data

Run-time of our log-determinant approximation

<table>
<thead>
<tr>
<th># of observations</th>
<th>run time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1.00</td>
</tr>
<tr>
<td>100.00</td>
<td>10,000</td>
</tr>
<tr>
<td>1,000,000</td>
<td></td>
</tr>
</tbody>
</table>

- **method**
  - Fiedler
  - Exact
  - low rank, r = 5
  - low rank, r = 15
  - low rank, r = 30
Experiments: synthetic data

Run-time of our algorithm vs. competitors

![Graph showing run-time of different methods vs. number of observations.](image)
Experiments: synthetic data

Accuracy of our algorithm vs. competitors

![Graph showing accuracy comparison between different methods.](image-url)
Local area crime forecasting

- Geocoded, data-stamped crime reports from Chicago\(^1\).
- Decade of data: 2004-2013. \(n = 233,088\) reported incidents of assault.
- After discretization: 1.6 million observations total

\(^1\)data.cityofchicago.org
Expressive spatiotemporal model

Impose grid: 1/2 mile × 1/2 mile × weekly counts.

Separable structure:

\[ k_{\theta}((x, y, t), (x', y', t')) = k_x(x, x')k_y(y, y')k_t(t, t') \]

\( k_x \) and \( k_y \) are Matérn-5/2 kernels.
\( k_t \) is a Spectral Mixture [Wilson & Adams 2013] kernel with 20 components:

\[ k(\tau) = \sum_{q=1}^{Q} w_q \exp(-2\pi^2 \tau^2 v_q) \cos(2\pi \tau \mu_q) \]

Negative Binomial observation model:

\[ y_i | f(s_i) \sim \text{NegBinom}(\exp(f(s_i)), r) \]
Assault in Chicago

Point pattern of assaults

Posterior Intensity
Bayesian posterior inference

Posterior Log-Intensity

Posterior Variance
Accurate small area forecasts


Forecasts: June 2012

Forecasts: December 2012

Assault Intensity (forecast)

- 2.0
- 1.5
- 1.0
- 0.5
- 0.0
Spatially varying trends

- AUSTIN
- ENGLEWOOD
- GREATER GRAND CROSSING
- HUMBOLDT PARK
- SOUTH CHICAGO
- WEST ENGLEWOOD (north)
- WEST ENGLEWOOD (south)
- WEST GARFIELD PARK
- WOODLAWN

Background
LGCP
Kronecker methods
Experiments
Crime forecasting
Expressive kernel learning

Background
LGCP
Kronecker methods
Experiments
Crime forecasting
# Numerical comparison

<table>
<thead>
<tr>
<th></th>
<th>KronNB SM-20</th>
<th>KronNB SM-20 Low Rank</th>
<th>KronGauss SM-20</th>
<th>FITC-100 NB SM-20</th>
<th>SSGPR-200</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training RMSE</strong></td>
<td>0.79</td>
<td>1.13</td>
<td>$10^{-11}$</td>
<td>2.14</td>
<td>1.45</td>
</tr>
<tr>
<td><strong>Forecast RMSE</strong></td>
<td>1.26</td>
<td>1.24</td>
<td>1.28</td>
<td>1.77</td>
<td>1.26</td>
</tr>
<tr>
<td><strong>Forecast log-likelihood</strong></td>
<td>-33,916</td>
<td>-172,879</td>
<td>-352,320</td>
<td>-42,897</td>
<td>-82,781</td>
</tr>
<tr>
<td><strong>Runtime</strong></td>
<td>2.8 hours</td>
<td>9 hours</td>
<td>22 min.</td>
<td>4.5 hours</td>
<td>2.8 hours</td>
</tr>
</tbody>
</table>
Future work

- Other inference methods? EP, VB, MCMC
- Other approximations? Inducing points, finite basis kernel representations
- Does computational grid work with rare events?
- Policy implications
Conclusion

- Extend Kronecker methods to GPs with non-Gaussian observation models
- Scalable, expressive kernel learning
- Kronecker + grid matches LGCP with computational grid
- Available in GPML [Rasmussen & Nickisch 2010]
  www.gaussianprocess.org/gpml/code
  (infGrid_Laplace)
- Tutorial online: www.cs.cmu.edu/~andrewgw/pattern/index.html
- Acknowledgments: NSF IIS-0953330, ONR N000141410684, NIH R01GM093156.