Final project: Scalable Log Determinants for Gaussian Process Kernel Learning

A. Bauman, F. Loginov, K. Nazirkhanova, K. Pirov, A.Savinov

Skoltech

December 20, 2018

What method was improved?

Regression for Gaussian Processes
In GP we have to deal with symmetric positive semi-definite covariance matrix $\widetilde{K}$:

$$\log|\widetilde{K}| = \operatorname{tr}(\log(\widetilde{K}))$$

How to estimate trace

Stochastic trace estimation

$$\mathrm{tr}(f(A)) \approx \frac{n}{n_\nu} \sum_{l=1}^{n_\nu} {v_l}^T f(A) v_l$$

Methods

1. Chebyshev
2. Lanczos
3. Taylor series

Chebyshev

z - probe vector, c_j - coefficients of Chebyshev decomposition

$$B = \frac{2\widetilde{K}}{\lambda_{max} - \lambda_{min}} - \frac{\lambda_{max} + \lambda_{min}}{\lambda_{max} - \lambda_{min}} I$$

$$w_0 = z, w_1 = Bz, w_{j+1} = 2Bw_j w_{j-1} \text{ for } j \geq 1$$

$$\log|\widetilde{K}| \approx \mathbb{E}\left[\sum_{j=0}^{m} c_j z^T w_j\right]$$

Chebyshev

$$\frac{\partial w_0}{\partial \theta_i} = 0, \quad \frac{\partial w_1}{\partial \theta_i} = \frac{\partial B}{\partial \theta_i} z,$$

$$\frac{\partial w_{j+1}}{\partial \theta_i} = 2\left(\frac{\partial B}{\partial \theta_i} w_j + B \frac{\partial w_j}{\partial \theta_i}\right) - \frac{\partial w_{j-1}}{\partial \theta_i} \text{ for } j \geq 1$$

$$\frac{\partial}{\partial \theta_i} \log |\widetilde{K}| \approx \mathbb{E}\left[\sum_{j=0}^{m} c_j z^T \frac{\partial w_j}{\partial \theta_i}\right]$$

Lanczos

$$\widetilde{K} Q_m = Q_m T + \beta_m q_{m+1} e_m^T$$

$$z^T \log(\widetilde{K}) z \approx e_1^T \log(\|z\|^2 T) e_1$$

$$\hat{g} = Q_m (T^{-1} e_1 \|z\|) \approx \widetilde{K}^{-1} z$$

$$tr(\widetilde{K}^{-1}(\frac{\partial \widetilde{K}}{\partial \theta_i})) = \mathbb{E}\left[(\widetilde{K}^{-1} z)^T \frac{\partial \widetilde{K}}{\partial \theta_i} z\right]$$

One more idea

$$\log(A) = 2^k \log(A^{\frac{1}{2^k}})$$

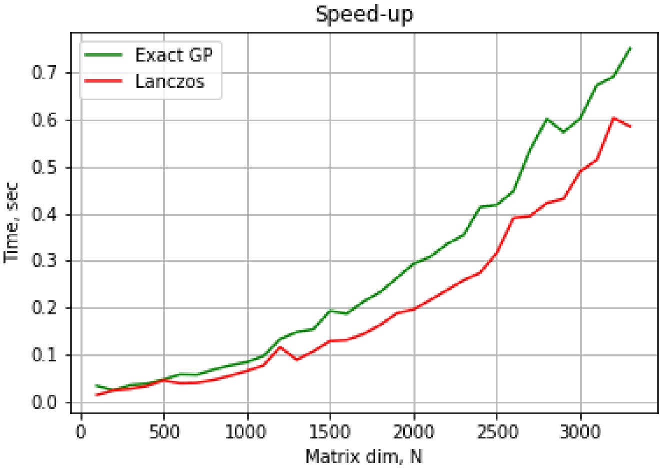
$$\log(I - W) = -\sum_{n=1}^{\infty} \frac{W^n}{n}, \text{if } ||W|| < 1$$

Due to $\sqrt[n]{n} \to 1$

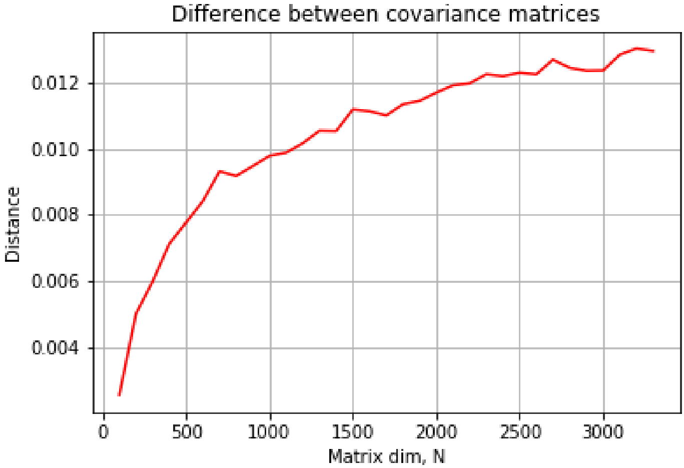
$$W = I - (A^{\frac{1}{2}})^k$$

And $A^{1/2} \to$ Newton

Results

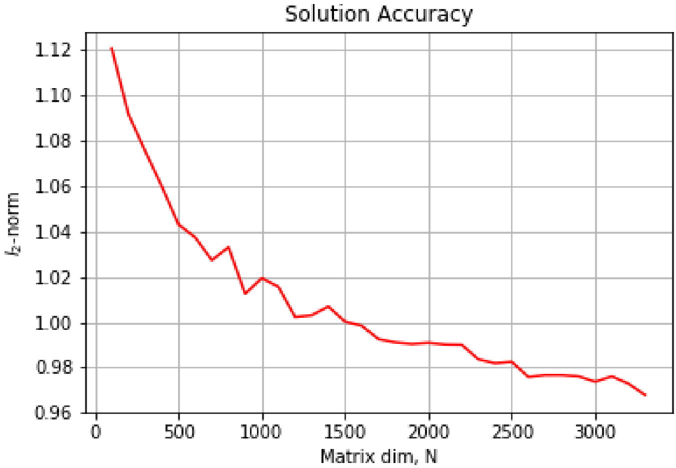

Speed-up
Exact GP
Lanczos
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Time, sec
0
500
1000
1500
2000
2500
3000
Matrix dim, N

Results

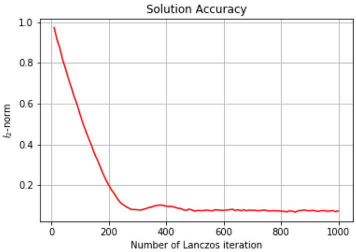


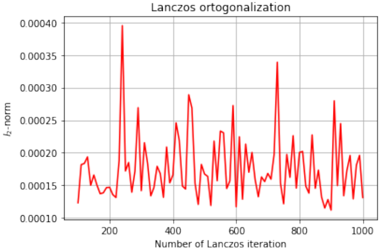
Difference between covariance matrices
0.012
0.010
0.008
0.006
0.004
Distance
0
500
1000
1500
2000
2500
3000
Matrix dim, N

Results

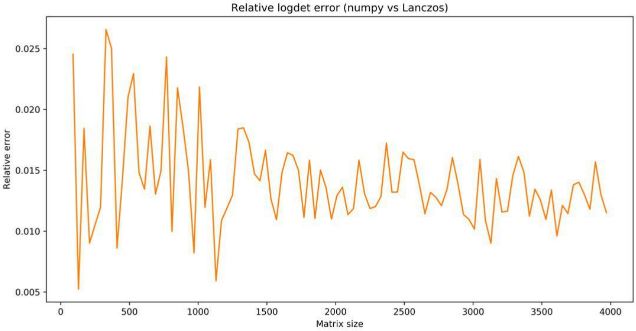

Solution Accuracy
1.12
1.10
1.08
1.06
1.04
1.02
1.00
0.98
0.96
l_2-norm
0
500
1000
1500
2000
2500
3000
Matrix dim, N

Accuracy and Stability





Comparison with built in function



Thank you for your attention!
Q & A

References

- Kun Dong, David Eriksson, Hannes Nickisch, David Bindel, Andrew Gordon Wilson, Scalable Log Determinants for Gaussian Process Kernel Learning, *Neural Information Processing Systems Conference 2017*, https://arxiv.org/pdf/1711.03481.pdf
- Geoff Pleiss, Jacob R. Gardner, Kilian Q. Weinberger, Andrew Gordon Wilson, *Constant-Time Predictive Distributions for Gaussian Processes, 2018*, https://arxiv.org/abs/1803.06058