Sparsity in Gaussian Processes Questions

Lehel Csató

Max-Planck Institute for Biological Cybernetics
Tübingen, Germany

Gaussian Processes Round Table[1]

[1] © N. Lawrence

Outline

1. Notations
   - Gaussian processes
   - Likelihoods
2. Approximations
   - Expectation propagation
   - Sparsification
   - Sparsity results
   - Comparisons
3. Examples
   - Predictive densities
   - Regression

Gaussian Processes

Gaussian processes – notations.

- For input locations $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, the associated random variables $\boldsymbol{f}_\mathcal{X} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)]^T$ are Gaussian:
$$p_0(\boldsymbol{f}_\mathcal{X}) = \mathrm{N}(\boldsymbol{\mu}_\mathcal{X}, \boldsymbol{K}_\mathcal{X})$$
- $\boldsymbol{\mu}_\mathcal{X}$ and $\boldsymbol{K}_\mathcal{X}$ are samples from the
  - Mean function $\mu(\boldsymbol{x}) = \langle f(\boldsymbol{x}) \rangle$
  - Covariance kernel function $K_0(\boldsymbol{x}, \boldsymbol{x}') = \langle f(\boldsymbol{x}) f(\boldsymbol{x}') \rangle$
- $p_0(\boldsymbol{f}|\theta_1)$ denotes the *prior* process, θ_1 are parameters of the kernel function.

Data likelihood

- $\mathcal{GP}$'s are *latent variables* in the inference process.
- Data $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ is factorising with *likelihood function*:

$$P(\mathcal{D}|\boldsymbol{f}_{\mathcal{X}}) = \prod_{n=1}^{N} P(y_n|f(\boldsymbol{x}_n), \theta_2)$$

θ_2 – parameters of the likelihood function

- The posterior process is computed using Bayes' rule

$$p_{\text{post}}(\boldsymbol{f}) = \frac{1}{Z} \int d\boldsymbol{f}_{\mathcal{X}} P(\mathcal{D}|\boldsymbol{f}_{\mathcal{X}}) p_0(\boldsymbol{f}_{\mathcal{X}}, \boldsymbol{f})$$

Problems when computing

$$p_{\text{post}}(\boldsymbol{f}) = \frac{1}{Z} \int d\boldsymbol{f}_{\mathcal{X}} P(\mathcal{D}|\boldsymbol{f}_{\mathcal{X}}) p_0(\boldsymbol{f}_{\mathcal{X}}, \boldsymbol{f})$$

- Non-Gaussian likelihoods lead to non-Gaussian processes
  - $p_{\text{post}}(\boldsymbol{f})$ not analytically computable.
  - Cannot compute the normalising

$$Z = \int d\boldsymbol{f}_{\mathcal{X}} P(\mathcal{D}|\boldsymbol{f}_{\mathcal{X}}) p_0(\boldsymbol{f}_{\mathcal{X}})$$

- For Gaussian likelihoods

$$\mu_{\text{post}}(\boldsymbol{x}) = \boldsymbol{k}_N(\boldsymbol{x})^T \left(\sigma_o^2 \boldsymbol{I} + \boldsymbol{K}_{NN}\right)^{-1} \boldsymbol{y}$$

the matrix inversion becomes prohibitive.

Approximation steps

- Approximating the non-Gaussian posterior with a Gaussian one:
  - retains information about values *and uncertainties*.
  - retains non-Gaussianity when predicting:

$$p(y_*|\boldsymbol{x}_*, \mathcal{D}) = \int df_* P(y_*|f_*, \theta_2) p_{\text{post}}(f_*)$$

- Further approximation – with a **sparse support set**
  - keeps the information up to the second order.
  - speeds up the computation of $p_{\text{post}}(f_*)$.

“Best” approximation to the posterior

Sparse $\mathcal{GP}$ ’s
Lehel Csató
Notations
Gaussian processes
Likelihoods
Approximations
Expectation propagation
Sparsification
Sparsity results
Comparisons
Examples
Predictive densities
Regression
Summary

Approximating the posterior process $p_{\text{post}}(\boldsymbol{f})$ with a Gaussian process $\hat{p}_{\text{post}}(\boldsymbol{f})$.

$\Longleftrightarrow$ minimising the Kullback-Leibler divergence:

$$d_{\text{KL}}\left(\hat{p}(\boldsymbol{f}) \| p_{\text{post}}(\boldsymbol{f})\right) = \int d\boldsymbol{f}\; p_{\text{post}}(\boldsymbol{f}) \log \frac{p_{\text{post}}(\boldsymbol{f})}{\hat{p}(\boldsymbol{f})}$$

The optimal $\mathcal{GP}$ has the first and second moments of the non-Gaussian posterior:

$$\begin{aligned} \hat{\mu}(\boldsymbol{x}) &= \mu_0(\boldsymbol{x}) + \sum_{n=1} \alpha_n K_0(\boldsymbol{x}, \boldsymbol{x}_n) \\ \hat{K}(\boldsymbol{x}, \boldsymbol{x}') &= K_0(\boldsymbol{x}, \boldsymbol{x}') + \sum_{m,n=1} K_0(\boldsymbol{x}, \boldsymbol{x}_m) C_{mn} K_0(\boldsymbol{x}_n, \boldsymbol{x}') \end{aligned}$$

Expectation propagation

- Iterating the following steps (ADATAP, Exp.Cons.):
  - For each n define the approximation that *excludes* n: $p_{\setminus n}(\boldsymbol{f})$
  - Build an approximation to the likelihood $P(y_n|f_n, \theta_2)$:

$$\frac{p_{\setminus n}(f_n)\, P(y_n|f_n, \theta_2)}{Z_n} \approx \mathrm{N}(f_n|\hat{\mu}_n, \hat{\sigma}_n) \stackrel{\text{def}}{=} \frac{p_{\setminus n}(f_n)\, \hat{t}(f_n|m_n, \lambda_n)}{\tilde{Z}_n}$$

  - The variational step: finding parameters (m_n, λ_n).

- Result – Gaussian approximation:

$$p_0(\boldsymbol{f}) \prod_n \frac{Z_n}{\tilde{Z}_n} \prod_n \hat{t}(f_n|m_n, \lambda_n)$$

The computational overload

The approximating $\mathcal{GP}$ is "anchored" at $\mathcal{X}$:

$$\hat{K}(\boldsymbol{x}, \boldsymbol{x}') = K_0(\boldsymbol{x}, \boldsymbol{x}') - \sum K_0(\boldsymbol{x}, \boldsymbol{x}_m) C_{mn} K_0(\boldsymbol{x}_n, \boldsymbol{x}')$$

- The computation time is **cubic** in data size.
- *Parameter* scale quadratically $\Longleftrightarrow$ over-parametrisation.

If data is "structured", then there is a less redundant representation of *the same* approximation.

KL-optimal sparsification

Sparse approximation step:

- The $\mathcal{GP}$ approximation $\hat{p}(\boldsymbol{f})$ is further reduced to a low-dimensional $\mathcal{GP}$ $p_{\mathcal{BV}}(\boldsymbol{f})$:

$$\hat{K}_{\mathcal{BV}}(\boldsymbol{x},\boldsymbol{x}') = K_0(\boldsymbol{x},\boldsymbol{x}') + \sum_{m,n\in\mathcal{BV}} K_0(\boldsymbol{x},\boldsymbol{x}_m) C_{mn} K_0(\boldsymbol{x}_n,\boldsymbol{x}')$$

- Kullback-Leibler divergence is used as minimiser:

$$p_{\mathcal{BV}}(\boldsymbol{f}) = \operatorname*{argmin}_{p_d\in\mathcal{GP}_d} \mathrm{KL}\left(p_d \| \hat{p}(\boldsymbol{f})\right)$$

  $\mathcal{GP}_d$ – $\mathcal{GP}$'s with d locations.

- Optimisation is NP-complete, greedy sequential approach.

KL-optimal sparsification II.

Sparse $\mathcal{GP}$'s
Lehel Csató
Notations
Gaussian processes
Likelihoods
Approximations
Expectation propagation
Sparsification
Sparsity results
Comparisons
Examples
Predictive densities
Regression
Summary

Assuming $|\mathcal{X}| = d + 1$

- Compute for each $n \in \mathcal{BV}_{d+1}$ (cheap approx.):

$$\mathrm{KL}\left(p_n \| \hat{p}_{d+1}(\boldsymbol{f})\right)$$

  Provides a measure of "how good" $\boldsymbol{x}_n$ is.
- Remove the one with the minimum KL-loss.

Result

$$\hat{p}(\boldsymbol{f}) \propto p_0(\boldsymbol{f}) \prod_n \hat{t}(\pi_n \boldsymbol{f}_{\mathcal{BV}} | m_n, \lambda_n)$$

with

$$\pi_n \boldsymbol{f}_{\mathcal{BV}} = E\left[f_n | \boldsymbol{f}_{\mathcal{BV}}\right]_0$$

KL-optimisation results

Reduced-rank $\mathcal{GP}$ using the **Basis Vectors – $\mathcal{BV}$ set**:

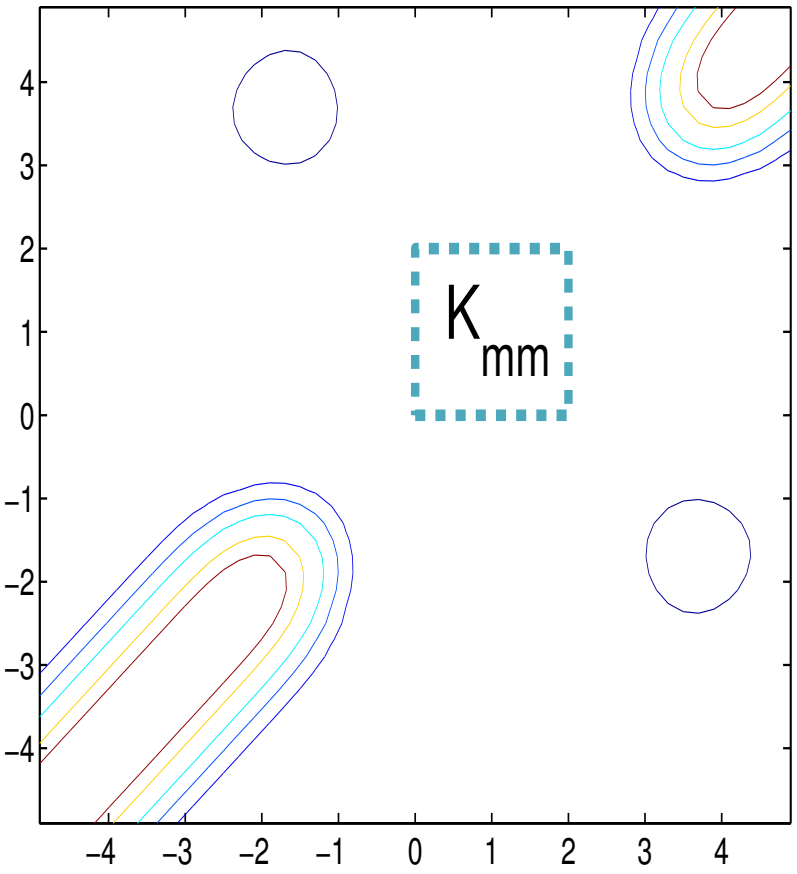
$$p_{\mathcal{BV}}(\boldsymbol{f}) = (\mu_{\mathcal{BV}}, K_{\mathcal{BV}}, \mathcal{BV})$$

- Probabilistic framework;
- Selection of optimal $\mathcal{BV}$ set via the "score" for $\boldsymbol{x} \in \mathcal{BV}$.
- Independent of the noise/likelihood model.

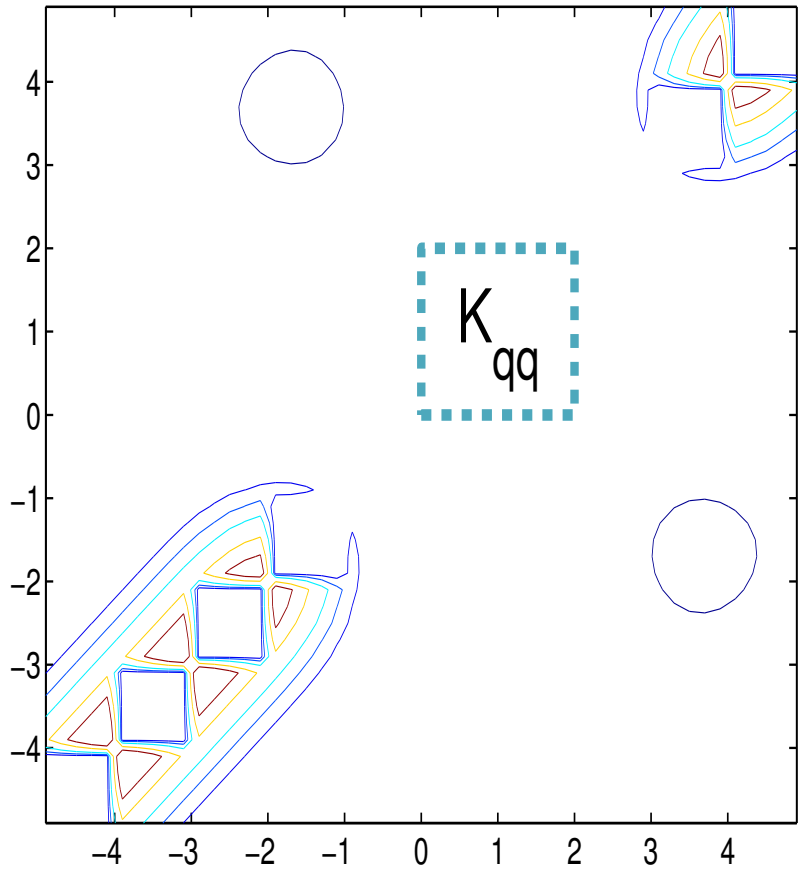
- **Sparse posterior variance shrinked.**
- Shrinkage – result of conditioning **but** is this good?
- A larger variance would probably be better.

Comparisons

GP approximations $\Longleftrightarrow$ viewed as approx to K_{NN}



Nystrom,
Sparse KL $\equiv$ PPA



Bayesian Committee,
Snelson & Ghahramani

Sparse M.L.II.

Sparse $\mathcal{GP}$'s
Lehel Csató
Notations
Gaussian processes
Likelihoods
Approximations
Expectation propagation
Sparsification
Sparsity results
Comparisons
Examples
Predictive densities
Regression
Summary

Optimising hyperparameters using

1. The log-Evidence is approximated as

$$\log \mathrm{Ev} = \sum_n \log Z_n - \sum \log \hat{Z}_n + \log \int d\boldsymbol{f}_{\mathcal{BV}} p_0(\boldsymbol{f}_{\mathcal{BV}}) \prod_n \hat{t}(\pi_n \boldsymbol{f}_{\mathcal{BV}})$$

2. The upper bound to the log-evid (EM)

$$\log \mathrm{Ev} = \int d\boldsymbol{f}_{\mathcal{BV}} p_{\mathcal{BV}}(\boldsymbol{f}) \log \; p_0(\boldsymbol{f}_{\mathcal{BV}}) P(\mathcal{D}|\Pi \boldsymbol{f}_{\mathcal{BV}})$$

The Evidence is better approximated using 1
but *the same test errors (class)*.

Inference Method

EM–EP:

Iterate the following steps

- Fix parameters $\Rightarrow$ *approximate the posterior*
- Fix posterior $\Rightarrow$ *new model parameters*.

Efficiency

Sparse $\mathcal{GP}$'s

Lehel Csató

Notations
Gaussian processes
Likelihoods
Approximations
Expectation propagation
Sparsification
Sparsity results
Comparisons
Examples
Predictive densities
Regression
Summary

Efficiency:

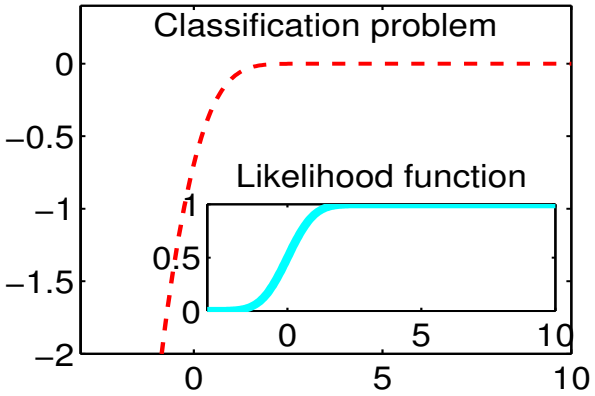
- selecting a “good subset”.
- KL-based selection – depends on the (KL–)loss.
- Can be inefficient for different problems.
- Measures related to the *loss function* – see IVM

Predictive densities for different models

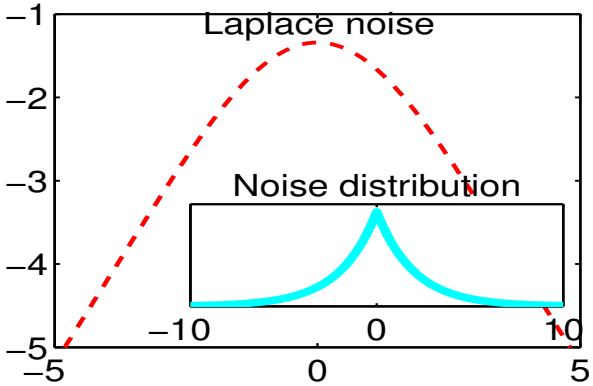
Sparse $\mathcal{GP}$'s
Lehel Csató
Notations
Gaussian processes
Likelihoods
Approximations
Expectation propagation
Sparsification
Sparsity results
Comparisons
Examples
Predictive densities
Regression
Summary

- Gaussian likelihood involves no approximation

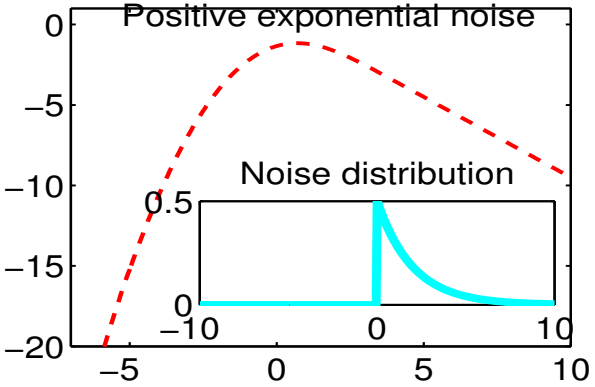
Classification



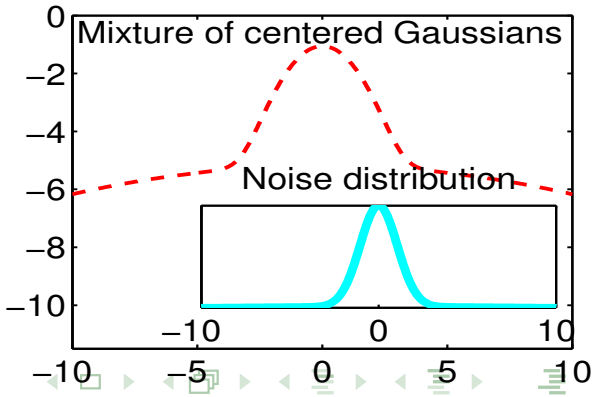
Laplace – robust



Pos. exponential



Mixture



Regression example

Sparse $\mathcal{GP}$'s

Lehel Csató

Notations
Gaussian processes
Likelihoods

Approximations
Expectation propagation
Sparsification
Sparsity results
Comparisons

Examples
Predictive densities
Regression

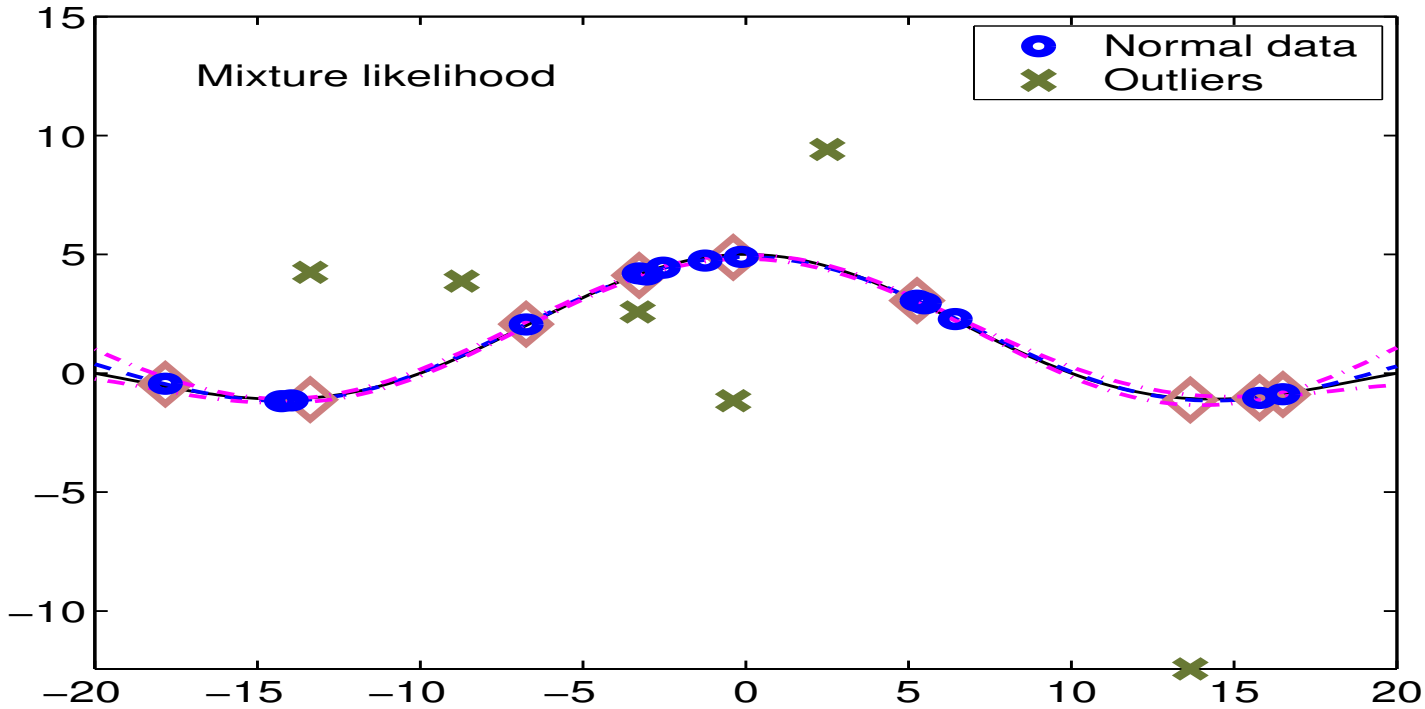
Summary

Detecting outliers using a mixture model:

$$P(y_n|f_n,\boldsymbol{\theta}) = \pi \mathrm{N}(y_n|f_n,\sigma_1^2) + (1-\pi)\mathrm{N}(y_n|f_n,\sigma_O^2)$$

Likelihood *not log-concave*.

π, σ_1^2, and σ_O^2 estimated using MLII.



Concerns

- Subset selection important for GP's
- Sparsification based on KL-divergence might not be better
- Approximations – bounds on the Evidence – speed up computation.
- Selection criterion *should be based* on model loss function or score.

Summary

- It is possible to infer the hyperparameters.
- Sparse approximation speeds up computation without significant loss.

- Outlook
  - Extension to two-level model specification.
  - Dynamical systems.

Software (matlab) and documentation available:

http://www.tuebingen.mpg.de/∼csatol