How priors of initial hyperparameters affect Gaussian Process regression model

Zexun Chen

Department of Mathematics
University of Leicester

zc74@le.ac.uk

August 1, 2015

Overview

1. Problems and Targets
2. Brief Introduction of Gaussian Process
   - GP and GPR
   - Kernels
   - Hyperparameters
3. Sensitivity of Initial Hypers
4. Priors
5. Experiment
   - Data&Method
   - GP samples
   - ARMA(2,1) samples
6. Conclusion

Problems

- Gaussian Process Regression (GPR) is a kernel-based non-parametric method
- GPR relies on kernel selection over undetermined hyperparameters (Hypers)
- Undetermined Hypers usually have to be estimated by maximum marginal likelihood using Conjugate Gradient (CG)
- Marginal likelihoods are not always convex and CG cannot guarantee global optima
- Consequently, the performance of GPR model might be affected by optimization.

Targets

A common strategy adopted by most researchers using GPR model is to randomly select some initial Hypers using their expert opinions and experiences. So the problem is how to randomly select initial Hypers.

As a result, the aims of this study are as follow,

- Introduce Gaussian Process Regression model
- Introduce kernels over undetermined Hypers
- Put different Priors on undetermined Hypers
- Qualify performance of GPR model

Gaussian Process

Definition (Gaussian Process)

For any set S, a Gaussian Process(GP) on S is a set of random variable $(f_t, t \in S)$, $s.t. \forall n \in \mathbb{N}, \forall t_1, \ldots, t_n \in S, (f_{t_1}, \ldots, f_{t_n})$ is (multivariate) Gaussian

A GP is fully specified by a mean function and covariance function.

Theorem (Gaussian Process)

For any set S, any mean function $\mu : S \mapsto \mathbb{R}$ and any covariance function $k : S \times S \mapsto \mathbb{R}$, there exists a GP f_t on S, s.t., $\mathbb{E}[f(t)] = \mu(t)$, $Cov(f(s), f(t)) = k(s,t), \forall s, t \in S$. It denotes $f(t) \sim \mathcal{GP}(\mu, k)$.

A collection of functions $[f(t_1), f(t_2), \ldots, f(t_n)]$ are realized through a multivariate normal distribution

$$[f(t_1), f(t_2), \ldots, f(t_n)]^{\mathrm{T}} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$$

where $\mu_i = \mu(t_i)$ and $K_{ij} = k(t_i, t_j)$.

Gaussian Process Regression

- Consider a noisy model $y = f(x) + \varepsilon$, where f is drawn from a GP and ε is Gaussian noise. Then, we wish to predict new observations $\mathbf{y}_*$ for new test points X_*.
- The joint distribution of training observations $\mathbf{y}$ and test data $\mathbf{y}_*$ is

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(X) \\ \mu(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X_*,X)^{\mathrm{T}} \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right),$$

from which, the predictive distribution is trivially

$$\begin{aligned} p(\mathbf{y}_* | X, \mathbf{y}, X_*) &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}), \\ \textit{where} \quad \hat{\boldsymbol{\mu}} &= K(X_*,X)^{\mathrm{T}}(K(X,X) + \sigma_n^2 I)^{-1}(\mathbf{y} - \mu(X)), & (1) \\ \hat{\boldsymbol{\Sigma}} &= K(X_*,X_*) - K(X_*,X)^{\mathrm{T}}(K(X,X) + \sigma_n^2 I)^{-1} K(X_*,X). & (2) \end{aligned}$$

- The marginal likelihood is defined by:$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f}$
- Model parameters (e.g.,length scale of covariance function) can be optimized to maximize the marginal likelihood

Basic Kernels

Kernel (also called covariance function) is the most important part of GPR model. Firstly, there are some basic kernels below,

- Linear(LIN), $k_{LIN}(x, x') = x \cdot x'$.
- Squared Exponential(SE), $k_{SE} = s_f^2 \exp(-\frac{(x-x')^2}{2\ell^2})$,where s_f is output-scale amplitude and ℓ is length scale of input.
- Periodic(PER), $k_{PER}(x, x') = s_f^2 \exp(-\frac{2\sin^2(\pi\frac{(x-x')}{p})}{\ell^2})$, where p is period and the other parameters are the same as SE.
- Rational Quadratic(RQ), $k_{RQ}(x, x') = s_f^2(1 + \frac{(x-x')^2}{2\alpha\ell^2})^{-\alpha}$, where α is known as the index and the other parameters are the same as SE. If $\alpha \to \infty$, the RQ kernel changes to SE.

Composite Kernels

Local Periodic(LP)

$$k_{LP} = k_{SE} \times k_{PER}.$$

Spectral Mixture(SM)

$$k_{SM}(\mathbf{x}, \mathbf{x}') = \sum_{q=1}^{Q} w_q \prod_{p=1}^{P} \exp(-2\pi^2 (x_p - x'_p)^2 \nu_q^{(p)}) \cos(2\pi (x_p - x'_p) \mu_q^{(p)}),$$

where the qth component has mean $\boldsymbol{\mu}_q = (\mu_q^{(1)}, \ldots, \mu_q^{(P)})$ and kernel $\mathbf{K}_q = diag(\nu_q^{(1)}, \ldots, \mu_q^{(P)})$. The inverse means $1/\mu_q$ are the component periods and inverse standard deviations $1/\sqrt{\nu_q}$ are length scales. More details about SM kernel are in [Wilson, 2014].

Hyperparameters

In Bayesian statistics, a hyperparameter is a parameter of a prior distribution of model. In GPR model, prior is Gaussian and Hypers are those undetermined parameters of kernels.
To recall model, marginal likelihood is also Gaussian,

$$p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0},\Sigma_{\boldsymbol{\theta}}), \tag{3}$$

where $\Sigma_\theta = K + \sigma_n^2 I$ and $\boldsymbol{\theta}$ are Hypers. Then the log marginal likelihood and the partial derivatives (Matrix Derivative) are below,

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\Sigma_\theta^{-1}\mathbf{y} - \frac{1}{2}\log\det\Sigma_\theta - \frac{n}{2}\log 2\pi \tag{4}$$

$$\frac{\partial}{\partial\theta_i}\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}\mathrm{tr}(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}) + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}\mathbf{y} \tag{5}$$

Sensitivity of Initial Hypers

- Eq.(4) shows that log marginal likelihood is not always convex
- Non-convex optimization cannot guarantee global optima
- Local optima may suffer sensitivity of initial Hypers

A common strategy adopted by most researchers using GPR is to randomly select a few initial Hypers using their expert opinions and experiences,for example [Wilson, 2014] and then compare them to achieve the best.

So the problem is how to randomly select a few initial Hypers.

- Select initial Hypers from uniform (0,1) (Mainstream method)
- Put different prior distributions on initial Hypers (My work)

Non-informative Priors

- Prior 1, $\theta_i \sim \mathrm{Uniform}(0, 1)$
- Prior 2, $\log(\theta_i) \sim \mathrm{Uniform}(-1, 1)$
- Prior 3, $\log(\theta_i) \sim \mathrm{Uniform}(-10, 10)$
- Prior 4, $\theta_i \sim \mathcal{N}(0, 10000)$
- Prior 5, $\frac{\pi}{\theta_i} \sim \mathrm{Uniform}(0, 1)$
- Prior 6, $\log(\frac{\pi}{\theta_i}) \sim \mathrm{Uniform}(-5, 5)$

Data-dominated Priors

- Prior 7, $\theta_i \sim \mathrm{Uniform}(0, \mathrm{Nyq})$, where Nyq from Nyquist frequency,equals 0.5 divided the minimum distance of inputs.
- Prior 8, $\frac{1}{\sqrt{\theta_i}} \sim \mathcal{TN}(0, \infty)$, where $\mathcal{TN}(0, \infty)$ is truncated standard normal distribution bounded above 0.
- Prior 9, $\frac{\pi}{\theta_i} \sim \mathrm{Uniform}(\frac{1}{\mathrm{MaxI}}, \mathrm{Nyq})$, where MaxI means the max distance of input data.

Remark

Prior 1,2,3 are used for SE kernel and Prior 1,5,6,7,9 are suitable for parameter period of PER/LP kernel. For SM kernel, there are two parameters from Prior 5,6,7,9 and Prior 1,8 respectively.

Data&Method

- Data, GPs first and then ARMA(2,1)
- Prediction, in-sample & out-of-sample model first and then only out-of-sample model
- Model, GPR first and then compare to ARMA true model
- Kernels, SE & PER first and then LP &SM
- Performance, RMSE and SMSE are to qualify GPR model

Remark

Root mean square error (RMSE) $= (\frac{1}{N}\sum(y_{est} - y_{actual})^2)^{\frac{1}{2}}$

Standardized mean square error(SMSE) $= \frac{\sum(y_{est} - y_{actual})^2}{Varanice(y_{actual})} = \frac{\sum(y_{est} - y_{actual})^2}{\sum(y_{actual} - \bar{y}_{actual})^2}$

Results of GP samples (SE)

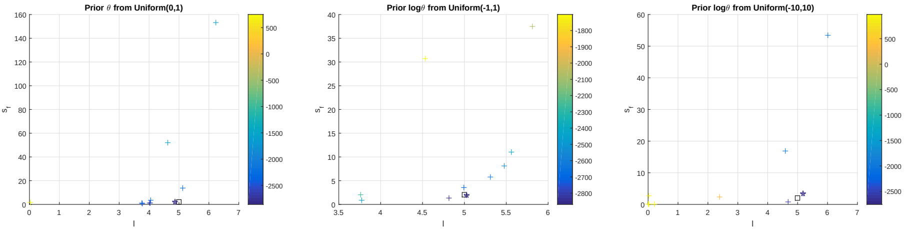

Figure 1: Position of hyperparameters from Prior 1,2,3

- Different color : negative log marginal likelihood
- The squared signal: actual parameters,
- The pentagram signal: finial optimized parameters
- Plus signals: Optimized parameters in process of optimization.

Results of GP samples (SE)

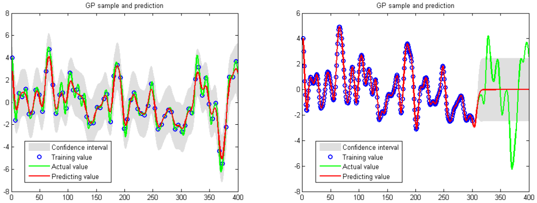

Figure 2: In-sample & Out-of-sample GPR,SE kernel parameters $[\ell, s_f] = [5, 2]$

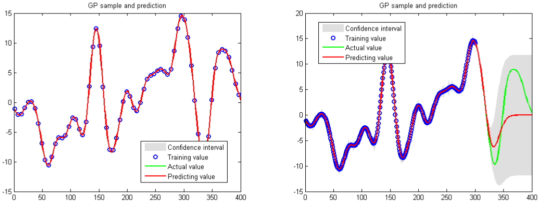

Figure 3: In-sample & Out-of-sample GPR,SE kernel parameters $[\ell, s_f] = [15, 7]$

Results of GP samples (SE)

Table 1: 20 samples' results from prior 1 over SE kernel

		In-Sample			Out-of-Sample		
$\boldsymbol{\theta}_{act}$		$\boldsymbol{\theta}_{final}$	RMSE	SMSE	$\boldsymbol{\theta}_{final}$	RMSE	SMSE
ℓ	5	4.74 ∓ 2.256	7.29E-02	1.71E+01	4.91 ∓ 0.085	1.82E-01	8.97E+01
s_f	2	1.93 ∓ 0.204			1.49 ∓ 0.370		
ℓ	15	11.47 ∓ 6.126	9.37E-05	1.25E-05	14.83 ∓ 0.325	1.66E-01	4.82E+01
s_f	7	6.77 ∓ 1.074			5.63 ∓ 1.632		

Remark

The actual parameters of model are in the range of 20 sample's final parameters.

Results of GP samples (PER)

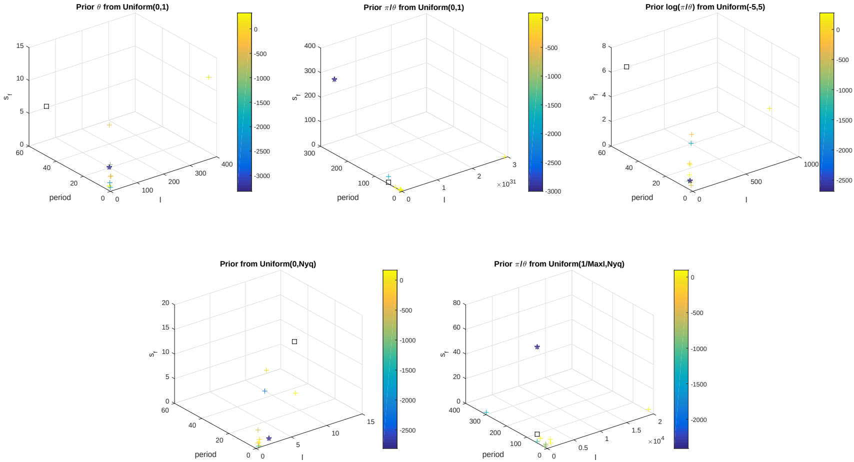

Figure 4: Position of hyperparameters from Prior 1,5,6,7,9

Results of GP samples (PER)

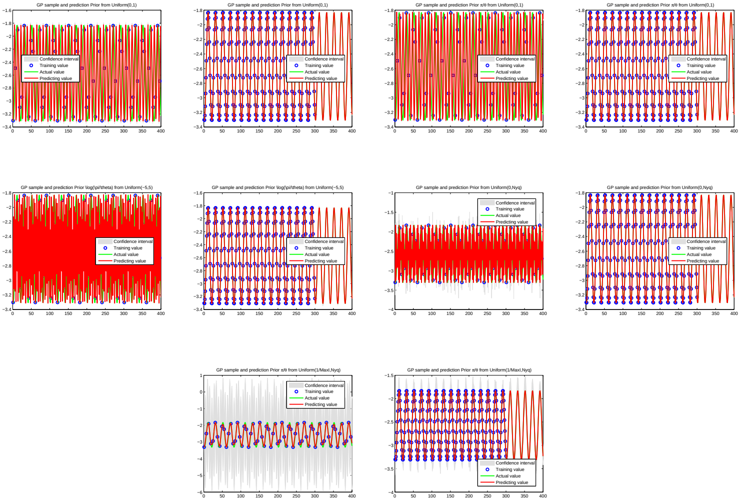

Figure 5: In&Out-of-sample GPR,PER kernel,prior from Prior 1,5,6,7,9

Results of GP samples (PER)

Table 2: 20 samples' results from prior 1,5,6,7,9 over PER kernel

			In-Sample			Out-of-Sample		
Prior	$\boldsymbol{\theta}_{act}$		$\boldsymbol{\theta}_{final}$	RMSE	SMSE	$\boldsymbol{\theta}_{final}$	RMSE	SMSE
Prior 1	ℓ	15	2.80	1.22E-01	7.96E+02	1.18	1.53E-06	3.30E-08
	period	50	0.56			0.70		
	s_f	7	6.45			2.54		
Prior 5	ℓ	15	4.98	1.41E-01	7.97E+02	1.64	1.81E-06	1.40E-08
	period	50	4.92			20.00		
	s_f	7	6.75			6.61		
Prior 6	ℓ	15	5.30	1.39E-01	7.96E+02	0.84	2.56E-06	1.79E-07
	period	50	2.17			2.08		
	s_f	7	10.67			3.88		
Prior 7	ℓ	15	6.74	1.29E-01	7.95E+02	1.05	1.92E-05	1.02E-05
	period	50	0.05			0.20		
	s_f	7	14.37			2.93		
Prior 9	ℓ	15	2.35	7.30E-03	1.67E+00	6.74	8.60E-07	1.24E-08
	period	50	60.00			20.00		
	s_f	7	53.81			31.81		

Results of ARMA samples (LP kernel)

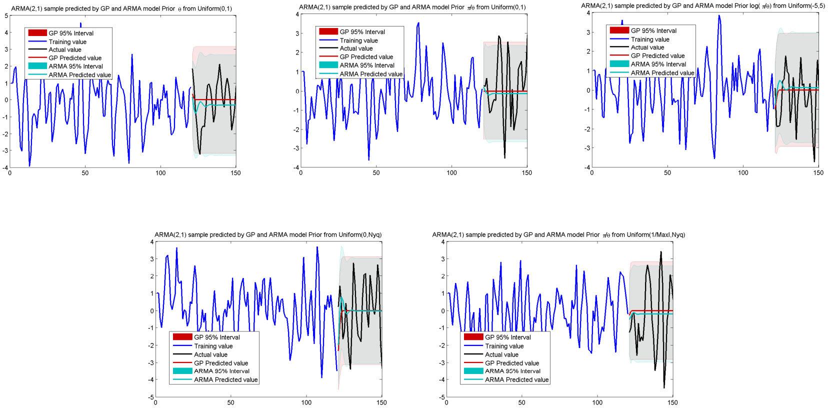

Figure 6: ARMA&GPR model,LP kernel,prior from Prior 1,5,6,7,9

Results of ARMA samples (LP kernel)

Table 3: 20 ARMA samples' results from different priors over LP kernel

	GPR model (LP kernel)		ARMA(2,1)	
	RMSE	SMSE	RMSE	SMSE
Prior 1	1.3726	30.9021	1.3657	30.7245
Prior 5	1.4443	29.9021	1.4301	29.3195
Prior 6	1.5503	30.2927	1.5549	30.5043
Prior 7	1.3772	29.9573	1.3804	30.2064
Prior 9	1.4423	30.4163	1.4514	30.8121

Results of ARMA samples (SM kernel)

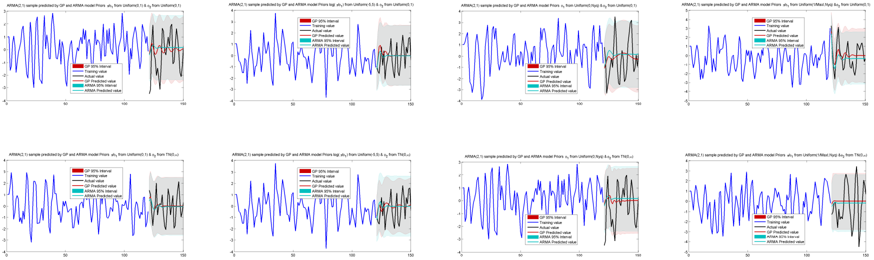

Figure 7: ARMA&GPR model, SM kernel,prior from Prior PS51, PS61, PS71, PS91, PS58, PS68, PS78, PS98

Results of ARMA samples (SM kernel)

Table 4: 20 ARMA samples' results from different priors over SM kernel

GPR model (SM kernel)			ARMA(2,1)	
Priors	RMSE	SMSE	RMSE	SMSE
PS51	1.5022	30.3764	1.4995	30.2384
PS61	1.5283	29.8368	1.5063	29.0083
PS71	1.5650	33.4866	1.4968	30.1307
PS91	1.4054	28.6271	1.4213	29.2684
PS58	1.4924	29.9930	1.4995	30.2384
PS68	1.5150	30.7773	1.5012	30.1248
PS78	1.5216	31.3949	1.5098	31.4075
PS98	1.4982	30.2374	1.4991	30.2396

Conclusion

Results:

- For SE kernel, the prior distribution of initial Hypers doesn't affect optimized Hypers, neither does the performance of GPR
- For PER kernel, the prior distribution of initial Hypers affect optimized Hypers seriously, but it doesn't affect the performance of GPR.
- For LP and SM kernel on ARMA(2,1) samples, the performance of GPR model isn't affected seriously by initial Hypers' priors and GPR model performs as good as ARMA true model.
- If kernel choice is wrong, no matter what priors of initial Hypers we choose are useless

References

Wilson A G. Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes[D]. PhD thesis, University of Cambridge, 2014.

Rasmussen C E. Gaussian processes for machine learning[J]. 2006.

Thanks to my supervisor:Prof.Gorban and Dr.Wang

Thanks to everyone

The End