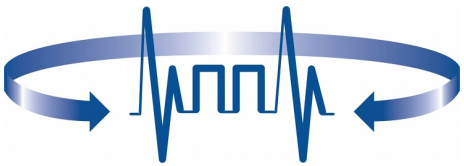
Gaussian processes regression for variable fidelity data

Alexey Zaytsev,
aleksey.zaitsev@phystech.edu

IITP RAS

«Bayesian methods in Machine Learning»,
MSU, 9 October, 2015



DATADVANCE
AN AIRBUS GROUP COMPANY

Approximation problem

- The given sample is $\boldsymbol{D} = (X, \mathbf{y}) = \{\mathbf{x}_i, y(\mathbf{x}_i) = y_i\}_{i=1}^n$, $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^d$.
- We want to construct a model $\widehat{y}(\mathbf{x})$ for the unknown function $y(\mathbf{x})$, such that

$$\widehat{y}(\mathbf{x}) \approx y(\mathbf{x}),$$

and *provide uncertainty estimation* $\widehat{\sigma}^2(\mathbf{x})$ for the prediction $\widehat{y}(\mathbf{x})$ (to construct confidence intervals).

Kernel ridge regression

- Suppose we have a positive-definite symmetric kernel $k(\mathbf{x}, \mathbf{x}')$.
- We consider the space $\mathcal{H}_k$ of functions $f(\mathbf{x})$

$$f(\mathbf{x}) = \sum_m w_m k(\mathbf{x}, \mathbf{x}_m) = \sum_{i=1}^{\infty} c_i \phi_i(\mathbf{x}),$$

  with functions $\phi_i(\mathbf{x})$ are from the eigen-decomposition of the kernel $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \gamma_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$.
- This space of functions $\mathcal{H}_k$ is a reproducing kernel Hilbert space (RKHS).
- The function $\widehat{y}(\mathbf{x})$ is a solution of the problem

$$\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|^2_{\mathcal{H}_k} \to \min_{f(\mathbf{x}) \in \mathcal{H}_k},$$

  where $\|f\|^2_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \frac{c_i^2}{\gamma_i}$.

Kernel ridge regression

- It turns out that there is an equivalent finite-dimension problem with $K = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$:

$$(\mathbf{y} - K\mathbf{w})^\top (\mathbf{y} - K\mathbf{w}) + \lambda \mathbf{w}^T K \mathbf{w} \to \min_{\mathbf{w}}.$$

- The solution has the form

$$\mathbf{w} = (K + \lambda I_n)^{-1}\mathbf{y}.$$

- The final approximation $\widehat{y}(\mathbf{x})$ has the form

$$\widehat{y}(\mathbf{x}) = \sum_{i=1}^n w_i k(\mathbf{x}, \mathbf{x}_i) = \mathbf{k}(\mathbf{x})^\top (K + \lambda I_n)^{-1}\mathbf{y},$$

where $\mathbf{k}(\mathbf{x}) = \{k(\mathbf{x}, \mathbf{x}_i)\}_{i=1}^n$.

Examples of $\mathcal{H}_k$ spaces

- Polynomial regression with penalty:

$$k(\mathbf{x}, \mathbf{z}) = 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2.$$

- Radial basis function with the kernel width θ:

$$k(\mathbf{x}, \mathbf{z}) = \exp(-\theta\|\mathbf{x} - \mathbf{z}\|_2^2).$$

- Support vector machine with loss function $\sum_{i=1}^n [1 - y_i f(\mathbf{x}_i)]_+ + \frac{1}{2}\mathbf{w}^T K \mathbf{w}$:

$$f(\mathbf{x}) = w_0 + \sum_{i=1}^n w_i k(\mathbf{x}, \mathbf{x}_i).$$

Problems with kernel ridge regression

- How to select the kernel function and the kernel width θ?
- How to select regularization λ?
- What is about uncertainty estimation (construction of confidence intervals)?

One way to solve these problems is to use Bayesian-driven *Gaussian processes regression*.

Gaussian processes

- We suppose that $y(\mathbf{x})$ is a realization of Gaussian process.
- Gaussian process is a random process for which any finite dimensional slice of the process is a random vector with a multivariate Gaussian distribution.
- To completely specify Gaussian process one has to specify its *mean* and *covariance function*.

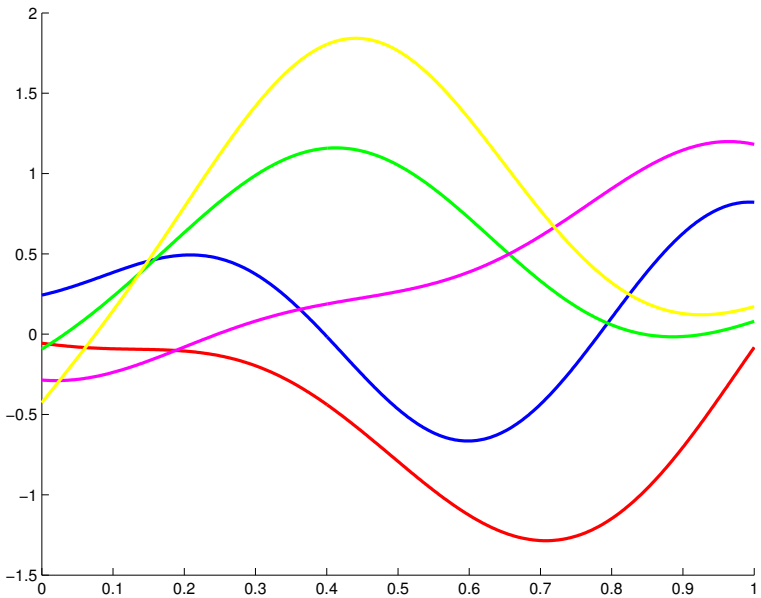
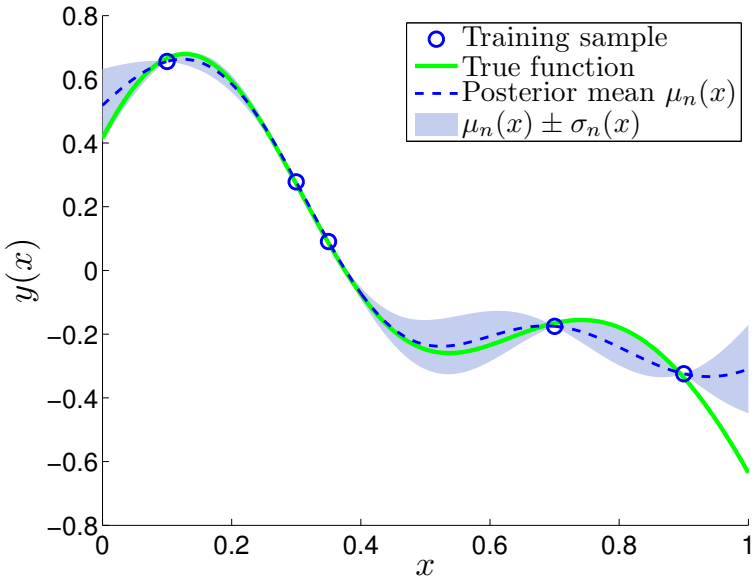


Figure : Gaussian process realizations (one-dimensional $\mathbf{x}$)

Gaussian processes regression

- The sample is
  $\boldsymbol{D} = (X, \mathbf{y}) = \{\mathbf{x}_i, y(\mathbf{x}_i) = y_i\}_{i=1}^n$,
  $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^d$.
- Suppose that
  - the function $y(\mathbf{x})$ is a realization of a Gaussian process,
  - *mean* of the Gaussian process equals zero,
  - *covariance function* has the form $\operatorname{cov}(y(\mathbf{x}), y(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}')$.
- For these assumptions joint distribution of $\mathbf{y}$ is Gaussian:
  $\mathbf{y} \propto \mathcal{N}(\mathbf{0}, K)$,
  $K = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.



Gaussian processes regression

- Posterior distribution $y(\mathbf{x})$ at $\mathbf{x}$ is Gaussian:

$$\mathrm{Law}(y(\mathbf{x})|\boldsymbol{D}) = \mathcal{N}(\mu_n(\mathbf{x}), \sigma_n^2(\mathbf{x})).$$

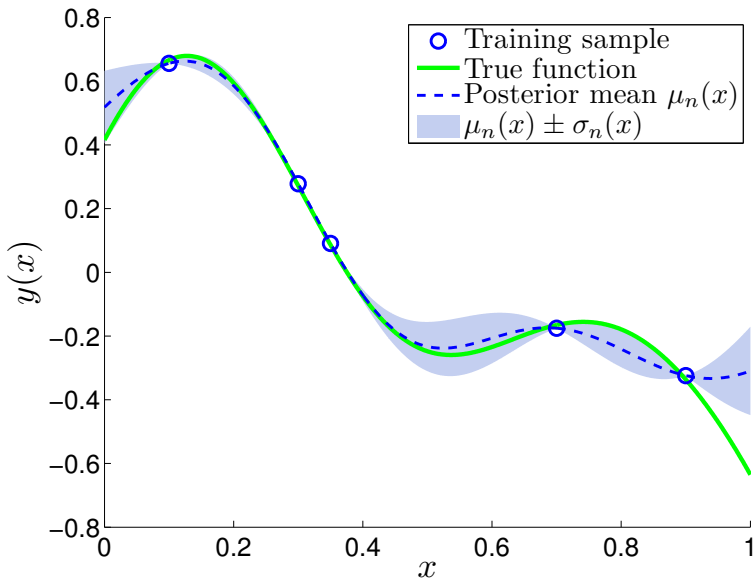
- Posterior mean $\mu_n(\mathbf{x})$ equals

$$\widehat{y}(\mathbf{x}) = \mu_n(\mathbf{x}) = \mathbf{k}^\top(\mathbf{x}) K^{-1} \mathbf{y}.$$

- Posterior variance $\sigma_n^2(\mathbf{x})$ equals

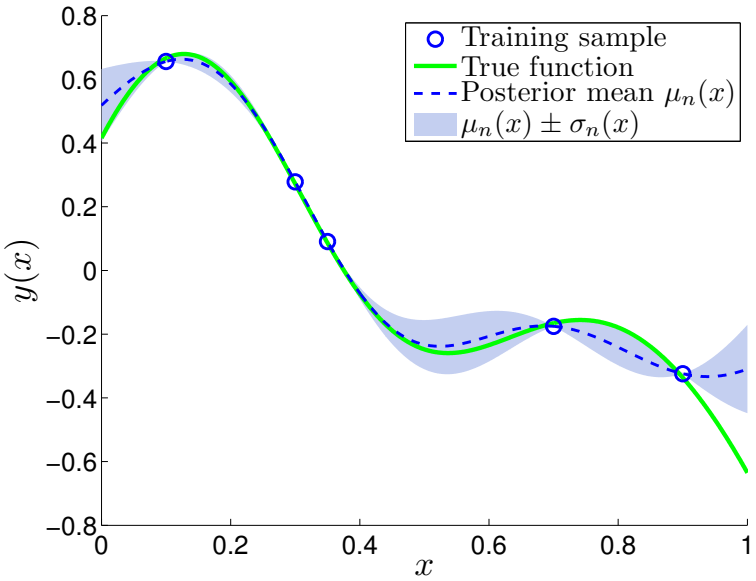
$$\sigma_n^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^\top(\mathbf{x}) K^{-1} \mathbf{k}(\mathbf{x}).$$

- Here $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n))^\top$.



Parametric assumption for Gaussian processes regression

- *Parametric assumption*: covariance function $k(\mathbf{x}, \mathbf{x}')$ coincides with $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ for some $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, and covariance matrix has the form $K_{\boldsymbol{\theta}} = \{k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.
- Parametric assumption is *possibly wrong*.

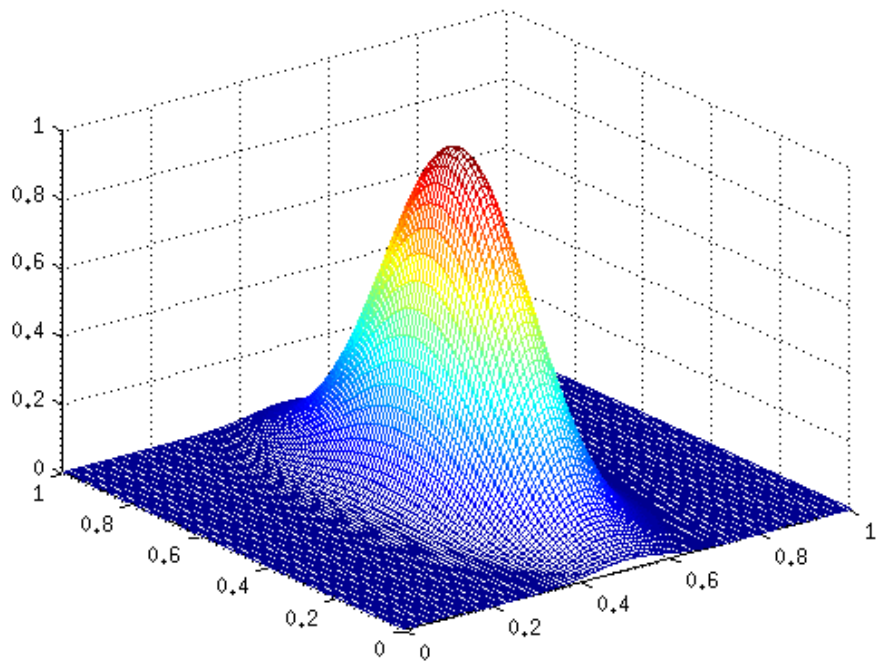


Example of covariance function

- Squared exponential covariance function

$$k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \theta_0^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \theta_i^2 (x_i - x_i')^2\right) + \sigma^2 \delta(\mathbf{x} - \mathbf{x}'),$$

here $\delta(\cdot)$ is Kroneker delta function. Number of parameters is $d + 2$.



Other covariance functions

There are other covariance functions:

- *A non-stationary covariance-based Kriging method for metamodelling in engineering design* by Y. Xiong, W. Chen, et al., *JNME*, 2007
- *Matern Cross-Covariance Functions for Multivariate Random Fields* by T. Gneiting, W.K. Leiber, M.S. Chlather, *JASA*, 2011
- *Additive Gaussian processes* by D. Duvenaud, H. Nickisch, C.E. Rasmussen, *arXiv preprint*, 2013

Estimation of parameters $\boldsymbol{\theta}$ of covariance function

- We need to estimate the vector of parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$.
- Maximum likelihood approach is a popular choice for estimation of $\boldsymbol{\theta}$:

$$\widetilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} L(\boldsymbol{\theta}),$$

where the logarithm of the likelihood has the form:

$$L(\boldsymbol{\theta}) = -\frac{1}{2}\left[n \log 2\pi + \log\det\left(K_{\boldsymbol{\theta}}\right) + \mathbf{y}^\top K_{\boldsymbol{\theta}}^{-1}\mathbf{y}\right].$$

- We hope that $\widetilde{\boldsymbol{\theta}}$ is close in some sense to the central point $\boldsymbol{\theta}^*$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \mathbb{E} L(\boldsymbol{\theta}),$$

where $\mathbb{E}$ is the mean with respect to $\mathcal{N}(\mathbf{0}, K)$.

Possibly wrong parametric assumption

- Gaussian process regression allows one to model a broad class of functions $y(\mathbf{x})$.
- However, parameter assumption is always wrong, i.e. true covariance function, as the true covariance function $k(\mathbf{x}, \mathbf{x}')$ doesn't belong to $\{k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'), \boldsymbol{\theta} \in \Theta\}$.
- In this case $\boldsymbol{\theta}^*$ minimizes Kullback–Leibler divergence between true distribution and distributions generated by $\boldsymbol{\theta} \in \Theta$

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathrm{KL}(\mathcal{N}(\mathbf{0}, K) | \mathcal{N}(\mathbf{0}, K_{\boldsymbol{\theta}})),$$

i.e. $\boldsymbol{\theta}^*$ is the best parametric fit.

Bayesian approach

- Let us select a prior $\Pi(d\boldsymbol{\theta})$ for the vector of parameters $\boldsymbol{\theta}$.
- Posterior distribution of $\boldsymbol{\theta}$ has the form

$$\mathrm{Law}\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{D}\right) \propto \exp\left\{L(\boldsymbol{\theta})\right\} \Pi(d\boldsymbol{\theta}).$$

- In this case $\overline{\boldsymbol{\theta}} \stackrel{\mathrm{def}}{=} I\!\!E\left\{\boldsymbol{\theta} \,\middle|\, \boldsymbol{D}\right\}$ is a reasonable estimate of $\boldsymbol{\theta}^*$, where $I\!\!E$ is a mean with respect to $\mathrm{Law}\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{D}\right)$.
- In this work we consider the case of *noninformative prior distribution* $\Pi(d\boldsymbol{\theta})$.

Computational limitations of Gaussian processes regression

- To calculate likelihood and posterior mean we need to inverse covariance matrix of size $n \times n$ — complexity is $O(n^3)$!
- The case of large samples $n \gtrsim 5000$ requires some kind of approximation.
- There exist some:
  - *Sparse Gaussian processes using pseudo-inputs*, E. Snelson and Z. Ghahramani, *NIPS*, 2005.
  - *On the Nyström method for approximating a Gram matrix for improved kernel-based learning*, P. Drineas and M.W. Mahoney, *JMLR*, 2006.
  - *Covariance tapering for interpolation of large spatial datasets*, R. Furrer, M.G. Genton and D. Nychka, *JCGS*, 2006.
  - *Gaussian processes for big data through stochastic variational inference*, J. Hensman, U.K. Sheffield, N. Fusi and N.D. Lawrence, *NIPS*, 2012.

State of the art. Theory

A number of papers consider asymptotic normality and consistency of estimates of covariance function parameters for Gaussian processes regression.

Author	Year	Estimate	
Mardia	1984	MLE	
Kaufman	2008	MLE	covariance tapering
Shabi	2010	MLE, Bayesian	covariance tapering
Chu	2011	MLE with penality	with and without covariance tapering

Problems to solve

State of the art papers consider the case:

- sample size $n \to \infty$ *(asymptotic results)*,
- parametric assumption holds *(true parametric assumption)*,
- covariance tapering (we suppose that $k(\mathbf{x}, \mathbf{x}') = 0$ if $\|\mathbf{x} - \mathbf{x}'\| > d$ for some d).

While, we need results for the case:

- MLE and Bayesian estimates;
- sample size is finite *(nonasymptotic results)*;
- parametric assumption can be wrong *(possibly wrong parametric assumption)*.

Regression for variable fidelity data

- A low fidelity sample is $D_l = (X_l, \mathbf{y}_l) = \left\{\mathbf{x}_i^l, y_l(\mathbf{x}_i^l)\right\}_{i=1}^{n_l}$, and a high fidelity sample is $D_h = (X_h, \mathbf{y}_h) = \left\{\mathbf{x}_i^h, y_h(\mathbf{x}_i^h)\right\}_{i=1}^{n_h}$ with $\mathbf{x}_i^l, \mathbf{x}_i^h \in \mathbb{R}^d$, $y_l(\mathbf{x}), y_h(\mathbf{x}) \in \mathbb{R}$.
- The low fidelity function $y_l(\mathbf{x})$ and the high fidelity function $y_h(\mathbf{x})$ model the same physical processes, but with different fidelity.
- Using both the low and the high fidelity samples we need to construct a regression model $\widehat{y}_h(\mathbf{x}) \approx y_h(\mathbf{x})$ of the high fidelity function and uncertainty estimation for it.

Examples of variable fidelity data

- $y_h(\mathbf{x})$ is a high fidelity function, $y_l(\mathbf{x})$ is a low fidelity function, $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$, $y_l \in \mathbb{R}$, $y_h \in \mathbb{R}$.

Low fidelity	High fidelity
CFD with coarse mesh	CFD with dense mesh
Full potential equations for CFD	Euler equations for CFD
Numerical experiments	Nature experiments
Noised data	Noise-free data

Example: surrogate model construction for an airfoil

- A lift coefficient C_l and a drag coefficient C_d describe quality of an airfoil.
- Fast evaluation of an airfoil quality is crucial to optimization of an airfoil.
- We can evaluate C_l and C_d using low and high fidelity solvers.

Equations	Euler	Full potentials
Fidelity	High	Low
Time (CPU), seconds	600	10
Sample size	100	10000

Gaussian processes regression for variable fidelity data

- At the moment there is no algorithm that can proceed large samples of variable fidelity data.
- *It is relevant* to construct an algorithm of surrogate model construction based on Gaussian processes regression for variable fidelity data.
- The algorithm should not be *ad hoc*.

Problems at hand

Problem 1 Provide theoretical foundation for Bayesian approach in Gaussian processes regression.

Problem 2 Develop an algorithm that can proceed large samples of variable fidelity data.

Problem 3 Access applicability of the presented approaches to real problems.

Problem 1: theoretical foundation for Bayesian approach in Gaussian processes regression

- Construction of Gaussian processes regression model equivalent to estimation of covariance function parameters.
- *It is relevant* to provide theoretical properties for covariance function parameters estimates based on Bayesian and MLE approaches.
- Obtained results should hold for the case of *finite sample sizes* and *possibly wrong parametric assumption*. The problem has not been solved yet for this statement.

Classic version of Bernstein-von Mises theorem (BvM)

- Specify some prior for parameters.
- For some regularity assumptions it holds that

$$\mathrm{Law}\left(\overline{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{D}\right) \approx \mathcal{N}\left(\widetilde{\boldsymbol{\theta}}, I\left(\widetilde{\boldsymbol{\theta}}\right)^{-1}\right) \text{ for } n \to \infty$$

for MLE $\widetilde{\boldsymbol{\theta}}$ and Bayesian estimate $\overline{\boldsymbol{\theta}}$ in terms of total variation distance. Fisher information $I\left(\widetilde{\boldsymbol{\theta}}\right)$ has the form:

$$I\left(\widetilde{\boldsymbol{\theta}}\right) = \left\{ \mathbb{E} \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j} \right\}_{i,j=1}^{p} \Bigg|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}}.$$

- BvM theorem advocates usage of Bayesian approach for point and interval estimates from probabilistic point of view.

Notation

- Spectral norm of vector $\|A\|_2 = \sigma_{\max}(A)$.
- ∞ norm of vector $\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|$.
- Euclidian norm of vector $\mathbf{a} = \{a_1, \ldots, a_p\}$ $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{p} a_i^2}$.
- Matrices D_0^2 and V_0^2 are analogues of Fisher information for possibly wrong parametric assumption:

$$D_0^2 = -\nabla^2 I\!\!E L(\boldsymbol{\theta}^*), \qquad V_0^2 = \mathrm{Var}\big\{\nabla L(\boldsymbol{\theta}^*)\big\}.$$

- If the parametric assumption is correct, then $D_0^2 = V_0^2 = I(\boldsymbol{\theta}^*)$.

Notation

- Posterior mean $\overline{\boldsymbol{\theta}}$ is Bayesian estimate of central point $\boldsymbol{\theta}^*$:

$$\overline{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \mathbb{E}\{\boldsymbol{\theta} \,|\, \boldsymbol{D}\}.$$

- Posterior variance has the form:

$$S^2 \stackrel{\text{def}}{=} \mathbb{E}\left\{(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})\,(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})^\top \,|\, \boldsymbol{D}\right\}.$$

- We consider a vicinity of the central point $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}L(\boldsymbol{\theta})$:

$$\Theta_0(\mathtt{r}_0) = \{\boldsymbol{\theta} \in \Theta : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}_0\}.$$

- In results below all constants have explicit form, while we drop this form for clarity and brevity.

Bernstein-von Mises theorem

Statement

Let assumptions (A1)–(A6) holds, ans sample size n (for fixed $C > 0$) is greater than:

$$n \geq 4C\mathtt{r}_0^2 p^3.$$

Then there exists a random set $\Omega(\mathtt{x})$ with probability at least $1 - 5e^{-\mathtt{x}}$, such that

$$\mathbb{P}(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)) \leq 3e^{-\mathtt{x}},$$
$$\mathbb{P}(\overline{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)) \leq 5e^{-\mathtt{x}}.$$

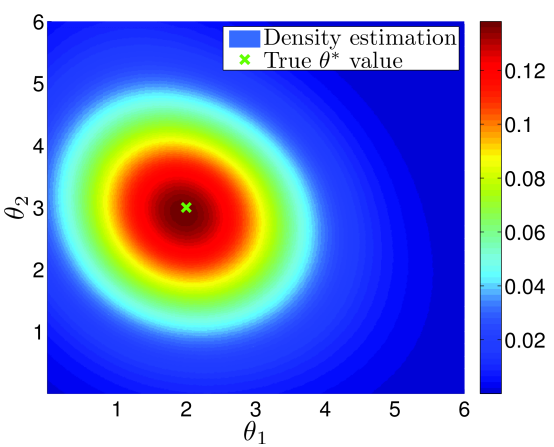
- Maximum likelihood estimate (MLE) $\widetilde{\boldsymbol{\theta}}$ and Bayesian estimate $\overline{\boldsymbol{\theta}}$ are close to the central point $\boldsymbol{\theta}^*$.

Distribution of $\widetilde{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\theta}}$

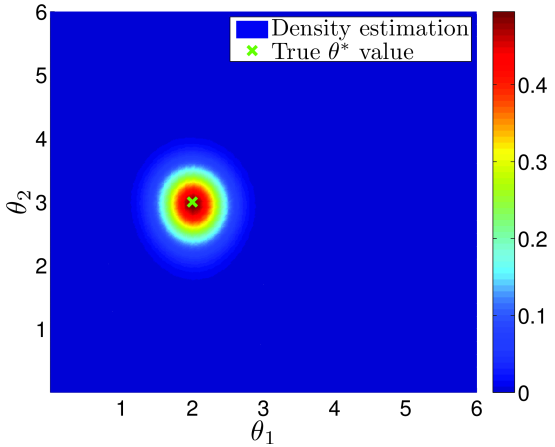
- Generate 400 samples $\boldsymbol{D}_i$ of Gaussian processes realizations.
- For each sample $\boldsymbol{D}_i$ obtain MLE $\widetilde{\boldsymbol{\theta}}_i$ and Bayesian estimate $\overline{\boldsymbol{\theta}}_i$.
- Using kernel density estimation (KDE) estimate density of $\widetilde{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\theta}}$ using obtained values $\{\widetilde{\boldsymbol{\theta}}_i\}_{i=1}^{400}$ and $\{\overline{\boldsymbol{\theta}}_i\}_{i=1}^{400}$.

Obtained estimates of $\widetilde{\boldsymbol{\theta}}$, parameters dimension $p = 2$.

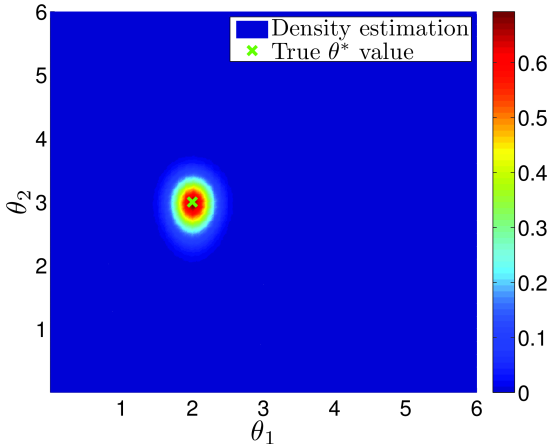
- We obtain kernel density estimate for $\widetilde{\boldsymbol{\theta}}$ using samples $\boldsymbol{D}_i$, $i = \overline{1, 400}$.
- While sample size n increases, estimates $\widetilde{\boldsymbol{\theta}}$ concentrate in the vicinity of the true value $\theta^* = [2, 3]$.



(a) $n = 10$



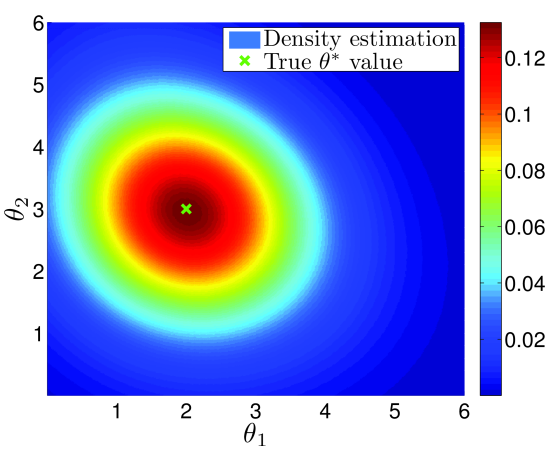
(b) $n = 200$



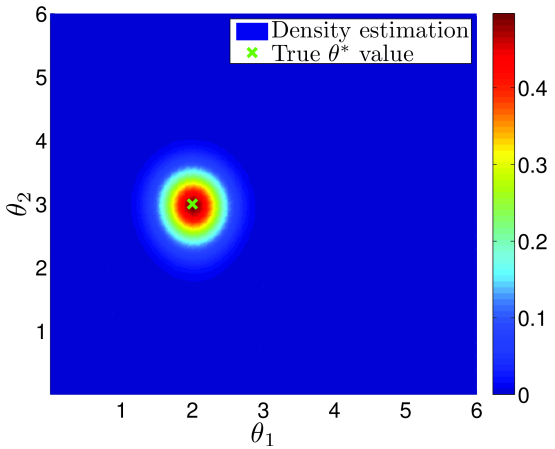
(c) $n = 1000$

Obtained estimates of $\overline{\boldsymbol{\theta}}$, parameters dimension $p = 2$.

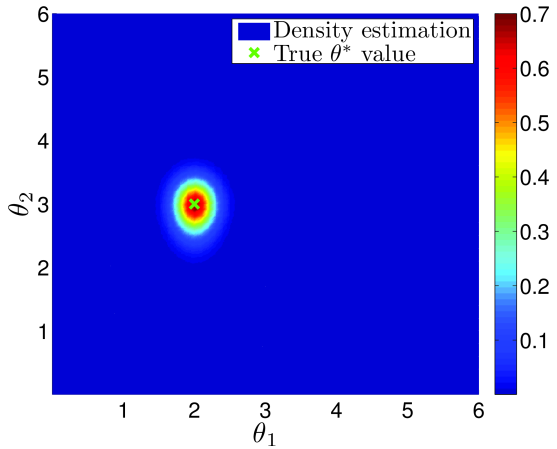
- We obtain kernel density estimate for $\overline{\boldsymbol{\theta}}$ using samples $\boldsymbol{D}_i$, $i = \overline{1, 400}$.
- While sample size n increases, estimates $\widetilde{\boldsymbol{\theta}}$ concentrate in the vicinity of the true value $\theta^* = [2, 3]$.



(a) $n = 10$



(b) $n = 200$



(c) $n = 1000$

Bernstein-von Mises theorem

Statement

There exists a constant $\diamondsuit(\mathtt{r}_0, \mathtt{x}) \le \frac{\diamondsuit_0(\mathtt{r}_0, \mathtt{x})}{\sqrt{n}}$, *such that for* $\Delta_\circ = \mathtt{r}_0 \diamondsuit(\mathtt{r}_0, \mathtt{x})$ *with probability at least* $1 - 5e^{-\mathtt{x}}$ *it holds:*

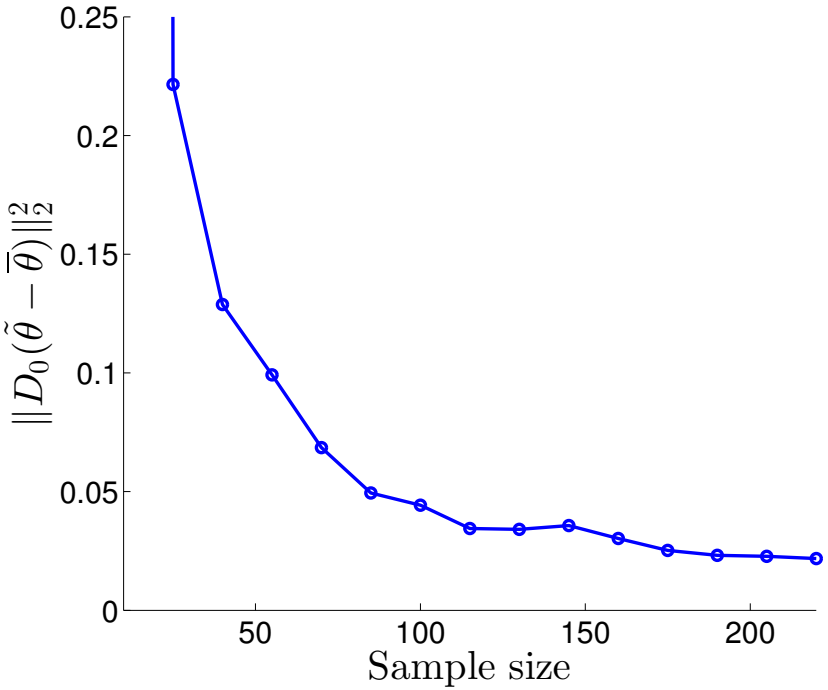
$$\|D_0(\overline{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})\|^2 \le 4\Delta_\circ(\mathtt{x}) + \diamondsuit^2(\mathtt{r}_0, \mathtt{x}) + 4e^{-\mathtt{x}},$$
$$\|I_p - D_0 S^2 D_0\|_2 \le 4\Delta_\circ(\mathtt{x}) + 4e^{-\mathtt{x}}.$$

- Bayesian estimate converges to MLE with speed $\frac{1}{\sqrt{n}}$.
- Mean value $\overline{\boldsymbol{\theta}}$ of posterior distribution is close to MLE (Maximum Likelihood estimate) $\widetilde{\boldsymbol{\theta}}$.
- Covariance matrix of posterior distribution of parameters S^2 is close to the matrix D_0^{-2}.

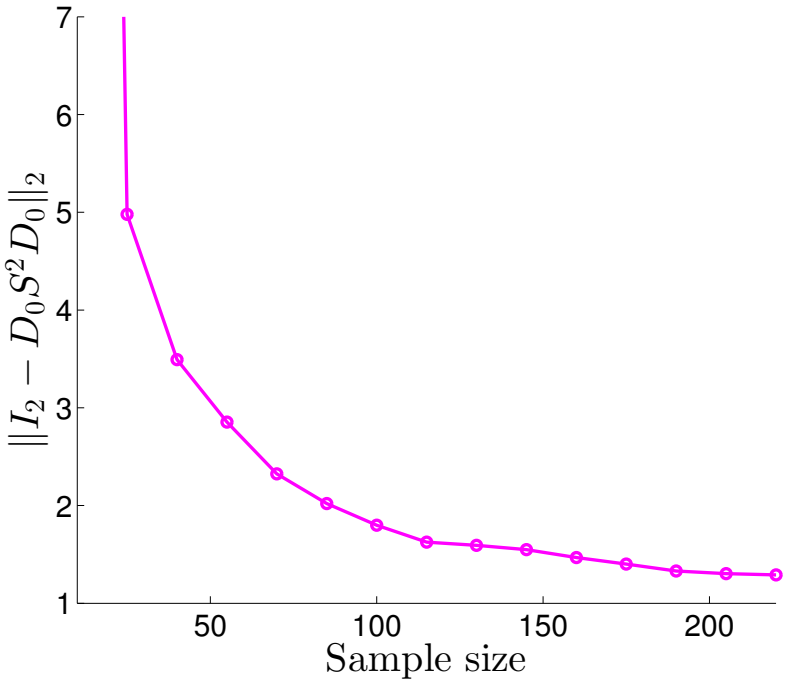
Statement of the theorem, $p = 2$.

$$\|D_0(\overline{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})\|^2 \leq 4\Delta_\circ(\mathtt{x}) + \diamondsuit^2(\mathtt{r}_0, \mathtt{x}) + 4e^{-\mathtt{x}},$$

$$\|I_p - D_0 S^2 D_0\|_2 \leq 4\Delta_\circ(\mathtt{x}) + 4e^{-\mathtt{x}}.$$



(a) Values $\left\|D_0\left(\widetilde{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}}\right)\right\|_2^2$ for the sample sizes n



(b) Values $\left\|I_p - D_0 S^2 D_0\right\|_\infty$ for the sample sizes n

Bernstein-von Mises theorem

Statement

For any measurable set $A \subset \mathbb{R}^p$ *and* $\boldsymbol{\gamma} \propto \mathcal{N}(\mathbf{0}, I_p)$ *with probability at least* $1 - 5e^{-\mathtt{x}}$ *it holds that:*

$$\mathbb{P}\left(D_0(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) \in A | \boldsymbol{D}\right) \geq$$
$$\geq \exp\left\{-2\Delta_\circ(\mathtt{x}) - 3e^{-\mathtt{x}}\right\}\left(\mathbb{P}(\boldsymbol{\gamma} \in A) - \diamondsuit(\mathtt{r}_0, \mathtt{x})\right) - e^{-\mathtt{x}},$$
$$\mathbb{P}\left(D_0(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) \in A | \boldsymbol{D}\right) \leq$$
$$\leq \exp\left\{2\Delta_\circ(\mathtt{x}) + 2e^{-\mathtt{x}}\right\}\left(\mathbb{P}(\boldsymbol{\gamma} \in A) + \diamondsuit(\mathtt{r}_0, \mathtt{x})\right) + e^{-\mathtt{x}}.$$

- Posterior distribution of parameters is close to Gaussian distribution with respect to Total variance distance.

Normality of posterior distribution of parameters $\boldsymbol{\theta}$

- We consider Total variance and Hellinger distances between posterior distribution of parameters and Gaussian distribution with the same mean and covariance matrix.
- For distributions with probability densities $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ total variance distance is

$$TV(p,q) = \frac{1}{2}\int_{\mathbb{R}^p} |p(\boldsymbol{\theta}) - q(\boldsymbol{\theta})| d\boldsymbol{\theta},$$

and Hellinger distance is

$$H(p,q) = \frac{1}{2}\int_{\mathbb{R}^p} \left(\sqrt{p(\boldsymbol{\theta})} - \sqrt{q(\boldsymbol{\theta})}\right)^2 d\boldsymbol{\theta}.$$

- To calculate these distances we use numerical integration. Then we average obtained distances for 200 runs.

Normality of posterior distribution of parameters θ

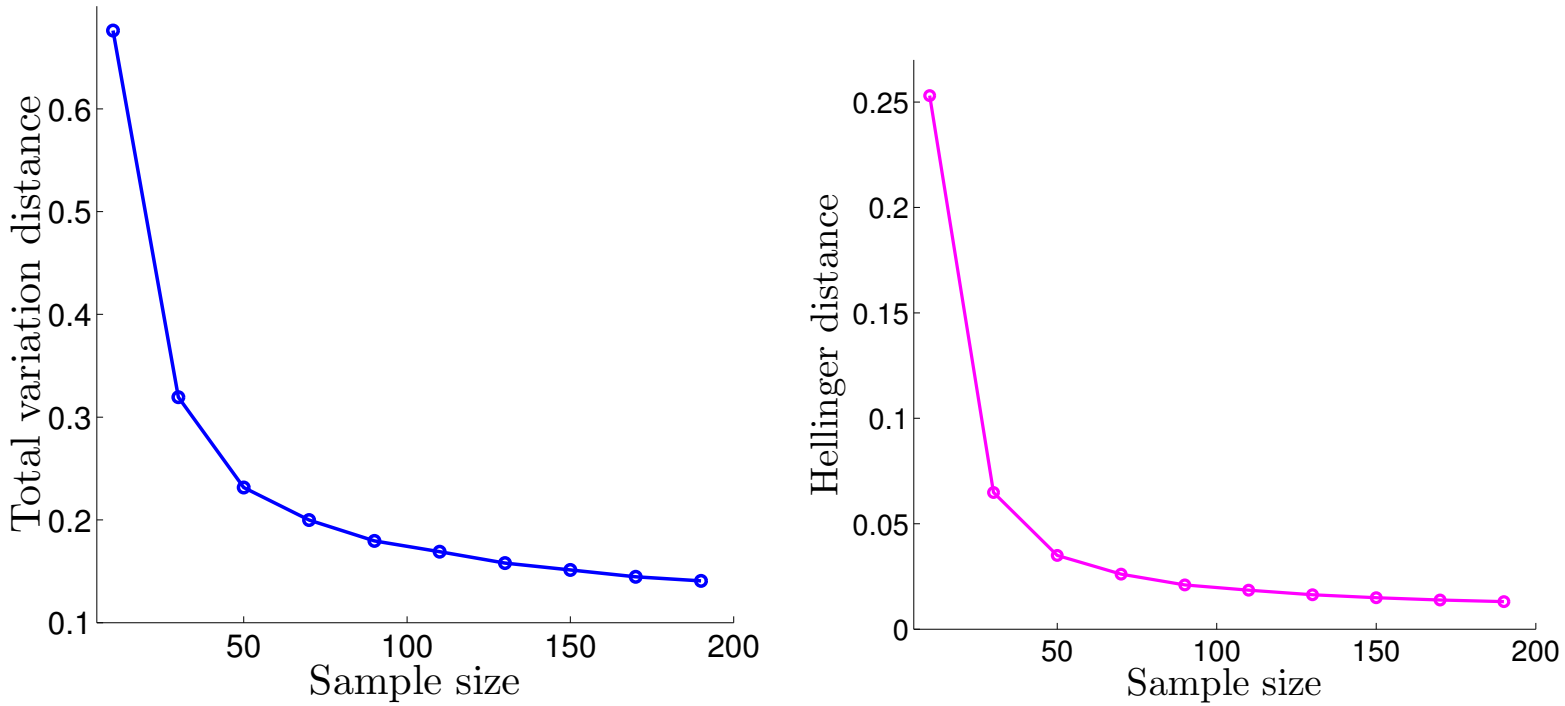


Figure : Total variance and Hellinger distances between the posterior distribution of parameters and Gaussian distribution with the same mean and covariance matrix, $p = 2$

Bernstein-von Mises theorem: results

- MLE and Bayesian estimate are close to the central point.
- MLE and Bayesian estimate approaches the central point with speed $\frac{1}{\sqrt{n}}$.
- Bayesian estimate $\overline{\boldsymbol{\theta}}$ is close to MLE $\widetilde{\boldsymbol{\theta}}$.
- Covariance matrix of posterior distribution S^2 is close to the matrix D_0^{-2}.
- The posterior distribution is close to Gaussian in terms of Total variance distance.

Idea of proof

- Idea of proof is to make upper bounds for exponential moments of logarithms of likelihood and its derivatives using introduced assumptions.
- Using obtained upper bounds for exponential moments of loglikelihood, apply modern empirical process theory to get the theorem.

Assumptions about the covariance function (A1)–(A3)

(A1) The central point $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} I\!\!E L(\boldsymbol{\theta})$ exists.

(A2) $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ is three times continuously differentiable with respect to $\boldsymbol{\theta} \in \Theta$ for $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$,

(A3) There exist constants $0 < \overline{\lambda} < \infty$ and $0 < \lambda_0 < \infty$, such that $\|K\|_2 \leq \overline{\lambda}, \|K_{\boldsymbol{\theta}}\|_2 \leq \overline{\lambda}$, $\|K^{-1}\|_2 \leq \lambda_0, \|K_{\boldsymbol{\theta}}^{-1}\|_2 \leq \lambda_0^{-1}$. for $\boldsymbol{\theta} \in \Theta$,

Assumptions about the covariance function (A4)–(A6)

(A4) There exist constants $0 < \lambda_1, \lambda_2, \lambda_3 < \infty$, such that $\left\| \frac{\partial K_{\boldsymbol{\theta}}}{\partial \theta_i} \right\|_2 \leq \lambda_1$, $\left\| \frac{\partial^2 K_{\boldsymbol{\theta}}}{\partial \theta_i \partial \theta_j} \right\|_2 \leq \lambda_2$, $\left\| \frac{\partial^3 K_{\boldsymbol{\theta}}}{\partial \theta_i \partial \theta_j \partial \theta_k} \right\|_2 \leq \lambda_3$ for $\boldsymbol{\theta} \in \Theta$ and all $i, j, k \in \{1, \ldots, p\}$.

(A5) There exists sufficiently small constant $C > 0$, such that $\frac{\lambda_i}{\lambda_0} < C$ for $i = 1, 2, 3$.

(A6) The smallest eigenvalue of the matrix $\frac{1}{n} D_0^2$ is greater than $\mathrm{d}_0 > 0$, and the smallest eigenvalue of the matrix $\frac{1}{n} V_0^2$ is greater than $\mathrm{v}_0 > 0$, where $D_0^2 = -\nabla^2 I\!\!E L(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ and $V_0^2 = \mathrm{Var}\{\nabla L(\boldsymbol{\theta})\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$.

Introduced assumptions are close to that used in the article by Shaby (2012).

Example of covariance function that satisfies the assumptions

Statement

Squared exponential covariance function of the form

$$k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \theta_1^2 \left(\exp\left(-\theta_2^2 \|\mathbf{x} - \mathbf{x}'\|_2^2 \right) + \sigma^2 \delta(\|\mathbf{x} - \mathbf{x}'\|_2^2) \right)$$

satisfies introduced assumptions for suitable design X and noise variance $\sigma^2 > 0$ and if parametric assumption holds.

Conclusions

- We introduced Bernstein-von Mises theorem for estimates of covariance function parameters for Gaussian processes regression
- Our Bernstein-von Mises theorem holds for finite sample size and possibly wrong parametric assumption.
  - Posterior mean is close to MLE estimate, and thus is close to the central point.
  - Posterior covariance matrix is close to D_0^{-2}.
  - Posterior distribution of parameters is close to Gaussian distribution with respect to Total variance distance.
- There exists a widely used covariance function with parameters space dimension $p > 1$, for which introduced assumptions hold.

Problem 2: regression for variable fidelity data

- A low fidelity sample is $D_l = (X_l, \mathbf{y}_l) = \left\{\mathbf{x}_i^l, y_l(\mathbf{x}_i^l)\right\}_{i=1}^{n_l}$, and a high fidelity sample is $D_h = (X_h, \mathbf{y}_h) = \left\{\mathbf{x}_i^h, y_h(\mathbf{x}_i^h)\right\}_{i=1}^{n_h}$ with $\mathbf{x}_i^l, \mathbf{x}_i^h \in \mathbb{R}^d$, $y_l(\mathbf{x}), y_h(\mathbf{x}) \in \mathbb{R}$.
- The low fidelity function$y_l(\mathbf{x})$ and the high fidelity function $y_h(\mathbf{x})$ model the same physical processes, but with different fidelity.
- Using both the low and the high fidelity samples we need to construct a regression model $\widehat{y}_h(\mathbf{x}) \approx y_h(\mathbf{x})$ of the high fidelity function and uncertainty estimation for it.

Gaussian processes regression model for mutifidelity data

- We use a common cokriging model [Forrester, 2007]:

$$y_l(\mathbf{x}) = f_l(\mathbf{x}) + \varepsilon_l,\ y_h(\mathbf{x}) = \rho y_l(\mathbf{x}) + y_d(\mathbf{x}),$$

$y_d(\mathbf{x}) = f_d(\mathbf{x}) + \varepsilon_d$.

- $f_l(\mathbf{x})$ is a Gaussian process with covariance function $k_l(\mathbf{x}, \mathbf{x}')$, a realization of this process is the low fidelity function.
- $f_d(\mathbf{x})$ is a Gaussian process with covariance function $k_d(\mathbf{x}, \mathbf{x}')$, a realization of this process is difference between low and high fidelity function multiplied by ρ.
- ε_l, ε_d are Gaussian white noises with variances σ_l^2 and σ_d^2, correspondingly.

Here $f_l(\mathbf{x})$, $f_d(\mathbf{x})$ are independent Gaussian process with zero means and covariance functions $k_l(\mathbf{x}, \mathbf{x}')$ and $k_d(\mathbf{x}, \mathbf{x}')$ correspondingly.

Prediction using Gaussian processes regression model for variable fidelity data

- Define $X = \begin{pmatrix} X_l \\ X_h \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} \mathbf{y}_l \\ \mathbf{y}_h \end{pmatrix}$.
- Posterior mean for the high fidelity function value at new points X^* has the form:

$$\widehat{\mathbf{y}}_h(X^*) = K(X^*, X)K^{-1}\mathbf{y},$$

- Posterior covariance matrix has the form:

$$\begin{aligned}\mathbb{V}(X^*) = \rho^2 K_l(X^*, X^*) + K_d(X^*, X^*) - \\ - K(X^*, X)K^{-1}\left(K(X^*, X)\right)^T.\end{aligned}$$

Notation for covariance matrices

$$K(X^*, X) = \begin{pmatrix} \rho K_l(X^*, X_l) \\ \rho^2 K_l(X^*, X_h) + K_d(X^*, X_h) \end{pmatrix},$$

$$K(X, X) = \begin{pmatrix} K_l(X_l, X_l) & \rho K_l(X_l, X_h) \\ \rho K_l(X_h, X_l) & \rho^2 K_l(X_h, X_h) + K_d(X_h, X_h) \end{pmatrix},$$

$K_l(X_a, X_b)$, $K_d(X_a, X_b)$ are matrices of pairwise covariances of $y_l(\mathbf{x})$ and $y_d(\mathbf{x})$ for points from sets X_a and X_b correspondingly.

Covariance functions parameters estimation

We use the following procedure to estimate parameters of $f_l(\mathbf{x})$ and $f_d(\mathbf{x})$ [Forrester, 2007]:

1. Estimate parameters of covariance function $k_l(\mathbf{x}, \mathbf{x})$, using common algorithm for Gaussian processes regression for the sample $D = D_l$.
2. Calculate differences between posterior means $\widehat{y}_l(\mathbf{x})$ for Gaussian process $y_l(\mathbf{x})$ and $\mathbf{x} \in X_h$.
3. Estimate parameters of Gaussian process $y_d(\mathbf{x})$ with covariance function $k_d(\mathbf{x}, \mathbf{x}')$ and parameter ρ by maximization of posterior density for $D = D_{\text{diff}} = (X_h, \mathbf{y}_d = \mathbf{y}_h - \rho\widehat{\mathbf{y}}_l(X_h))$.
4. Complexities of parameters estimation and posterior mean evaluation are $O(n^3)$, where $n = n_l + n_h$.

Base points subsample

- Size of subsample of base points $n_1 = n_h^1 + n_l^1$ is such, that we can do casual inference for the basic points subsample
- Fix subsample of *base* points from initial sample
  $D_1 = (X^1, \mathbf{y}^1)$, $X^1 = \begin{pmatrix} X_l^1 \\ X_h^1 \end{pmatrix}, \mathbf{y}^1 = \begin{pmatrix} \mathbf{y}_l(X_l^1) \\ \mathbf{y}_h(X_h^1) \end{pmatrix}$
- Now we can use Nystrom approximation for covariance matrices.

Nystrom approximation for covariance matrices

We use subsample of base points and the following matrices:

$$K_{11} = \begin{pmatrix} K_l(X_l^1, X_l^1) & \rho K_l(X_l^1, X_h^1) \\ \rho K_l(X_h^1, X_l^1) & \rho^2 K_l(X_h^1, X_h^1) + K_d(X_h^1, X_h^1) \end{pmatrix},$$

$$K_1 = \begin{pmatrix} K_l(X_l^1, X_l) & \rho K_l(X_l^1, X_h) \\ \rho K_l(X_h^1, X_l) & \rho^2 K_l(X_h^1, X_h) + K_d(X_h^1, X_h) \end{pmatrix},$$

$$K_1^* = \begin{pmatrix} \rho K_l(X^*, X_l^1) \\ \rho^2 K_l(X^*, X_h^1) + K_d(X^*, X_h^1). \end{pmatrix}$$

Then for new points $X^* = \{\mathbf{x}_i^*\}_{i=1}^{n^*}$ we get approximation of the form $K(X^*, X)$, K and $K(X^*, X^*)$:

$$\widehat{K}(X^*, X\ ) = K_1^* K_{11}^{-1} K_1,\ \widehat{K} = (K_1)^T K_{11}^{-1} K_1,$$
$$\widehat{K}(X^*, X^*) = K_1^* K_{11}^{-1} (K_1^*)^T.$$

Prediction using Nystrom approximation

Define

- $$R = \begin{pmatrix} \frac{1}{\sigma_l} I_{n_l} & 0 \\ 0 & \frac{1}{\sqrt{\rho^2 \sigma_l^2 + \sigma_d^2}} I_{n_h} \end{pmatrix},$$

  where I_k is a unity matrix of size k;
- $C_1 = RK_1$;
- V_{11} is the Cholesky decomposition of matrix K_{11};
- $V = C_1 V_{11}^{-T}$.

Statement

For posterior mean using Nystrom approximation we get:

$$\widehat{\mathbf{y}}_h(X^*) \approx K_1^* V_{11} (I_{n_1} + V^T V)^{-1} V^T \mathbf{y}.$$

Uncertainty of prediction using Nystrom approximation

Use also the following approximation $k(\mathbf{x}^*, X) \approx K_1^* K_{11}^{-1} K_1^T$, $k(X, X) \approx R^{-2} + K_1 K_{11}^{-1} K_1^T$.

Statement

For posterior variance Nystrom approximation has the form:

$$\widehat{\sigma}^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - K_1^* V_{11}^{-1} (I + V^T V)^{-1} (V^T V) V_{11}^{-T} K_1^{*T}.$$

Statement

Computational complexities of posterior mean and variances calculation are $O(nn_1^2)$.

Approach to proceed variable fidelity data with a blackbox

- Suppose there is blackbox for the low fidelity function $y_l(\mathbf{x})$, so we can get low fidelity function value for any point from the design space $\mathbb{X} \subseteq \mathbb{R}^d$.
- Then we can include in regression model the low fidelity function value at the target point $\mathbf{x}$.

- Add to the sample a pair of point and the low fidelity function value at this point.
- Then prediction and uncertainty estimation for this prediction has the form:

$$\widehat{y}_h^{\text{exp}}(\mathbf{x}) = \mathbf{k}_{\text{exp}} K_{\text{exp}}^{-1} \mathbf{y}_{\text{exp}},$$
$$\mathbb{V}_{\text{exp}}(\mathbf{x}) = \rho^2 K_l(\mathbf{x}, \mathbf{x}) + K_d(\mathbf{x}, \mathbf{x}) - \mathbf{k}_{\text{exp}}^T K_{\text{exp}}^{-1} \mathbf{k}_{\text{exp}}.$$

Statement

Complexity of the Cholesky decomposition update is $O(n^2)$.

Problem 3: Access applicability of the presented approaches to real problems

We compare the following approaches:

- GP — Gaussian processes regression,
- VFGP — Gaussian processes regression for variable fidelity data,
- **SVFGP** — Sparse Gaussian processes regression for variable fidelity data,
- **BB VFGP** — Gaussian processes regression for variable fidelity data with low fidelity function blackbox available,

We estimate quality of models using RRMS and a test sample $D_{\text{test}} = \{\mathbf{x}_i^{\text{test}}, y_i^{\text{test}} = f_h(\mathbf{x}_i^{\text{test}})\}_{i=1}^{n_t}$:

$$RRMS(D_{\text{test}}, \widehat{y}) = \frac{\sum_{i=1}^{n_t} (\widehat{y}_h(\mathbf{x}_i^{\text{test}}) - y_i^{\text{test}})^2}{\sum_{i=1}^{n_t} (\overline{y} - y_i^{\text{test}})^2},$$

Performance of BB VFGP for artificial data

We use the following high and low fidelity functions:

$$y_h(x) = (6x-2)^2 \sin(12x-4),$$
$$y_l(x) = 0.5 y_h(x) + 10(x-1).$$

n_h	6	15	30
GP	0.7102	0.0159	$3.83e-04$
VFGP	0.3036	$7.42e-04$	$1.38e-04$
BB VFGP	0.1610	$6.90e-07$	$1.67e-07$

Table : RRMS for different sample sizes of the high fidelity sample n_h

Performance of SVFGP and BB VFGP for artificial data

Let high fidelity and low fidelity functions be (ε_h, ε_l are Gaussian white noise with variance 0.001, 0.002):

$$y_h(\mathbf{x}) = 20 + \sum_{i=1}^{d}(x_i^2 - 10\cos(2\pi x_i)) + \varepsilon_h,\ \mathbf{x} \in [0,1]^d,$$

$$y_l(\mathbf{x}) = y_h(\mathbf{x}) + 0.2\sum_{i=1}^{d}(x_i + 1)^2 + \varepsilon_l,\ \mathbf{x} \in [0,1]^d.$$

n_l	1000	3000	5000
VFGP	30.46	852.70	7283.27
SVFGP	30.46	33.42	37.50
BB VFGP	30.38	842.97	7672.60

Table : Training times (in seconds) for VFGP, SVFGP, and BB VFGP.

Performance of SVFGP and BB VFGP for artificial data

n_l	1000	3000	5000
VFGP	0.0502	0.0170	0.0058
SVFGP	0.0502	0.0305	0.0260
BB VFGP	0.0010	0.00029	0.00017

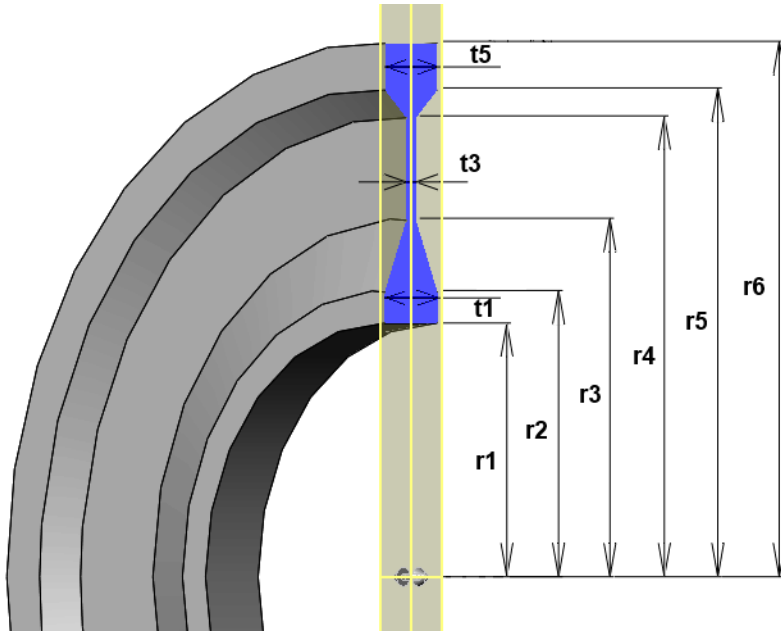
Table : Comparison of RRMS errors

n_l	1000	3000	5000
VFGP	0.3636	0.1351	0.1028
SVFGP	0.3636	0.3281	0.3586
BB VFGP	0.000998	0.00113	0.00034

Table : Comparison of extrapolation RRMS errors

Rotating disk problem

- We predict maximum radial displacement $u_{\max}$ and maximum load $s_{\max}$ for a rotating disk in an airplane engine.
- Geometry description includes 9 parameters: radii r_i, $i = 1, \ldots, 6$ define where thickness of disk changes, and thicknesses t_1, t_3, t_5 define thicknesses.
- Radii r_4, r_5 and thickness t_3 are fixed for this problem.



Results for rotating disk problem

n_h	20	40	60	80	100
GP	0.3368	0.1826	0.1305	0.1091	0.0756
VFGP	0.1679	0.0998	0.0822	0.0564	0.0435
SVFGP	0.1018	0.0658	0.0494	0.0427	0.0339
BB VFGP	0.0964	0.0717	0.0503	0.0434	0.0347

Table : RRMS errors for $u_{\max}$

n_h	20	40	60	80	100
GP	0.5261	0.3181	0.2164	0.2095	0.1643
VFGP	0.2336	0.2326	0.2058	0.1321	0.1088
SFGP	0.1674	0.1095	0.1023	0.0939	0.0812
BB VFGP	0.1583	0.1283	0.1295	0.0899	0.0793

Table : RRMS errors for $s_{\max}$

Optimization using surrogate models

$$
\begin{aligned}
m, u_{\max} &\to \min_{r_1,\ldots,r_6,t_1,t_3,t_5}, \qquad (1)\\
u_{\max} &\le 0.3, s_{max} \le 600,\\
10 \le r_1 &\le 110, 120 \le r_2 \le 140,\\
150 \le r_3 &\le 168, 170 \le r_4 \le 200,\\
4 \le t_1 &\le 50, 4 \le t_3 \le 50,\\
r_5 &= 210, r_6 = 230, t_5 = 32.
\end{aligned}
$$

Optimization algorithm

- Generate initial sample of size 30 points using LHS.
- Construct surrogate models using GP, VFGP, SVFGP and BB VFGP approaches.
- Solve multiobjective optimization problem at hand using these surrogate models as the target functions and constraints.
- To estimate quality of models we calculate true values at Pareto frontiers obtained during optimization using high fidelity solver.

Obtained Pareto frontiers

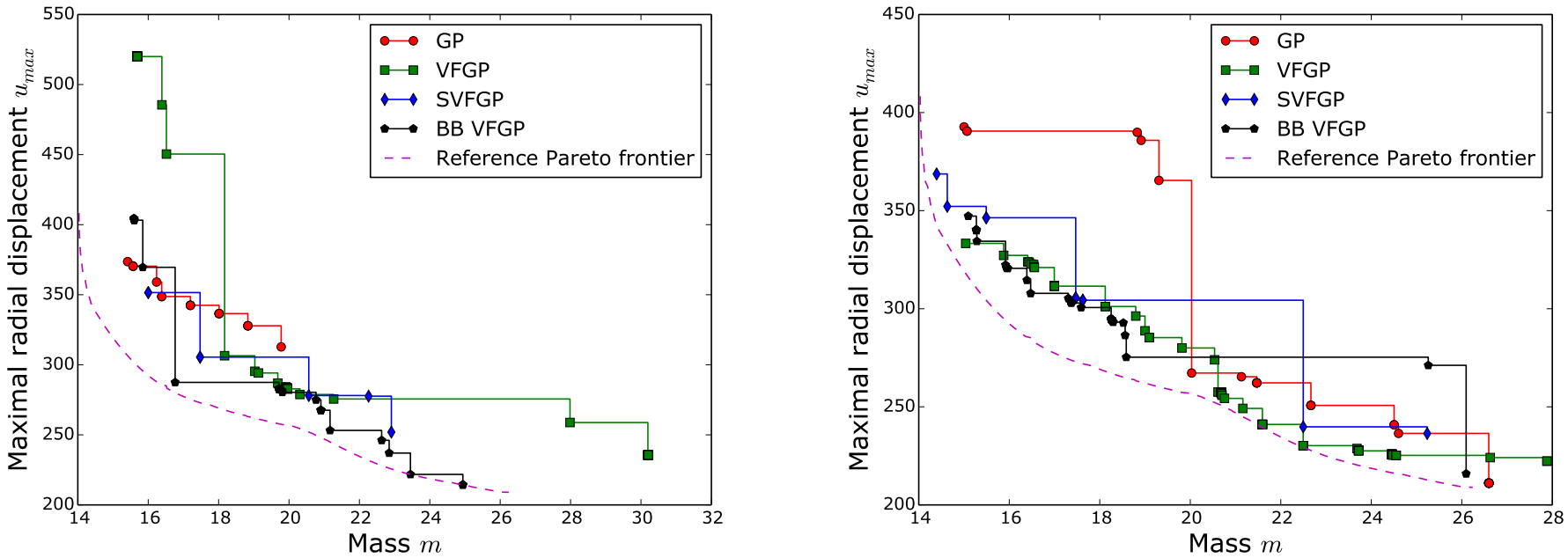


Figure : Pareto frontiers obtained using optimization of surrogate models

Optimization using surrogate models

Objective	GP	VFGP	SVFGP	BB VFGP
m	16.62	15.69	**15.09**	15.63
$0.8m + 0.2u_{\max}$	73.65	70.74	70.71	**68.10**
$0.6m + 0.4u_{\max}$	125.10	117.37	116.21	**112.55**
$0.4m + 0.6u_{\max}$	176.55	163.89	161.18	**156.99**
$0.2m + 0.8u_{\max}$	228.00	210.33	206.12	**201.44**
$u_{\max}$	279.44	256.77	251.05	**245.89**
Feasible points share	0.54	0.57	0.55	**0.75**

Table : Optimization results for three different surrogate models used with minimal values for different optimization objectives. Also we present share of feasible points in the final Pareto frontier. The best values are in bold font.

Obtained results: *Problem 1*

- We introduced Bernstein-von Mises theorem for estimates of covariance function parameters for Gaussian processes regression
- Our Bernstein-von Mises theorem holds for finite sample size and possibly wrong parametric assumption.
  - Posterior mean is close to MLE estimate, and thus is close to the central point.
  - Posterior covariance matrix is close to D_0^{-2}.
  - Posterior distribution of parameters is close to Gaussian distribution with respect to Total variance distance.
- There exists a widely used covariance function with parameters space dimension $p > 1$, for which introduced assumptions hold.

Obtained results: *Problems 2 and 3*

- We proposed an approach for proceeding of variable fidelity data with large sample sizes.
- We proposed an approach for update of model if blackbox for low fidelity function is available.
- MACROS library by DATADVANCE includes introduced approaches.

Publications

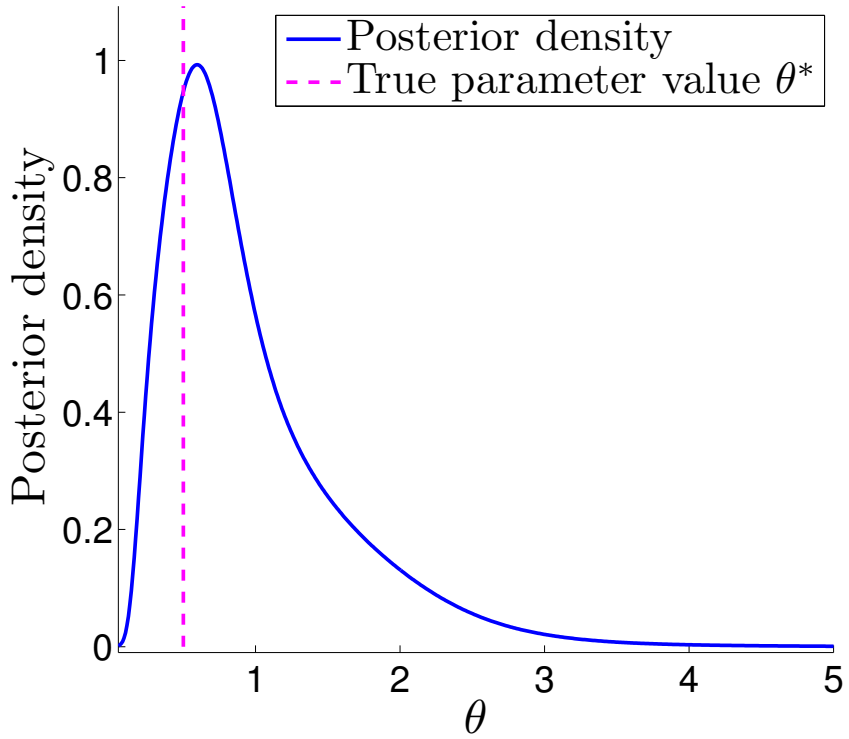
- A.A. Zaitsev, E.V. Burnaev and V.G. Spokoiny, *Properties of the posterior distribution of a regression model based on Gaussian random fields*, Automation and Remote control, 2013
- A.A. Zaitsev, E.V. Burnaev and V.G. Spokoiny, *The Bernstein-von Mises theorem for regression based on Gaussian Processes*, Russian Mathematical Surveys, 2013
- A.A. Zaitsev, E.V. Burnaev and V.G. Spokoiny, *Properties of the Bayesian Parameter Estimation of a Regression Based on Gaussian Processes*, Journal of Mathematical Sciences, 2013
- A.A. Zaitsev and E.V. Burnaev, *Variable fidelity surrogate modeling using low fidelity function blackbox and sparsification*, preprint, 2015

State of the art

- Forrester, A. et al., *Engineering Design via Surrogate Modelling: a Practical Guide*, Chichester, Wiley, 2008.
- Rasmussen C.E., *Gaussian processes for machine learning*, Cambridge, The MIT press, 2006.
- Shaby, B., Ruppert, D., Tapered Covariance: Bayesian Estimation and Asymptotics, J. Comp. Graph. Stat., 2012, 21, 2, 433–452.
- Mardia, K.V., Marshall, R.J., Maximum likelihood estimation of models for residual covariance in spatial regression, Biometrika, 1984, 71, 1, 135–146.
- Spokoiny, V., Parametric estimation. Finite sample theory, Ann. Statist., 2012, 50, 6, 2877-2909.

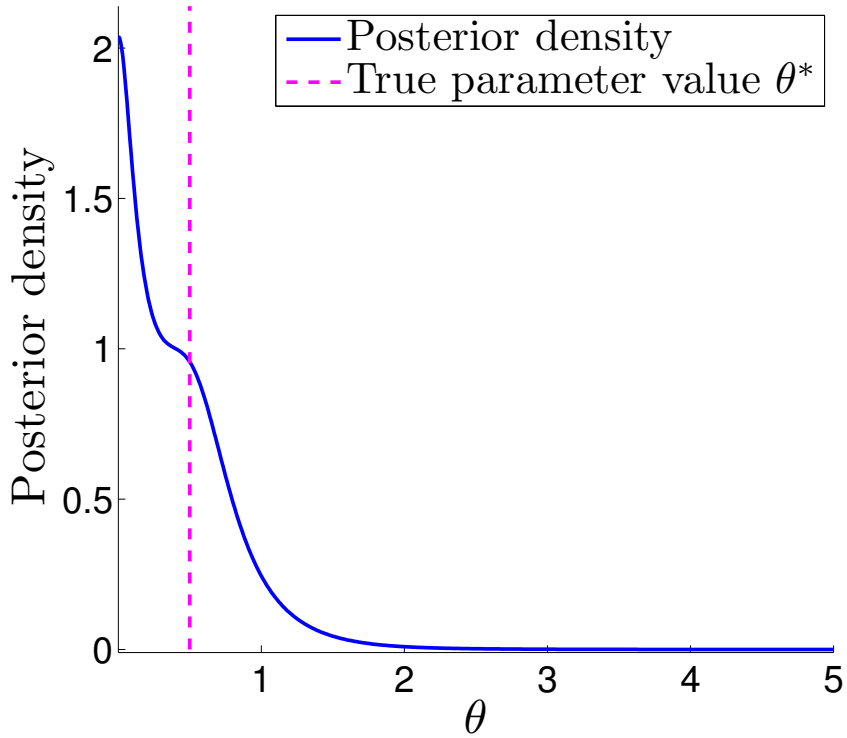
Common behaviour of posterior density of $\boldsymbol{\theta}$

- We use squared exponential covariance function.
- Noise variance σ^2 is fixed and equals 0.01.
- Parameter $\boldsymbol{\theta}^*$ equals 0.5.
- Sample size n equals 50.
- We use noninformative prior distribution.



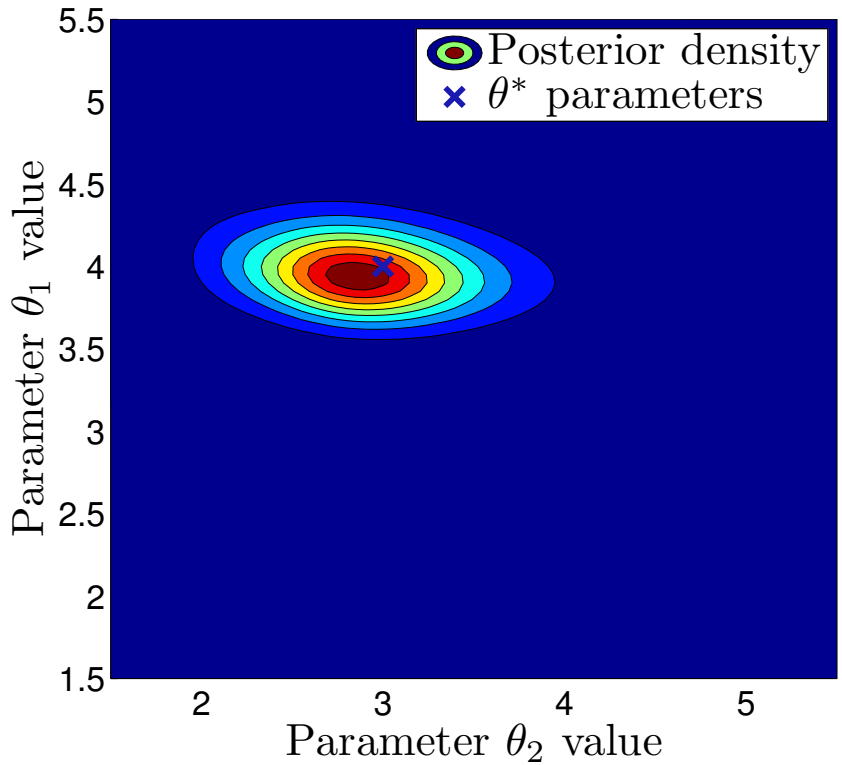
Degenerate behaviour of posterior density of $\boldsymbol{\theta}$

- We use squared exponential covariance function.
- Noise variance σ^2 is fixed and equals 0.01.
- Parameter $\boldsymbol{\theta}^*$ equals 0.5.
- Sample size n equals 50.
- We use noninformative prior distribution.
- Due to design X covariance matrix is close to singular.



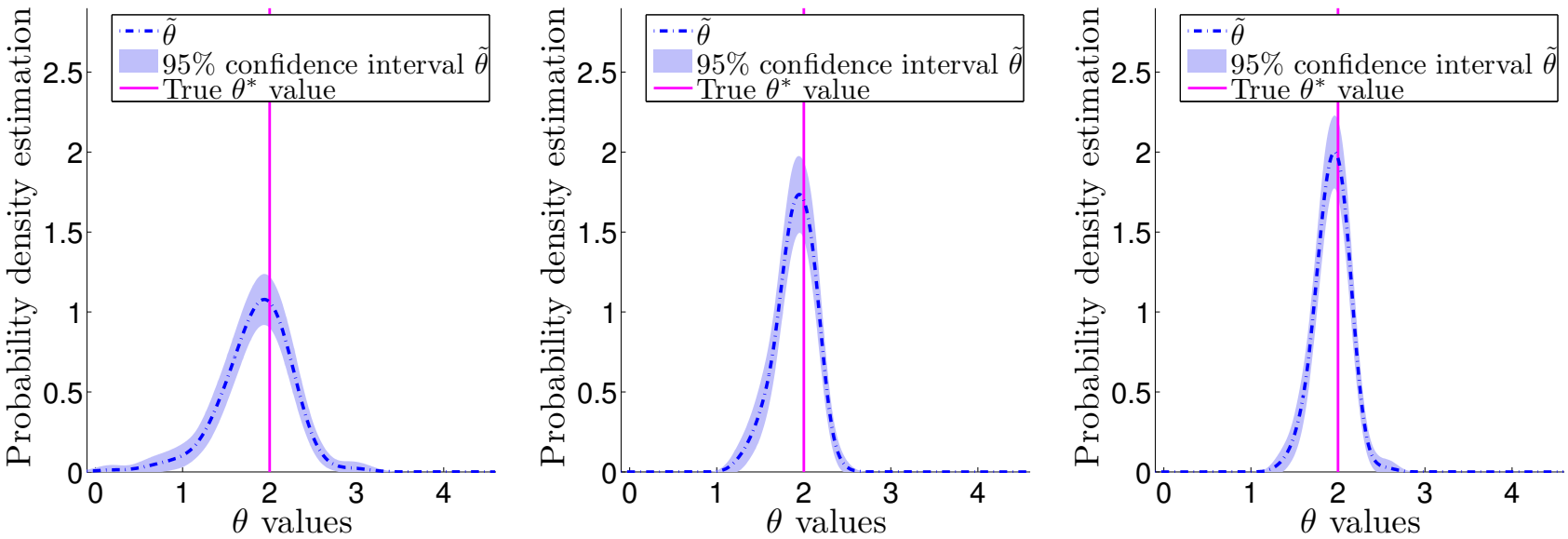
Common behaviour of posterior density of $\boldsymbol{\theta}$

- We use squared exponential covariance function.
- Noise variance σ^2 is fixed and equals 0.01.
- Parameter $\boldsymbol{\theta}^*$ equals $[2, 3]$.
- Sample size n equals 500.
- We use noninformative prior distribution.



Obtained estimates of $\widetilde{\boldsymbol{\theta}}$, parameters space dimension is $p = 1$.

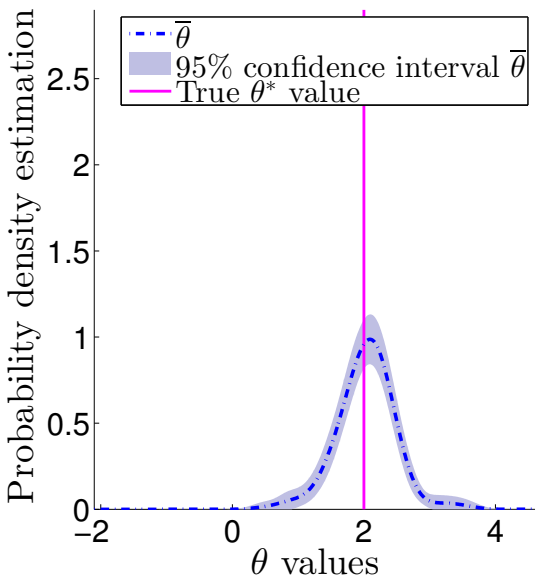
- We get KDE for $\widetilde{\boldsymbol{\theta}}$ using samples $\boldsymbol{D}_i$, $i = \overline{1, 400}$.
- While sample size n increases, estimates of $\widetilde{\boldsymbol{\theta}}$ concentrates in a vicinity of true value $\theta^* = 2$.



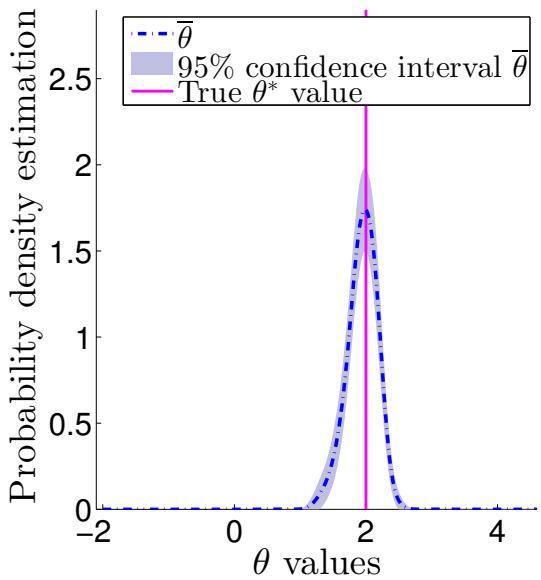
(a) $n = 10$ (b) $n = 200$ (c) $n = 1000$

Obtained estimates of $\overline{\boldsymbol{\theta}}$, parameters space dimension is $p = 1$.

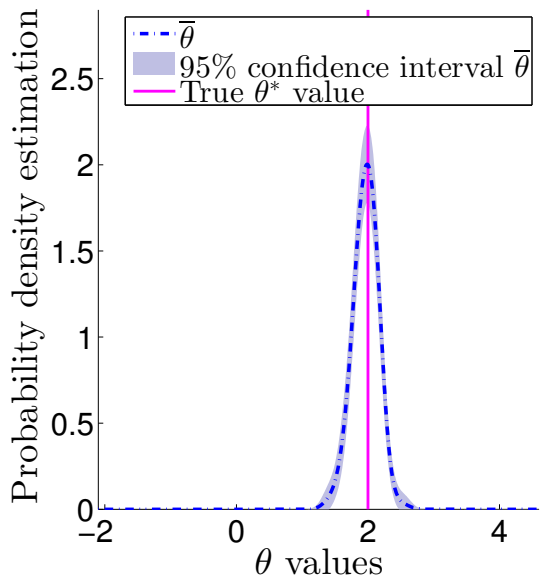
- We get KDE for $\overline{\boldsymbol{\theta}}$ using samples $\boldsymbol{D}_i$, $i = \overline{1, 400}$.
- While sample size$n$ increases, estimates of $\overline{\boldsymbol{\theta}}$ oncentrates in a vicinity of true value $\theta^* = 2$.



(a) $n = 10$



(b) $n = 200$



(c) $n = 1000$

Theorem statement, $p = 6$.

$$\|D_0(\overline{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})\|^2 \leq 4\Delta_\circ(\mathtt{x}) + \diamondsuit^2(\mathtt{r}_0, \mathtt{x}) + 4e^{-\mathtt{x}}.$$

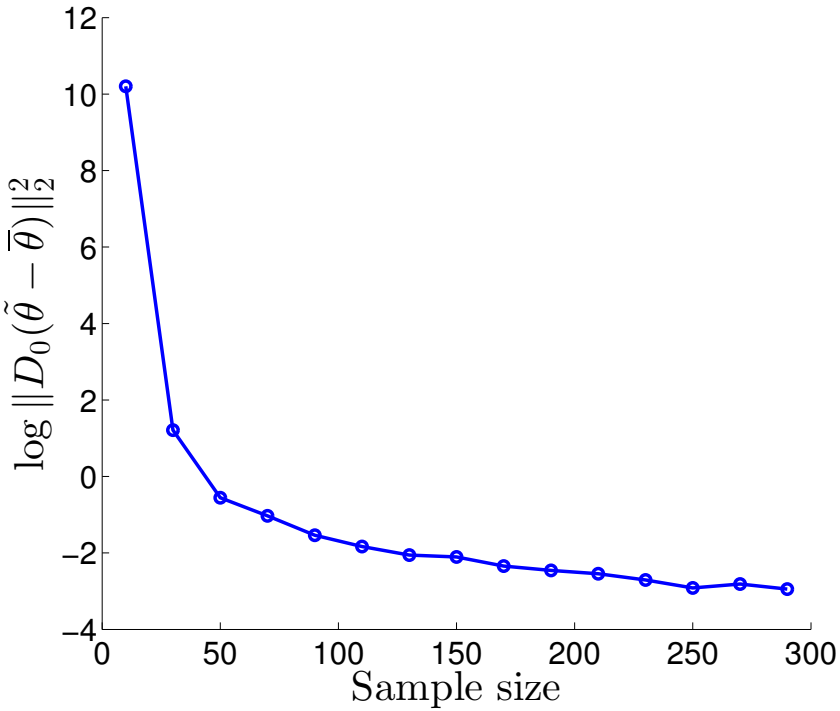


Figure : Size of parameters space $p = 6$

Critical sample size for Gaussian processes regression

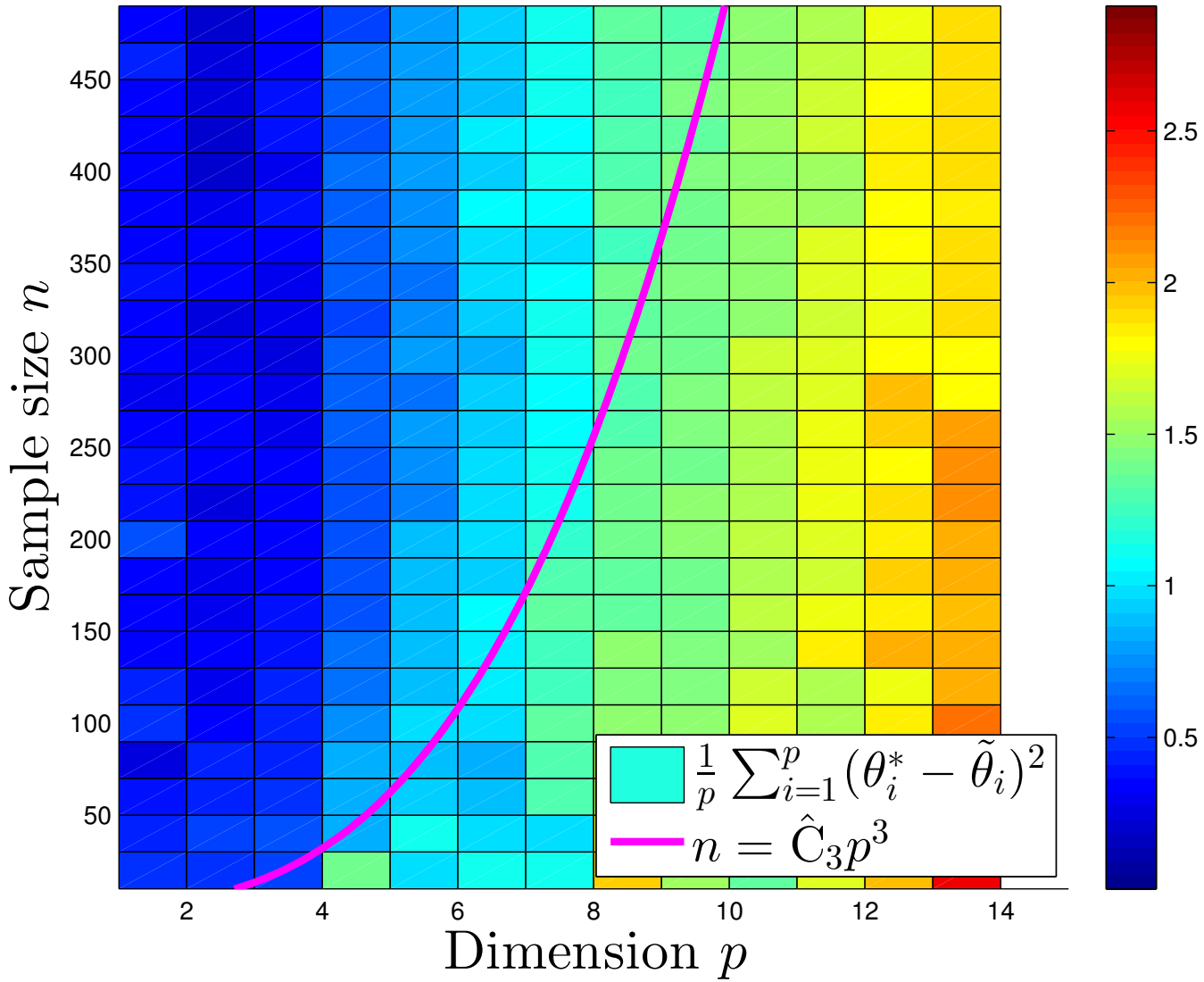
- The following inequality holds for the sample size n and the parameters space dimension p is required for BvM theorem to hold:

$$n \geq 4C\mathbf{r}_0^2 p^3$$

- We use the following squared exponential covariance function

$$k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\sum_{i=1}^{d} \theta_i^2 (x_i - x_i')^2\right) + \sigma^2 \delta(\mathbf{x} - \mathbf{x}').$$

Critical sample size for Gaussian processes regression



Constants in BvM theorem

$$\mathtt{g} \stackrel{\text{def}}{=} \frac{\sqrt{n\mathtt{v}_0}}{2\sqrt{p}\frac{1}{\lambda_0^2}\overline{\lambda}\lambda_1}, B \stackrel{\text{def}}{=} D_0^{-1}V_0^2D_0^{-1},$$

$$\mathtt{v} \geq \frac{1}{n}\|V_0^2\|_2 \text{ и } \mathtt{d} \geq \frac{1}{n}\|D_0^2\|_2,$$

$$\mathtt{p}_B \stackrel{\text{def}}{=} \operatorname{tr}(B) \leq p\left(\mathtt{v}/\mathtt{d}_0\right)^2, \mathtt{v}_B^2 \stackrel{\text{def}}{=} 2\operatorname{tr}(B^2) \leq 2\left(\mathtt{v}/\mathtt{d}_0\right)^4,$$

$$\lambda_B \stackrel{\text{def}}{=} \lambda_{\max}(B) \leq p\left(\mathtt{v}/\mathtt{d}_0\right)^2, \mathtt{g}_c \stackrel{\text{def}}{=} \sqrt{\mathtt{g}^2 - 2\mathtt{p}_B/3}, \mathtt{y}_c^2 \leq \mathtt{p}_B + 6\lambda_B\mathtt{x}_c,$$

$$2\mathtt{x}_c \stackrel{\text{def}}{=} \left(3/2\mathtt{g}^2 - \mathtt{p}_B\right)/\lambda_B + \log\|I_p - 2B/(3\lambda_B)\|_2,$$

$$z(B,\mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \mathtt{p}_B + 2\mathtt{v}_B\mathtt{x}^{1/2}, & \mathtt{x} \leq \mathtt{v}_B/(18\lambda_B), \\ \mathtt{p}_B + 6\lambda_B\mathtt{x}, & \mathtt{v}_B/(18\lambda_B) < \mathtt{x} \leq \mathtt{x}_c, \\ |\mathtt{y}_c + 2\lambda_B(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c|, & \mathtt{x} > \mathtt{x}_c, \end{cases}$$

Constants in BvM theorem

$$z_{\mathbb{H}}(\mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \sqrt{4p + 2\mathtt{x}}, & 4p + 2\mathtt{x} \leq \mathtt{g}^2, \\ \mathtt{g}^{-1}\mathtt{x} + \frac{1}{2}(\mathtt{g}^{-1}4p + \mathtt{g}), & 4p + 2\mathtt{x} > \mathtt{g}^2, \end{cases}$$

$$C_3 \stackrel{\text{def}}{=} 4\frac{1}{\lambda_0^6}\lambda_1^3\overline{\lambda}^3 + 5.5\frac{1}{\lambda_0^4}\overline{\lambda}^2\ \lambda_1\lambda_2 + \frac{1}{\lambda_0^2}\overline{\lambda}\lambda_3,$$

$$\delta(\mathtt{r}) \stackrel{\text{def}}{=} \frac{C_3 p^{\frac{3}{2}}}{\sqrt{n}\mathtt{d}_0^{\frac{3}{2}}}, \nu_0^2 \stackrel{\text{def}}{=} \max\left(1, 2\left(\frac{\sqrt{p}}{\sqrt{\mathtt{v}_0}}\frac{1}{\lambda_0^2}\overline{\lambda}\lambda_1\right)^2\right), \mathtt{b} = \frac{\mathtt{d}_0}{2\mathtt{d}},$$

$$\diamondsuit(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} \left(\delta(\mathtt{r}) + 6\frac{1}{n}\nu_0 z_{\mathbb{H}}(\mathtt{x})\right)\mathtt{r}.$$

Constants in BvM theorem

Statement

For BvM theorem to hold required sample size has to be

$$n \geq 4\mathtt{d}_0^3 C_3^2 \mathtt{r}_0^2 p^3.$$