Data Fusion based on Gaussian Processes

Alexey Zaytsev, Evgeny Burnaev

June 22, 2012

Table of contents

Engineering problem statement: Sample-based Data Fusion

- $y_h(\mathbf{x})$ — high fidelity function, $y_l(\mathbf{x})$ — low fidelity function, $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n,\ y_l \in \mathbb{R},\ y_h \in \mathbb{R}$.
- $D_l = (X_l, \mathbf{y}_l) = \{(\mathbf{x}_i^l, y_l(\mathbf{x}_i^l))\}_{i=1}^{N_l}$ — low fidelity function sample,
- $D_h = (X_h, \mathbf{y}_h) = \{(\mathbf{x}_j^h, y_h(\mathbf{x}_j^h))\}_{j=1}^{N_h}$ — high fidelity function sample, $N_h \ll N_l$.
- It is assumed that $y_l(\mathbf{x})$ and $y_h(\mathbf{x})$ model the same physical phenomenon.
- Our goal is to build an approximation

$$\hat{y}_h(\mathbf{x}) \approx y_h(\mathbf{x})$$

using both high fidelity and low fidelity samples.

State of the art

- The sample-based data fusion problem is elaborated in Forrester's paper in 2007. However, we have proposed a better algorithm for parameter learning, than the state of the art algorithms (this algorithm isn't an issue of the current talk).
- For Sparse gaussian processes regression Foster reviewed existed techniques in 2010 article.
- We consider the Data Fusion problem for large sample sizes. Suppose

$$N_h \gg 1, \text{or}$$
$$N_l \gg 1.$$

We can't evaluate VFGP regression because problem becomes computationally intractable.

Table of contents

Gaussian processes regression

- We assume that $y_h(\mathbf{x})$ is a realization of GP with apriory selected mean and covariance functions.
- Gaussian process is defined by it's mean and covariance function.
- Suppose mean to be zero.
- Covariance function $k(\mathbf{x}, \mathbf{x}')$ of any GP is supposed to be of the form

$$k(\mathbf{x}, \mathbf{x}') = \theta_0 \exp\left(-\sum_{i=1}^{n} \theta_i^2 (x_i - x_i')^2\right) + \sigma^2 \delta(\mathbf{x}, \mathbf{x}'),$$

x_i — i-th element of a vector $\mathbf{x} \in \mathbb{X}$, $\delta(\mathbf{x}, \mathbf{x}')$ is the delta function.

Gaussian processes regression

- $D_N = \{(\mathbf{x}_i, y_i = y(\mathbf{x}_i))\}_{i=1}^N$ — sample generated by GP.
- $\mathrm{Law}(y(\mathbf{x})|D_N) = \mathcal{N}(\mu_N(\mathbf{x}), V_N(\mathbf{x}))$, where
- posterior mean, used for forecast $\hat{y}(\mathbf{x})$, has the form:

$$\hat{y}(\mathbf{x}) = \mu_N(\mathbf{x}) = \mathbf{k}(\mathbf{x}, X) K^{-1} \mathbf{y},$$

  where
  $K = K(X, X) = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^N$, $\mathbf{k}(\mathbf{x}, X) = \{k(\mathbf{x}, \mathbf{x}_i)\}_{i=1}^N$.
- posterior covariance, used for accuracy estimation, has the form:

$$V_N(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x}, X) K^{-1} \left(\mathbf{k}(\mathbf{x}, X)\right)^T.$$

- Parameters $\{\boldsymbol{\theta}, \sigma\}$, $\boldsymbol{\theta} = \{\theta_0, \theta_1, \ldots, \theta_n\}$, of the covariance function $k(\mathbf{x}, \mathbf{x}')$ are estimated by likelihood maximization:

$$-\frac{1}{2} \ln |K| - \frac{\mathbf{y}^T K^{-1} \mathbf{y}}{2} \rightarrow \max_{\boldsymbol{\theta}, \sigma}.$$

Table of contents

The model of variable fidelity data

- Variable fidelity GP (VFGP) model has the form:

$$
\begin{aligned}
y_l(\mathbf{x}) &= f_l(\mathbf{x}), \\
y_h(\mathbf{x}) &= \rho f_l(\mathbf{x}) + f_d(\mathbf{x}),
\end{aligned}
$$

- $f_l(\mathbf{x})$ and $f_d(\mathbf{x})$ are gaussian processes,
- hyperparameters of gaussian process $f_l(\mathbf{x})$ covariance function $k_l(\mathbf{x}, \mathbf{x}')$ are $\boldsymbol{\theta}_l$ and σ_l,
- hyperparameters of gaussian process $f_d(\mathbf{x})$ covariance function $k_d(\mathbf{x}, \mathbf{x}')$ are $\boldsymbol{\theta}_d$ and σ_d,
- ρ is a link coefficient.

Table of contents

Variable fidelity gaussian processes learning

- Let us denote by

$$X = \begin{pmatrix} X_l \\ X_h \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} \mathbf{y}_l \\ \mathbf{y}_h \end{pmatrix}.$$

- For a gaussian process likelihood has the form:

$$-\frac{1}{2} \ln |\det K| - \frac{\mathbf{y}^T K^{-1} \mathbf{y}}{2},$$

- To maximize joint likelihood for high and low fidelity samples we use the following procedure:
    - Find the maximum likelihood hyperparameters $\boldsymbol{\theta}_l$ and σ_l for sample $D_l = (X_l, \mathbf{y}_l)$ and obtain approximation $\hat{y}_l(\mathbf{x})$.
    - Find the maximum likelihood hyperparameters $\boldsymbol{\theta}_d$, σ_d and ρ for sample $(X_h, \mathbf{y}_h - \rho \hat{\mathbf{y}}_l(X_h))$.

Variable fidelity gaussian processes evaluation

- For the given samples D_l and D_h ($N = N_h + N_l$)

$$\mathrm{Law}(y_h(\mathbf{x})|D_l, D_h) = \mathcal{N}(\mu_N(\mathbf{x}), V_N(\mathbf{x})).$$

- posterior mean, used for forecast $\hat{y}_h(\mathbf{x})$, has the form:

$$\hat{y}_h(\mathbf{x}) = \mu_N(\mathbf{x}) = \mathbf{k}(\mathbf{x}, X)K^{-1}\mathbf{y},$$

where

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

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

- posterior covariance, used for accuracy estimation, has the form:

$$V_N(\mathbf{x}) = \rho k_l(\mathbf{x}, \mathbf{x}) + k_d(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x}, X)K^{-1}\left(\mathbf{k}(\mathbf{x}, X)\right)^T.$$

Complexity of GP model

Denote sample size by N ($N = N_h + N_l$ in case of Variable Fidelity GP model). Time complexity of operations during hyperparameters optimization:

- $O(N^3)$ — likelihood evaluation (we need to inverse covariance matrix),
- $O(N^2)$ — calculate likelihood derivatives.

Time complexity of operations during model evaluation:

- $O(N)$ — estimate mean,
- $O(N^2)$ — estimate variance.

So, we need $O(N^3)$ to estimate hyperparameters and $O(N^2)$ to obtain posterior for a new point. Using *Sparse Variable Fidelity GP* we can reduce time and memory complexity.

Table of contents

Sparse variable fidelity gaussian processes

- For the given samples D_l and D_h

$$\mathrm{Law}(y_h(\mathbf{x})|D_l, D_h) = \mathcal{N}(\mu_N(\mathbf{x}), V_N(\mathbf{x}))$$

- posterior mean, used for forecast $\hat{y}_h(\mathbf{x})$, has the form:

$$\hat{y}_h(\mathbf{x}) = \mu_N(\mathbf{x}) = \mathbf{k}(\mathbf{x}, X) K^{-1} \mathbf{y}.$$

- posterior covariance, used for accuracy estimation, has the form:

$$V_N(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x}, X) K^{-1} \left(\mathbf{k}(\mathbf{x}, X)\right)^T,$$

$k(\mathbf{x}, \mathbf{x}) = \rho k_l(\mathbf{x}, \mathbf{x}) + k_d(\mathbf{x}, \mathbf{x})$.

- We select samples of *support points* from the training samples: $X_l^1 \subset X_l$, $X_h^1 \subset X_h$ with sizes N_1^l, N_1^h correspondingly.
- Denote

$$X^1 = \begin{pmatrix} X_l^1 \\ X_h^1 \end{pmatrix}, \mathbf{y}^1 = \begin{pmatrix} \mathbf{y}_l(X_l^1) \\ \mathbf{y}_h(X_h^1) \end{pmatrix}.$$

Sparse variable fidelity gaussian processes

Using Nöstrom approximation we obtain that

$$\begin{aligned}
\mathbf{k}(\mathbf{x}, X) \approx & \hat{\mathbf{k}}(\mathbf{x}, X) = K_1^* K_{11}^{-1} K_1, \\
K(X, X) \approx & \hat{K}(X, X) = (K_1)^T K_{11}^{-1} K_1, \\
k(\mathbf{x}, \mathbf{x}) \approx & \hat{k}(\mathbf{x}, \mathbf{x}) = K_1^* K_{11}^{-1} (K_1^*)^T,
\end{aligned}$$

where

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

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

$$K_1^* = \begin{pmatrix} \rho \mathbf{k}_l(\mathbf{x}, X_l^1) \\ \rho^2 \mathbf{k}_l(\mathbf{x}, X_h^1) + \mathbf{k}_d(\mathbf{x}, X_h^1) \end{pmatrix}.$$

Sparse model evaluation

- Let us denote
  1. R has the form:

$$R = \begin{pmatrix} \frac{1}{\sqrt{\theta_0^l \sigma_l^2}} I_{N_l} & 0 \\ 0 & \frac{1}{\sqrt{\rho^2 \theta_0^l \sigma_l^2 + \theta_0^d \sigma_d^2}} I_{N_h} \end{pmatrix},$$

  2. V_{11} is a Cholesky decomposition of the matrix K_{11},
  3. $V_1 = \tilde{K}_1 V_{11}^{-T}$, $\tilde{K}_1 = K_1 R$.

Proposition For the introduced assumptions an approximation for aposteriory mean and variance holds:

- Approximate posterior mean:

$$\hat{y}_h(\mathbf{x}) = \mu_N(\mathbf{x}) \approx K_1^* V_{11} (I + V^T V)^{-1} V^T \mathbf{y}.$$

- Approximate posterior variance:

$$V_N(\mathbf{x}) \approx (\rho^2 \sigma_l^2 + \sigma_d^2) + K_1^* V_{11} (I + V^T V)^{-1} V_{11}^{-1} (K_1^*)^T.$$

Sparse Variable Fidelity GP complexity

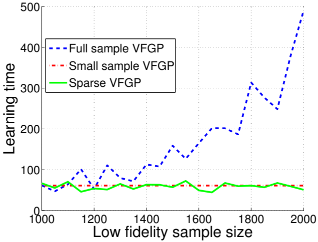
- Initial sample size $N = N_h + N_l$.
- Size of support sample $N_1 = N_h^1 + N_l^1$.
- Time complexity of operations during hyperparameters optimization:
  - $O(N_1^2 N)$ — likelihood evaluation (we need to inverse covariance matrix),
  - $O(N_1^2)$ — calculate likelihood derivatives.
- Time complexity of operations during model evaluation:
  - $O(N)$ — mean estimation,
  - $O(N_1 N)$ — variance estimation.

Table of contents

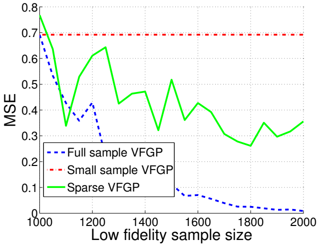
Model data

$$y_h(\mathbf{x}) = 20 + \sum_{i=1}^{2} \left(x_i^2 - 10\cos(2\pi x_i)\right),$$

$$y_l(\mathbf{x}) = y_h(\mathbf{x}) + 0.2 \sum_{i=1}^{2} \left(\frac{\tilde{x}_i}{c} + 1\right)^2, \tilde{x}_i = \frac{x_i}{10.28} + 0.5.$$



(a) Learning time



(b) Approximation quality

Real data problem

- We approximate airfoil drug C_d and lift C_l coefficients.
- As high and low fidelity function we use tight and coarse meshes correspondingly for Euler equations solving.
- A usual airfoil parametrization specifies ordinates for a fixed set of absciss points. Typically, dimensionality of such a parametrization is ≈ 59.
- To get input variables $\mathbf{x}$ of dimension 12 we compress a depicted parametrization using PCA and add angle of attack α to the list of variables.

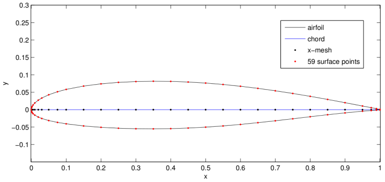


Figure: Airfoil parametrization

Real data results

- Low fidelity sample size is 200, high fidelity learning sample size is 100 for **GP** and **VFGP** and 350 for **Sparse GP** and **Sparse VFGP**.
- Errors are averaged over ten splits of the whole sample to learning and control.

	GP	Sparse GP	VFGP	Sparse VFGP
C_l	0.1671	0.0826	0.1284	0.0788
C_d	0.0114	0.0068	0.0084	0.0050

Table: Averaged approximation MSE for control sample

Results

- We generalize sparse gaussian processes regression approach to the data fusion problem.
- Provided algorithm demonstrates superior performance to previously used.