Extraction of the most sensitive directions via Gaussian Process Modelling

PreMoLab
Research laboratory in Predictive Modeling and Optimization at PhysTech

Evgeny Burnaev, Pavel Prikhodko, Pavel Erofeev
DATADVANCE, PreMoLab, IITP

Introduction

- Effective dimension reduction problem is considered
- VEGA algorithm is proposed, based on gaussian process regression. Its main features are:
  - No need for prior selection of reduced dimensionality since data can be compressed to any selected reduced dimensionality without re-training the model
  - Better accuracy compared to the state-of-the-art methods

Effective dimension reduction (EDR) problem statement

- Let the data $D = (X, Y) = \{\mathbf{x}_i, y_i(\mathbf{x}_i)\}_{i=1}^n$ be generated by the process
$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon(\mathbf{x}),$$
where $f(\mathbf{x}) = g(B\mathbf{x})$, ε is a random noise with $E\varepsilon = 0$, $B \in \mathbb{R}^{d\times m}$, $d < m$, $BB^T = I_{d\times d}$
- The problem of Effective Dimension Reduction is to estimate $S = span\{B\}$ (Central Mean Subspace, CMS)

State-of-the-art methods

- Inverse regression based: SIR, SAVE
- Local linear model based: SAMM, MAVE, OPG
- PLS

Gaussian process based regression

Let $y(\mathbf{x})$ be modeled as
$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon(\mathbf{x}),$$
where $f(\mathbf{x})$ is a some realization of a Gaussian Process (GP), $\varepsilon(\mathbf{x})$ is a gaussian white noise with a variance σ_1^2.
Let us assume that the covariance function $K_0(\mathbf{x}, \mathbf{x}')$ of the gaussian field $f(\mathbf{x})$ belongs to some parametric family
$$K_0(\mathbf{x}, \mathbf{x}') = \sigma_0^2 K_0(\mathbf{x}, \mathbf{x}'|\Theta),$$
where Θ is a some set of parameters, σ_0^2 is a scale parameter of the covariance function.
In this case the covariance function of the process $y(\mathbf{x})$ can be represented as
$$K(\mathbf{x}, \mathbf{x}') = K_0(\mathbf{x}, \mathbf{x}') + \sigma_1^2\delta(\mathbf{x}, \mathbf{x}'),$$
where $\delta(\mathbf{x}, \mathbf{x}')$ is a kroneker symbol

Covariance function parameters tuning

In order to estimate parameters $\mathbf{a} = \{\Theta, \sigma_0, \sigma_1\}$ of the covariance function usually maximum likelihood estimate is used. Log-likelihood has the form
$$\log p(Y|X, \mathbf{a}) = -\frac{1}{2}Y^T\mathrm{K}^{-1}Y - \frac{1}{2}\log|\mathrm{K}| - \frac{n}{2}\log 2\pi,$$
where $|\mathrm{K}|$ is a determinant of $\mathrm{K} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.
Parameters $\mathbf{a}$ are estimated by maximizing the log-likelihood $\log p(Y|X, \mathbf{a})$.

Covariance function

Here we use exponential family to model the covariance function
$$K_0(\mathbf{x}, \mathbf{x}') = \sigma_0^2 \exp(-d(\mathbf{x}, \mathbf{x}')),$$
where $d(\mathbf{x}, \mathbf{x}')$ is a distance between points $\mathbf{x}$ and $\mathbf{x}'$. Usually the following metrics are considered

- Weighted Euclidean distance:
  - $$d(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^n \theta_i^2(x_i - x_i')^2,$$
    where $\theta_i \in \mathbb{R}, i = 1, \ldots, m$
  - With this metric one has to tune m hyperparameters
- Mahalanobis distance:
  - $$d(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T A(\mathbf{x} - \mathbf{x}'),$$
    where $A \in \mathbb{R}^{m\times m}$ is a some positive definite matrix
  - Due to positive definiteness of A Cholesky decomposition holds true
    $$A = L^T L,$$
    where L is an upper triangular matrix.
  - So one has to tune $\frac{m(m+1)}{2}$ hyperparameters, parameterizing the Cholesky factor L

VEGA (Variable Extraction via Gradient Approximation)

- Let us define the metric as
$$d(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T B^T \Lambda B(\mathbf{x} - \mathbf{x}'),$$
where B is an orthogonal matrix $m \times m$, Λ is a diagonal positive definite matrix $m \times m$
- With this parameterization
  - B - rotates coordinate axes
  - Λ - tunes kernel widthes along the axes
- To reduce dimensionality to k, with already given matrices B and Λ, one needs to keep k columns of B corresponding to the largest elements of Λ.

VEGA parameter estimation

With considered parameterization an approximation $\hat{f}(\mathbf{x})$ of $f(\mathbf{x})$ can be estimated as
$$\hat{f}(\mathbf{x}) = \hat{g}(B\mathbf{x}),$$
where $\hat{g}$ is a GP approximation (a posterior mean) with covariance based on weighted Euclidean distance.

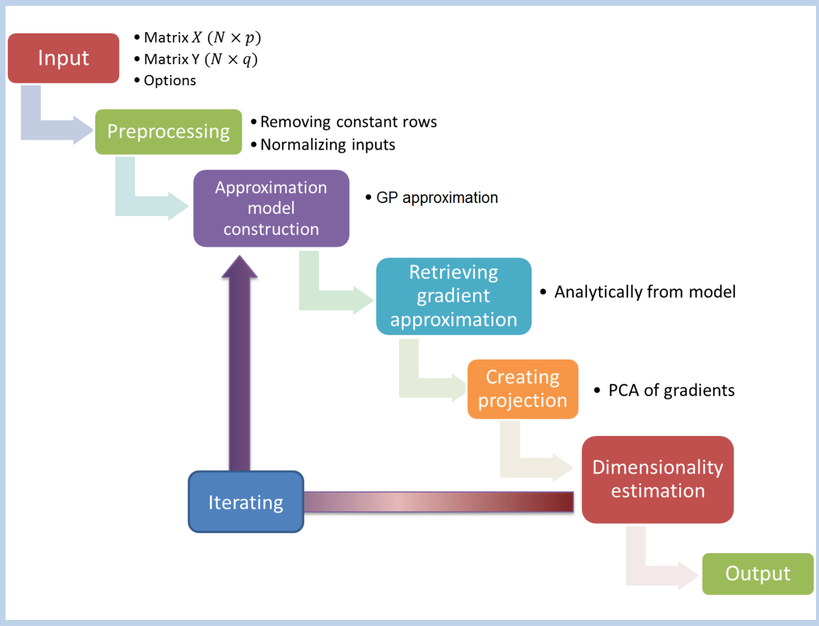
- With fixed B hyperparameters of $\hat{g}$ (diagonal matrix Λ) can be found by log-likelihood maximization
- Matrix B can be explicitly found using the following considerations. Let us consider some approximation $\hat{f}$. It allows us to get the estimates of gradients of f as
$$\hat{\Gamma} = \left\{\frac{\partial\hat{f}(\mathbf{x})}{\partial x^1}\Big|_{\mathbf{x}=\mathbf{x}_i}, \ldots \frac{\partial\hat{f}(\mathbf{x})}{\partial x^m}\Big|_{\mathbf{x}=\mathbf{x}_i}\right\}_{i=1}^n.$$
  In this case we may align axes the way that derivatives along them would not be correlated, i.e. to find B as an eigenvectors of the empirical gradient covariance matrix $\hat{\Gamma}^T\hat{\Gamma}$
- Also it might be beneficial to adjust Λ and B in an iterative manner, i.e. to find optimal Λ with fixed B and vice-versa several time.

Dimensionality estimation

Method performs a principal component analysis on the covariance matrix of the gradient estimates.

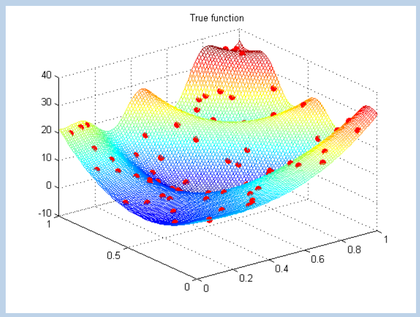
- To estimate reduced dimension one may use any of standard approaches from the principal component analysis framework.
- Also underlying probabilistic model makes it possible to develop specific statistical tests in order to estimate consistently a number of non-zero elements of Λ.

VEGA Workflow diagram

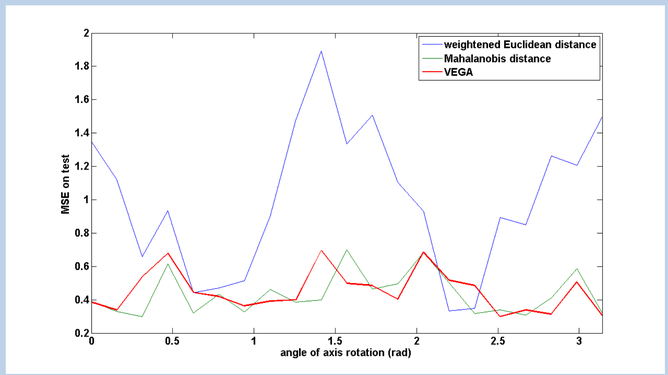


Rotation invariant

VEGA as well as a Mahalanobis distance based model provides invariance towards axes rotation. We demonstrate this on the example of Mystery function.



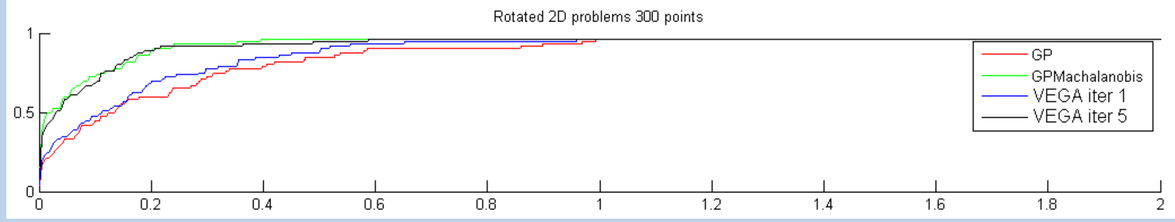
Mystery function



Accuracy of models built for rotated axes

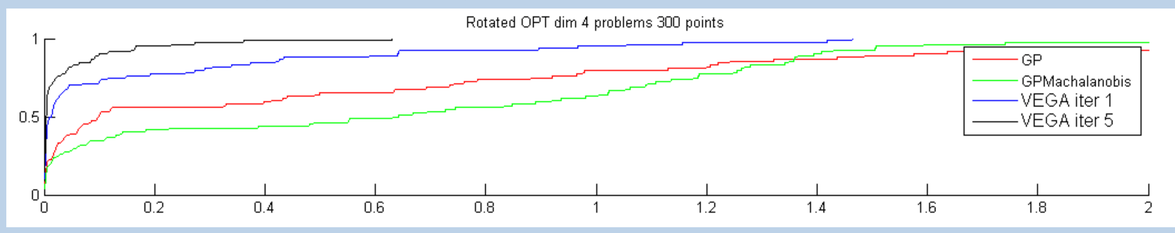
Comparison with other covariance functions

We provide comparison of techniques in terms of Mean Square Error. Dolan-More curves are drawn for a set of 30 smooth functions often used for an unconstrained optimization benchmark.
On 2D problem model based on Mahalanobis distance covariance function and VEGA works with comparable accuracy.



Dolan-moore profiles for 2D functions

However as problems dimensionality increases models based on Mahalanobis distance covariance function starts to decrease in accuracy (already in 4D case) being inferior to other methods.

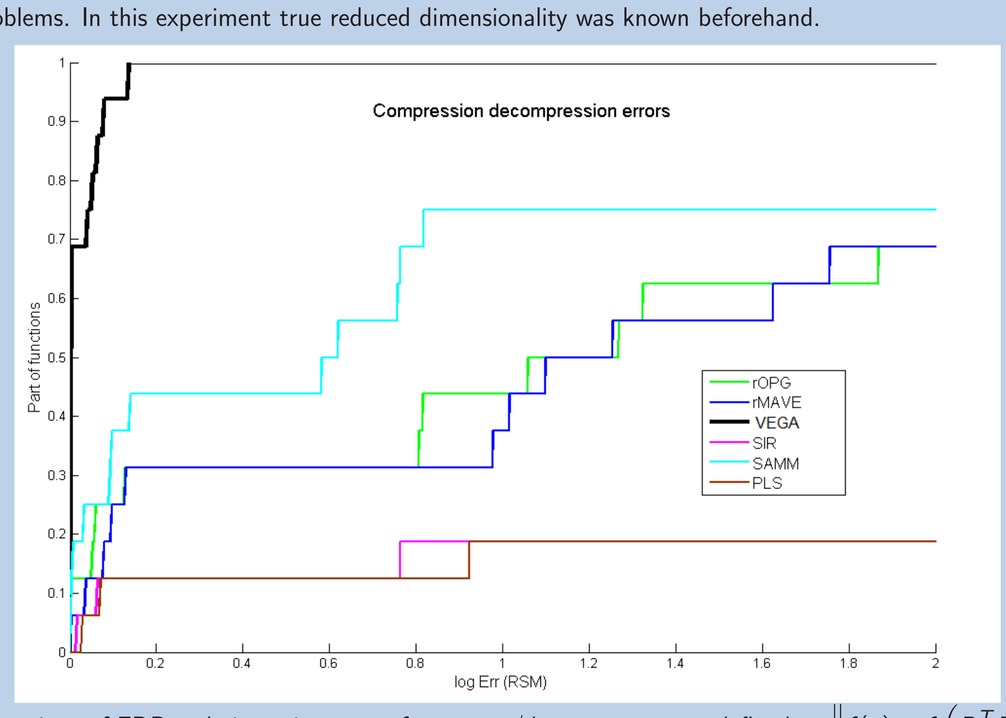


Dolan-moore profiles for 4D functions

This effect is due to a considerable growth in number $\sim m^2$ of hyperparameters for Mahalanobis distance covariance function instead of $\sim m$ for VEGA.

Comparison with different EDR techniques

Here using Dolan-More curves we provide a comparison with state-of-the-art EDR techniques for a number of test problems. In this experiment true reduced dimensionality was known beforehand.



Comparison of EDR techniques in terms of compress/decompress error, defined as $\left\|f(\mathbf{x}) - f\left(B^T B\mathbf{x}\right)\right\|$

Conclusions

- Accuracy of the proposed technique is better than the state-of-the-art effective dimension reduction approaches
- Model quality is robust to the rotation of coordinate axes
- Provides natural way to extract the most sensitive directions corresponding to the highest output variations
- When maximizing likelihood only m hyperparameters are needed to be tuned

http://datadvance.net
{evgeny.burnaev, pavel.prikhodko, pavel.erofeev}@datadvance.net