Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

GAUSSIAN PROCESSES FOR HIGH-DIMENSIONAL, LARGE DATA SETS: A REVIEW

Mengrui (Mina) Jiang
Giulia Pedrielli

School of Computing and
Augmented Intelligence
Arizona State University
699 S Mill Avenue
Tempe, AZ 85281, USA

Szu Hui Ng

Department of Industrial Systems
Engineering and Management
National University of Singapore
1 Engineering Drive 2
Singapore, 117576, SINGAPORE

ABSTRACT

Gaussian processes, known to have versatile uses in several fields across engineering, science, economics, show important advantages to several alternative approaches while controlling model complexity. However, the use of this family of models is hindered for inputs that are high dimensional as well as large sample sizes due to the intractability of the likelihood function, and the growth of the variance covariance matrix. This article investigates state-of-art solutions to these challenges according classifying them into categories. The goal is to select several algorithms covering each category and perform empirical experiments to compare their performances on the same set of test functions. Our preliminary results focus on deterministic implementations of a set of selected approaches. The results of the experiments may serve as a guidance to future readers who want to study and use Gaussian process in problems with high dimensions and big data sets.

1 INTRODUCTION & BACKGROUND

Gaussian process (GP) modeling (or kriging) is a type of nonparametric statistical model which takes any smooth function as realization of a stationary Gaussian process (Mathesen et al. 2021). Surrogates are used to approximate an, unknown, potentially nonlinear noncovex function learned from a set of inputs for which outputs (possibly noisy) are available (Bouhlel and Martins 2019). Through their hyperparameters Gaussian processes have flexibility to fit many different response types, making them among the most commonly used surrogates.

A problem that has become increasingly critical is that Gaussian processes have decaying performances when applied in problems with high dimensionality and large data sets. The difficulty of hyperparameter estimation grows exponentially with the dimensionality of data sets, while the complexity of prediction is caused by size of data sets. Recently, these challenges have drawn attention of scholars to developing contributions in scaling GPs.

Background: Gaussian process modeling Given n locations $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, each of d dimensional, and associated outputs $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))$, a Gaussian process $M(\mathbf{X}) = \mu + Z(\mathbf{X})$ can be constructed around the function, with μ the constant mean and $Z(\mathbf{X}) \sim GP\left(0, \tau^2\mathbf{K}\right)$, where τ^2 is the process variance and $\mathbf{K} = \mathbf{k}(\mathbf{X}, \mathbf{X}') =$ is the correlation matrix. One of the commonly adopted correlation function is, for each entry of $\mathbf{K}$, $K_{ij} = \prod_{l=1}^{d} e^{-(\theta_l |x_{il} - x_{jl}|)^2}$, where θ is the d-dimensinoal smoothing length-scale parameter.

Noiseless Evaluations If the true function is deterministic, which means there is no random noise added to the observations, the estimation of μ and τ^2 is $\hat{\mu} = \left(\mathbf{1}^T\mathbf{K}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}^T\mathbf{K}^{-1}\mathbf{f}$ and $\hat{\tau}^2 = \frac{1}{n}\left(\mathbf{f} - \mathbf{1}\hat{\mu}\right)^T\mathbf{K}^{-1}\left(\mathbf{f} - \mathbf{1}\hat{\mu}\right)$,

where $\mathbf{f}$ is the n-dimensional true function values at sample locations and $\mathbf{1}$ is n-vector of ones (Santner et al. 2013).

The best linear unbiased predictor and model variance result are:

$$\hat{f}(\mathbf{x}) = \hat{\mu} + \mathbf{r}(\mathbf{x})^T \mathbf{K}^{-1}(\mathbf{f} - \mathbf{1}\hat{\mu}), \tag{1}$$

$$\hat{s}^2(\mathbf{x}) = \tau^2 \left(1 - \mathbf{r}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{r}(\mathbf{x}) + \frac{\left(1 - \mathbf{1}^T \mathbf{K}^{-1} \mathbf{r}(\mathbf{x})\right)^2}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}} \right). \tag{2}$$

where $\mathbf{r}(\mathbf{x}) = r_i(\mathbf{x})$ for each $r_i(\mathbf{x}) = \mathrm{Corr}(Z(\mathbf{x}), Z(\mathbf{x}_i))$ and $i = 1, \ldots, n$ (Santner et al. 2013).

Noisy Evaluations The functions values in this case, are observed with noise, and Gaussian process model $M(\mathbf{x})$ is written as $M(\mathbf{x}) = \mu(\mathbf{x}) + \delta(\mathbf{x}) + \varepsilon(\mathbf{x})$, where $\mu(\mathbf{x})$ and $\delta(\mathbf{x})$ are mean and variance, $\varepsilon(\mathbf{x})$ is the models the noise associated to the observations. In many applications, such noise is considered as i.i.d. $N\left(0, \sigma_\varepsilon^2\right)$ (Santner et al. 2013). A rich literature is also available for the case of heterogeneous, non i.i.d. noise, where the noise variance is a function of the sample locations (Yin et al. 2011), though heterogeneous noise is not in the scope of this paper. The inference of hyperparameters θ in this case can be obtained through numerical optimization of the log likelihood function $-\frac{n}{2}\log(2\pi) - \frac{1}{2}|\mathbf{K}| - \frac{1}{2}(\mathbf{f} - \mathbf{1}\mu)^T \mathbf{K}^{-1}(\mathbf{f} - \mathbf{1}\mu)$, where $\mathbf{f}$ is the n-dimensional true function values with noises at sample locations (Liu et al. 2020). While such optimization is usually tractable for dimensionality $d \leq 20$, the task becomes quickly intractable in high dimensional problems.

Objective and paper outline In this paper, we present the recent literature that addresses these challenges, leading to a categorization of approaches, namely scaling matrix computation, embedding and projection-based methods, and additive modeling. A preliminary computational comparison of three models was also implemented to provide insights on the effectiveness of approaches in diverse categories.

The remainder of the paper is structured as follows: Section 2 provides an overview of Gaussian processes in the model structure and likelihood estimation. We highlight the challenges that arise with high dimensionality and large sample size, and propose a new categorization of recent approaches that tackle at least one of these challenges. Section 3 dives into model formulation of three approaches that we aim to compare. Section 4 presents our preliminary analysis and comparison of the different algorithms. Finally, section 5 concludes the paper providing future directions for analysis and research avenues.

2 GAUSSIAN PROCESSES AND THE CHALLENGE OF HIGH DIMENSIONALITY AND LARGE DATA SETS

As previously mentioned, the efficiency in the estimation and prediction using Gaussian processes is impacted by both the sample size as well as the dimensionality of the input. More specifically, the complexity of computing the inverse covariance matrix grows in $O(n^3)$ complexity with the number of points in the sample set n. On the other hand, the size of the hyperparameters θ grows linearly with the dimensionality of the input d. Such growth impacts the algorithm used to maximize the likelihood function (Bouhlel and Martins 2019).

In fact, scalability is a challenge, in general, for statistical learning (Donoho et al. 2000). Advances in sensor technology have made available large amounts of high dimensional data (a huge amount of features can be measured and/or extracted - see the case of deep neural networks where inputs are purposely made larger), while carrying the challenge to scale statistical models to large and high dimensional data sets. Several approaches have been proposed in the literature with some of the approaches to solve different aspects of the multiple challenges arising from large and high dimensional data sets. In order to scale models to high dimensional problems, *dimensionality reduction* is often used, some works of which include multidimensional scaling (Cox and Cox 1991), principal component analysis (Lee et al. 2012) and nonlinear dimensionality reduction (Roweis and Saul 2000). More comparative reviews can be found in (Van Der Maaten et al. 2009; Sorzano et al. 2014; Reddy et al. 2020). The most developed and commonly

used approaches include *variable selection* or *feature extraction* (Fan and Lv 2010; Wasserman and Roeder 2009; Pang et al. 2008), used to recover information from massive data. A family of approaches within the class of feature selection is based on penalized least squares (Fan et al. 2009), examples of which include Akaike and Bayesian information criterion. In the past decade, a plethora of penalizing score functions have been proposed, with a notable example being the Lasso penalized regularization (Tibshirani 1996; Zhao and Yu 2006; Greenshtein and Ritov 2004; Donoho 2006; Greenshtein 2006; Meinshausen and Bühlmann 2006; Wainwright 2006). LASSO was demonstrated to perform well in feature selection under scenarios where solutions to regression models are sparse. Also, a comparison between LASSO and other selection methods can be found in (Bickel et al. 2009). A framework for *penalized likelihoods* was proposed in (Fan and Li 2006) showing good performance across widely different statistical problems. (Fan and Lv 2008) proposed to reduce high dimensionality by means of a sure screening method called sure independence screening, which promises that, with probability going to 1, all important variables survive after applying a variable screening procedure. Another family of methods are based on the use of eigenvalues and eigenvectors of the covariance matrix (Johnstone and Titterington 2009), including canonical correlation analysis (Zhu et al. 2012), and discriminant analysis (Bouveyron et al. 2007).

2.1 Proposed Classification for Scalable Gaussian Processes

In this paper, we distinguish three main classes of approaches that tackle scalability of Gaussian processes with respect to the size of the training sample, and dimensionality of the input. We will highlight in the following sections how the different approaches can be mixed by different learning algorithms in order to solve both challenges (a representation of such interaction is in Figure 1).

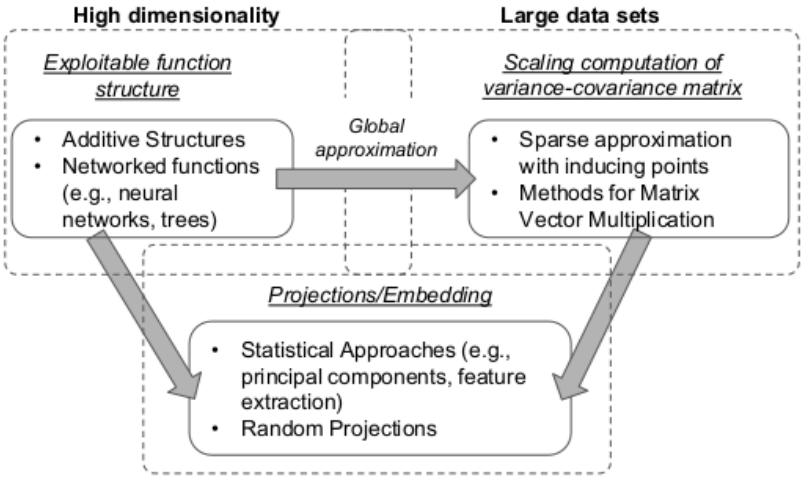


Figure 1: Graphical representation for the classification dimensions.

Specifically, we propose the following approach-classes: (i) *scaling matrix computation*, which focuses on reducing complexity coming from the computation of correlation matrix; (ii) *embedding and projection-based methods*, which reduce the dimensionality of the original input space; and (iii) additive modeling, which exploits the additive structure of the original function to produce kernels that can be efficiently estimated. The intuition behind the classification comes from an investigation of how each algorithm tackles the challenge of either high dimensionality or big data sets, or both. A graphical representation of classification details is presented in Figure 1, with all approaches restricted to the scope of literature and experiments within this review. The left and right sections are areas of challenges, high dimensionality and big data sets, respectively. The three categories mentioned earlier are in dashed boxes, with several

examples. The horizontal position of boxes indicates which challenge each category belongs to. Arrows represent directions of scaling approaches in the direction of other challenges.

2.1.1 Scaling Matrix Computation

An important task to improve estimation scalability is to efficiently process the variance covariance matrix. A common family of approaches within this category is *approximate/sparse Gaussian processes*, which build the model based on a small set of m inducing points thus reducing the time complexity for the variance covariance matrix inversion to $O\left(nm^2\right)$, where n is the, large, sample size. (Snelson and Ghahramani 2006) parametrizes the GP variance by generating pseudo-inputs that are learnt through a gradient based optimization. The authors demonstrate that even small number of such points give good performance. A variation of the inducing point method was developed by (Titsias 2009) using joint inference of inducing points and maximization of the likelihood lower bound. (Bouhlel and Martins 2019) improved method of sampling points during previous gradient-based models to achieve controllable correlation matrix size. Another family of methods, rather than approximation, focuses on improving *Matrix-Vector Multiplications* (MVMs), specifically through linear solves and log determinants based on iterative methods to reduce the complexity significantly. Expressing the covariance matrix as a Kronecker product allows the eigen-decomposition of the matrix to be efficiently discovered. If the covariance matrix is generated from a stationary covariance kernel, a Toeplitz matrix can be obtained which provides opportunity for faster matrix vector products (Wilson and Nickisch 2015). Additionally, (Gardner et al. 2018) developed fast MVMs through Lanczos decomposition of components of kernels with product structure.

2.1.2 Embedding and Projection-based Models

Embedding and projection-based methods work with the basic idea to reduce the dimensions by mapping the original input into low-dimensional spaces. An example of approach in this family is the *active subspace method* (Chen et al. 2021), which detects *active subspaces*, i.e., low dimensional subspaces, by exploiting the gradients of the function. The approach follows two steps: first, the active subspace is identified using the gradient defined with respect to the original space. subsequently, the Gaussian process is estimated over the resulting low-dimensional space. Partial Least Squares (PLS) handle high dimensional problems through maximization of the variance between input and output variables by representing them into a smaller subspace using principal components or latent variables. PLS applied with kriging achieved reduction of hyperparamter space (Bouhlel et al. 2016). Stochastic Gaussian process modeling average (SGPMA) (Xuereb et al. 2020) uses embedding on the original input space to achieve a number of subsets of lower dimensions *and* with smaller sample sizes. The number of subsets, the associated dimensionality, and the, smaller, sample sizes are user defined. This framework addresses challenges from both high dimensionality and large sample size. Similarly, random embedding Bayesian optimization (Wang et al. 2016) draws a random embedding matrix and performs optimization on the embedded space, resulting in good invariance properties; hashing-enhanced subspace Bayesian optimization (Nayebi et al. 2019), on the other hand, implements a hashing based mapping strategy that allows to map to low dimensions. Recently, (Letham et al. 2020) reexamined the linear embeddings in high dimensional Bayesian optimization and achieved performance improvement through a Mahalanobis kernel tailored for the linear embeddings and adding polytope bounds to the embedding.

2.1.3 Additive Models

An alternative way to reduce the complexity of the likelihood for the Gaussian process hyperparameter estimation is exploiting additive structure of the model. Some recent works use the additive structure of the original function to accelerate the modeling and optimization process. (Duvenaud et al. 2011) introduces a tractable kernel parametrization method exploiting the additivity of the squared exponential kernel, resulting in efficient evaluation of hyperparameters. Also based on additive kernel structure, (Durrande et al. 2011)

proposed a numerical method for data-driven parameter estimation combined with kernel additivity. (Meng and Ng 2015) combines a global Gaussian process model and a piecewise local GP model into an additive GP model with a composite covariance structure, during which inducing points are selected based on a mechanism of dividing the space into local regions (clustering). (Gardner et al. 2017) proposes a Monte Carlo Markov Chain (MCMC) based algorithm that uses Metropolis-Hastings to sample several models from the proposal distribution. Given a partition that is assumed to have an additive structure, a proposal distribution is defined over all the subpartitions that could be possibly formed. The model posterior is computed as an average of a user-defined number of samples from the proposal distribution. (Eriksson et al. 2019) uses Gaussian process as local surrogate model for robust noisy observations, followed by an implicit multi-armed bandit strategy at each optimization iteration that allocates samples across trust regions. For an additive structure of the model, (Kandasamy et al. 2015) proposes an additive Gaussian process for Bayesian optimization.

3 PRELIMINARY TESTING PLATFORM

In this section, we present a comparison platform for different modeling approaches. In this preliminary version, we consider three approaches, which we used for testing and comparing performance of a sample of methods. These were selected as representative of the categories presented in Section 2.1. In particular, the methods we selected were shown to have good performance with respect to at least one of the two challenges, high dimensionality and big data size. In this preliminary implementation, we considered function evaluations without noise, and compared the following approaches:

(Scaling matrix computation) For this class, we chose the Gradient-enhanced Kriging with Partial Least Squares (Bouhlel and Martins 2019) (GE-KPLS) method (Section 3.1);

(Embedding/Projection methods) For this class, we analyze the Stochastic Gaussian process with model averaging (Xuereb et al. 2020) (SGPMA) approach (Section 3.2);

(Additive models) For this class, we chose the approach in (Gardner et al. 2017), which we refer to as Gaussian process with MCMC sampling (in short GPMCMC) (Section 3.3).

3.1 Gradient-enhanced Kriging with Partial Least Squares (GE-KPLS)

Bouhlel and Martins (2019) proposes a model based on gradient-enhanced kriging (GEK), where the basic idea is to accelerate the Gaussian process estimation using partial least squares (kriging with partial least squares -KPLS). The existing GEK methods have intractable computation of correlation matrix when either the number of sampling points or number of inputs becomes large. To tackle this issue, the gradient-enhanced kriging with partial least squares (GE-KPLS) method suggests the following components. At each sampling point, it uses PLS regression for the hyperparameter estimation over a lower dimensional problem derived using principal component analysis. First-order Taylor approximation is used to generate a set of additional locations around each sample point. PLS is then applied to subsets of points obtaining a local influence of each component. The correlation function is

$$k\left(\mathbf{x},\mathbf{x}'\right)=\tau^2\prod_{l=1}^{h}\prod_{i=1}^{d}\exp\left[-\theta_l\left(w_{av_i}^{(l)}x_i-w_{av_i}^{(l)}x_i'\right)^2\right],\quad\forall\theta_l\in[0,+\infty)$$

where h is the number of principal components from PLS, d is the dimensionality of the input, and $w_{av_i}^{(l)}$ represents the weight associated to the *l-th* principal component in the *i-th* dimension. $n_{ext}\in[1,d]$ is the number of approximating points generated around each sampling point, and it corresponds to the highest coefficients obtained from PLS. Although not shown in the kernel form, it is a user-defined parameter during GE-KPLS model training. The same predictors as in equations (1)-(2) is then used.

3.2 Stochastic Gaussian Process Modeling Average (SGPMA)

The SGPMA approach was first proposed in (Xuereb et al. 2019), and extended in (Xuereb et al. 2020) with the objective to address both the dimensionality and data size challenges. The algorithm tackles high input dimensionality by decomposing the original dimensions into multiple, lower-dimensional, embeddings. The predictions are then obtained for each embedding using a subset of samples. Finally, the predictions are averaged using different weights for each model. Two projection operators were used by the authors: (i) the π^{random} that generates lower dimensional models randomly selecting a subset of dimensions for each predictor; (ii) the π^{PCA}, instead, uses principal component analysis over a subset of the training samples. The number of desired predictors is decided by the user, while the locations (samples) to be assigned to each model are selected uniform at random no matter the projection method used. Then, a Bayesian model averaging approach is used to coalesce the predictions obtained from each of the models defined with a potentially different projection method and sample data. The user defined parameters are the number of subsets n_{sub}, and, for each model $i = 1, \ldots, n_{sub}$, the number of data points n_i and dimensionality d_i of each subset, which will be defined by users. The estimation of the low dimension models includes the estimation of the weights associated to the sub-model. The authors defined a proxy pair, $(n_i, d_i), i = 1, \ldots, n_{sub}$, to characterize the quantity of the prediction power of each lower dimensional model. The predictor is:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \mu_i \left(f_i \left(\pi_i^t (\mathbf{x}) \right) | \mathbf{f_i} \right) w_i$$

where $\mathbf{x}$ is one d-dimensional sample location, $\mathbf{f}_i$ is the n_i-dimensional vector of true function evaluations, and $w_i = \frac{p_i}{\sum_{j=1}^{n} p_j}$, $p_i = \left(\frac{n_i}{n}\right)^2 \times \left(\frac{d_i}{d}\right)^{\eta}$, $i \in 1, ..., n_{sub}$, and n, d are the original input data size and dimensionality. $\eta > 0$ is another user-defined parameter that characterizes the model prediction power. Finally, the superscript t of the projection operator can be either set to "PCA" or "random" based on the user choice.

3.3 Gaussian Processes with Monte Carlo Markov Chain (GPMCMC)

This models attempts to face the issue of high dimensional inputs by decomposing the full dimensional model into the sum of several components (non-overlapping) each with a subset of the original dimensions. Gardner et al. (2017) develops a decomposition approach that, combined with Markov chain Monte Carlo (MCMC), attempts to learn and exploit the additive structure of the true function. If the true function is indeed additive, then each model can be uniquely defined by the additive decomposition of its underlying dimensions. However, given the exponentially growing size of possible additive decompositions, it remains challenging to evaluate all possible models. Metropolis-Hasting sampling was used to obtain a proposal distribution over all the possible partitions of a given model. For each sampled partition the additive GP is estimated, and the likelihood of the model is used as evidence associated to the specific partition. Such evidence is used to update the posterior over the space of all possible dimensions combinations. Given an additive model with dimensions forming the partition P, the kernel can be formulated as $k(\mathbf{x}, \mathbf{x}^*) = \sum_{i=1}^{|P|} k(\mathbf{x}[P_i], \mathbf{x}'[P_i])$. As an example, given a 4-dimensinoal space, $P = \{[1], [2,3], [4]\}$, the additive model will be $f(\mathbf{x}) = f_1(x_1) + f_2(x_2, x_3) + f_3(x_4)$. The predictive distribution, given k models is:

$$p\left(f\left(\mathbf{x}^*\right) | D, \mathbf{x}^*\right) \approx \frac{1}{k} \sum_{j=1}^{k} p\left(f\left(\mathbf{x}^*\right) | D, \mathbf{x}^*, M_j\right)$$

where $D = \{\mathbf{X}, \mathbf{f}\}$ is the training set $\mathbf{X}$, and evaluations $\mathbf{f}$. M_j is a model sampled from the proposal, i.e., a specific additive decomposition, and $p\left(f\left(\mathbf{x}^*\right) | D, \mathbf{x}^*, M_j\right)$ represents the posterior probability to observe the function value $f(\mathbf{x}^*)$ conditional on the samples D, and the chosen additive model M_j.

4 NUMERICAL EXPERIMENTS

4.1 Experiment Design

As previously mentioned, this preliminary comparison presents the results over deterministic versions of the three models, GE-KPLS, SGPMA, and GPMCMC introduced in Sections 3.2, 3.1, and 3.3, respectively. We tested the approaches against three analytic functions, the Griewank (Griewank 1981) which has widely spread and regularly distributed local minima, the Styblinski-Tang (Styblinski and Tang 1990), a multimodal additive function, and the Michalewicz with parameter that describes the valley and ridges steepness set to 10 (Molga, Marcin and Smutnicki, Czesaw), a multimodal additive function with $d!$ local minima, where d is dimensionality. The metrics used to compare the performance of each model are root mean squared error (RMSE) and mean absolute error (MAE), which are commonly used error metrics to measure prediction accuracy. Runtime (in seconds) is also considered to understand the computation efficiency of each model.

Experiment results are presented in two parts, the individual model performance (Section 4.2), and the comparison across different approaches (Section 4.3). We consider different combinations of sample sizes (n) and input dimensions (d) in ascending order, averaged over 50 macro-replications. The computing resources used for all experiments were run on the AGAVE computing cluster. We consider the following possible input parametrizations for the different methods:

- Dimensionality was set to $d = 10, 20, 30, 50, 100, 150, 200$;
- Choices of sample size n based on dimensionality include $10d$, $20d$;
- Concerning method-dependent parameters, we considered the following:
  - GE-KPLPS: number of approximating points used in PLS, $n_{ext} = 0, 2, 4$;
  - SGPMA: number of subsets, $n_{sub} = 0.2d, 0.5d$; embedding method, π^{pca}, π^{random}; prediction power parameter, $\eta = 5, 8, 12$;
  - GPMCMC: number of models sampled from MCMC, $k = 5, 20, 50$; number of MCMC iterations was set to 500 for all experiments.

Hence, a total of $(168 + 36 + 4) \times 3 = 624$ experiments were run, each with 50 replications. Implementations can be found on Github.

4.2 Individual Algorithms Analysis

In the following we analyze the performance of GE-KPLS (Section 4.2.1), SGPMA (Section 4.2.2), and GPMCMC (Section 4.2.3). Since RMSE and MAE showed similar patterns across the cases, we are only providing the results for the RMSE.

4.2.1 Gradient-enhanced Kriging with Partial Least Squares

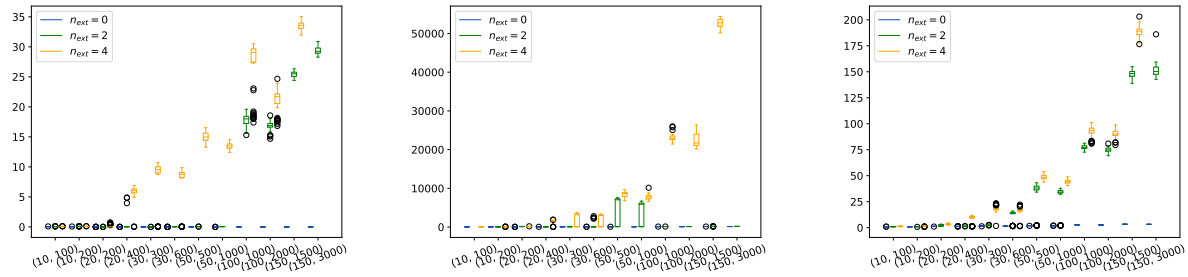
The prediction performance for GE-KPLS is presented in Figure 2.

While $n_{ext} = 2 - 4$ results in large error, not adding such points improves performance. As dimensionality goes above 10, the RMSE is below $1e$-3 (similar to SGPMA for $d = 10$). This observation applies to both choices of n.

GE-KPLS shows a consistent behavior in terms of runtime (Figure 3): increased dimensionality with sample size hurts runtime, consistently across different test functions. The experiments only ran up to $d = 150$ due to computation overhead.

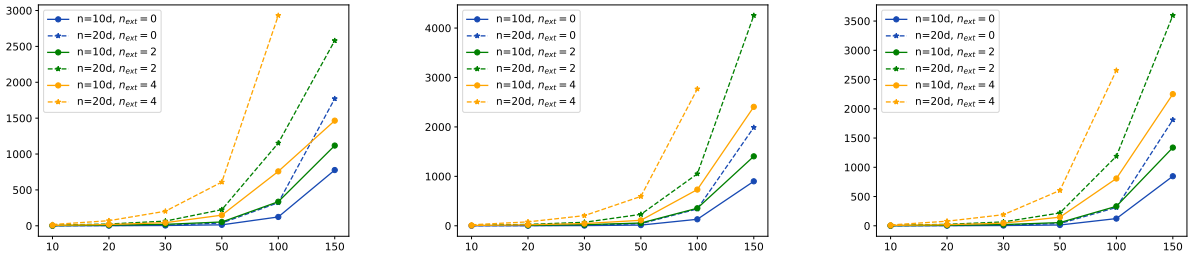
4.2.2 Stochastic Gaussian Process with Model Averaging

Figures 4–5 show the performance of SGPMA with PCA embedding (similar results were observed for the random projection case). The error decreases and the runtime increases with dimensionality and data sizes, with an obviously larger variance associated with the RMSE values for the Griewank function.



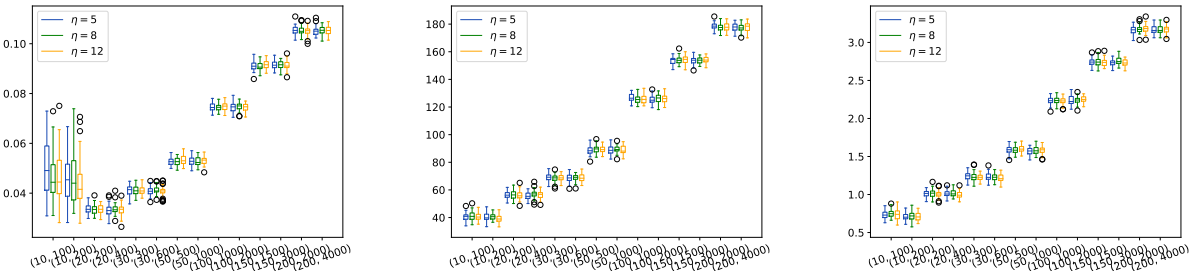
(a) Griewank, RMSE values. (b) Styblinski-Tang, RMSE values. (c) Michalewicz, RMSE values.

Figure 2: RMSE (y-axis) as a function of dimensionality and sample size (d,n) (x-axis) for GE-KPLS.



(a) Griewank, runtime values. (b) Styblinski-Tang, runtime values. (c) Michalewicz, runtime values.

Figure 3: Runtime (y-axis) as a function of dimensionality d (x-axis) Performance for GE-KPLS.



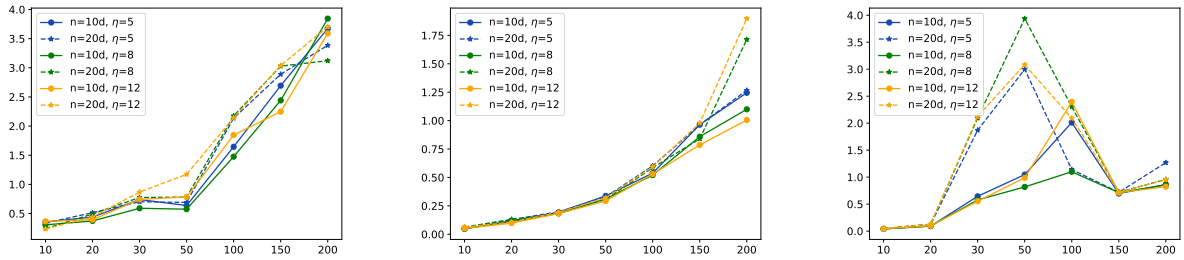
(a) Griewank function, RMSE values. (b) Styblinski-Tang, RMSE values. (c) Michalewicz, RMSE values.

Figure 4: RMSE (y-axis) as a function of dimensionality and sample size (d,n) (x-axis) for SGPMA.

The runtime behavior associated to the Michalewicz function shows abnormal peaks at $d = 50$ and 100 with an inverse relationship between data size/dimension and runtime. Since experiments were replicated multiple runs, it is unlikely for this behavior to be caused by randomness. Instead, we infer that it might be related to certain conditions of the computing cluster, and how the data is allocated to subsets.

4.2.3 Gaussian Process with Monte Carlo Markov Chain

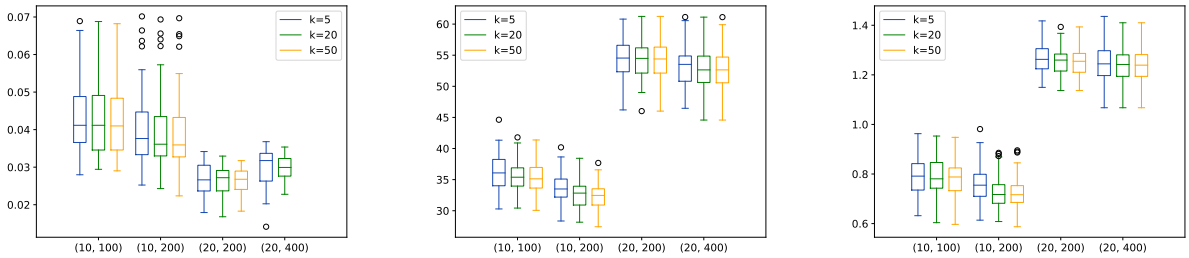
GPMCMC shows good performance in terms of error (see Figure 6). In the RMSE plots, different color represents the different number of models k sampled used for prediction, but since this parameter is not



(a) Griewank, runtime values. (b) Styblinski-Tang, runtime values. (c) Michalewicz, runtime values.

Figure 5: Runtime (y-axis) as a function of dimensionality d (x-axis) for SGPMA.

of concern during modeling, varing its value does not have any impact in runtime. We observe that the RMSE of Griewank decreases as d increases, given fixed n. This observation is consistent with the unusual behavior of the performance other two models with Griewank when $d = 10$ and 20, thus, given the results so far, we infer it is a consequence of the Griewank function performance in lower-dimensional settings.



(a) Griewank, RMSE values. (b) Styblinski-Tang, RMSE values. (c) Michalewicz, RMSE values.

Figure 6: RMSE (y-axis) as a function of dimensionality, sample size (d,n) (x-axis) for GPMCMC.

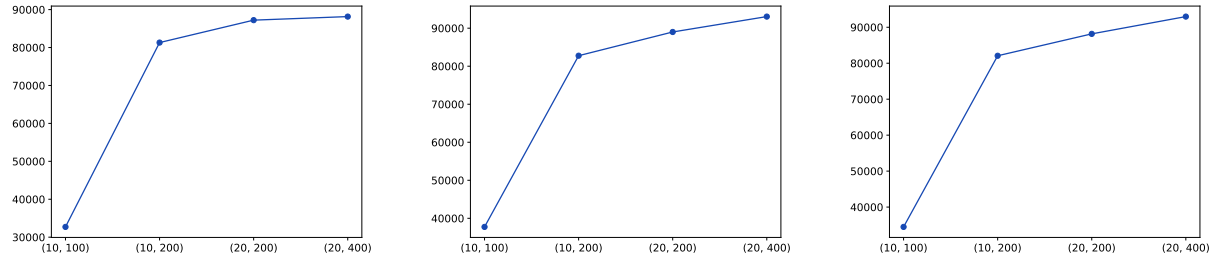
However, Figure 7 shows how the runtime is extremely high even for one iteration of the MCMC algorithm. As a result, we only tested this algorithm up to $n = 400$ and $d = 20$. Since the prameter k does not influence the runtime, the plots in Figure 7 only report one value for each (d,n) condition. We can see that the runtime plots show an extremely large gap between $n = 10$ and $n = 20$ for $d = 10$. This gives some insight on the promising performance of GPMCMC only restricted to when both dimension and sample size are small; as soon as either gets larger, the runtime performance decays tremendously.

4.3 Comparative Algorithms Analysis

The output statistics, averaged across 50 macro-replications, are summarized in Table 1 for a comprehensive analysis of the three models. SGPMA and GE-KPLS outputs were obtained from the model with parameters that returned the best accuracy and runtime, i.e., $\eta = 5$ and $n_{sub} = 0.5d$ subsets, and $n_{ext} = 0$, respectively. For dimensions up to 20, all three models have equally good accuracy, while SGPMA has the best runtime and GPMCMC performs extremely poorly. Both SGPMA and GE-KPLS scale well to increasing dimensionality and sample size, but the trade-off between good accuracy and efficient runtime is noticeable.

5 CONCLUSIONS AND FUTURE WORKS

In this paper, three HDGP models were implemented and compared in their deterministic version, which provides a foundation to the usage of GP models in high dimensionality with big data size. The experiment



(a) Griewank, runtime values. (b) Styblinski-Tang, runtime values. (c) Michalewicz, runtime values.

Figure 7: Runtime (y-axis) as a function of dimensionality and sample size (d,n) (x-axis) for GPMCMC.

Table 1: Performance comparison; $*$ means the model could not be run for certain dimensionality d and sample size n; statistically better performance are highlighted in **bold**.

Test function	n		$10*d$							$20*d$						
	d		10	20	30	50	100	150	200	10	20	30	50	100	150	200
Griewank	RMSE	SGPMA	.045	.033	.040	.052	.074	.091	.105	.044	.032	.039	.052	.074	.091	.105
		GE-KPLS	.044	**.001**	.002	.003	.004	.005	*	.040	.0005	**.00005**	.0008	.002	.002	*
		GPMCMC	.043	.026	*	*	*	*	*	.039	.030	*	*	*	*	*
	MAE	SGPMA	.029	.026	.032	.042	.059	.072	.084	.028	.026	.031	.041	.059	.073	.084
		GE-KPLS	.025	**.0008**	.001	.002	.003	.004	*	.023	.0003	**.00004**	.0006	.001	.002	*
		GPMCMC	.026	.021	*	*	*	*	*	.022	.024	*	*	*	*	*
	Runtime	SGPMA	**.326**	.436	.715	.741	1.87	2.80	3.34	**.22**	.454	.702	.9	1.86	2.78	3.96
		GE-KPLS	.56	1.89	4.08	13.3	123	777	*	1.26	4.23	7.27	38.8	322	1771	*
		GPMCMC	32706	87221	*	*	*	*	*	81308	88158	*	*	*	*	*
Styblinski-Tang	RMSE	SGPMA	39.2	56.8	69.5	89.1	126	154	177	39.1	56.0	68.2	89.2	125	154	177
		GE-KPLS	**35.4**	48.9	59.4	75.9	106	128	*	34.8	48.3	58.1	74.1	103	126	*
		GPMCMC	35.5	54.1	*	*	*	*	*	**32.3**	52.9	*	*	*	*	*
	MAE	SGPMA	31.6	45.3	55.7	71.1	101	124	142	31.6	44.8	54.7	71.2	100	123	142
		GE-KPLS	**28.3**	39.3	47.3	60.6	84.3	102	*	27.7	38.5	46.5	59.1	81.9	100	*
		GPMCMC	28.4	43.1	*	*	*	*	*	**25.9**	42.2	*	*	*	*	*
	Runtime	SGPMA	**.05**	.125	.184	.318	.535	.653	.954	**.068**	.134	.201	.272	.601	.864	1.25
		GE-KPLS	.576	2.03	3.99	13.4	131	900	*	1.55	4.81	10.6	42.0	341	1989	*
		GPMCMC	37736	89002	*	*	*	*	*	82763	93050	*	*	*	*	*
Michalewicz	RMSE	SGPMA	**.729**	1.04	1.25	1.58	2.25	2.76	3.17	**.711**	1	1.22	1.59	2.25	2.74	3.17
		GE-KPLS	.83	1.19	1.45	1.89	2.66	3.27	*	.783	1.16	1.43	1.85	2.62	3.17	*
		GPMCMC	.781	1.26	*	*	*	*	*	.725	1.25	*	*	*	*	*
	MAE	SGPMA	**.587**	.832	1	1.27	1.80	2.21	2.53	**.575**	.804	.979	1.27	1.80	2.19	2.53
		GE-KPLS	.669	.961	1.16	1.51	2.13	2.61	*	.632	.927	1.15	1.47	2.09	2.53	*
		GPMCMC	.629	1.01	*	*	*	*	*	.581	.995	*	*	*	*	*
	Runtime	SGPMA	**.040**	.097	.649	.93	2.20	.687	.855	**.051**	.129	2.05	3.56	4.37	.719	.946
		GE-KPLS	.536	2.08	4.24	13.4	123	847	*	1.29	4.44	10.4	38.3	314	1812	*
		GPMCMC	34485	88162	*	*	*	*	*	82075	92966	*	*	*	*	*

results show that SGPMA and GEKPLS have acceptable runtime while maintaining good accuracy even when scaled up. GPMCMC has good accuracy but with much worse runtime. All algorithms fail to achieve as good performance for Styblinski-Tang as the other two functions. Therefore, it is of our interest to investigate into additional functions to find out how intrinsic characteristics impact model performance. Based on this preliminary comparison, future directions would include an extension to the stochastic models over an additional set of functions, as well as improving existing framework by automating choices of SGPMA parameters for better results.

REFERENCES

Bickel, P. J., Y. Ritov, and A. B. Tsybakov. 2009. "Simultaneous Analysis of Lasso and Dantzig Selector". *The Annals of Statistics* 37(4):1705–1732.

Bouhlel, M. A., N. Bartoli, A. Otsmane, and J. Morlier. 2016. "Improving Kriging Surrogates of High-dimensional Design Models by Partial Least Squares Dimension Reduction". *Structural and Multidisciplinary Optimization* 53(5):935–952.

Bouhlel, M. A., and J. R. Martins. 2019. "Gradient-enhanced Kriging for High-dimensional Problems". *Engineering with Computers* 35(1):157–173.

Bouveyron, C., S. Girard, and C. Schmid. 2007. "High-dimensional Discriminant Analysis". *Communications in Statistics—Theory and Methods* 36(14):2607–2623.

Chen, L., H. Qiu, L. Gao, Z. Yang, and D. Xu. 2021. "Exploiting Active Subspaces of Hyperparameters for Efficient High-dimensional Kriging Modeling". *Mechanical Systems and Signal Processing*:108643.

Cox, T. F., and M. A. Cox. 1991. "Multidimensional Scaling on a Sphere". *Communications in Statistics-Theory and Methods* 20(9):2943–2953.

Donoho, D. L. 2006. "High-dimensional Centrally Symmetric Polytopes with Neighborliness Proportional to Dimension". *Discrete & Computational Geometry* 35(4):617–652.

Donoho, D. L. et al. 2000. "High-dimensional Data Analysis: The Curses and Blessings of Dimensionality". *AMS math Challenges Lecture* 1(2000):32.

Durrande, N., D. Ginsbourger, and O. Roustant. 2011. "Additive Kernels for Gaussian Process Modeling". *arXiv preprint arXiv:1103.4023*.

Duvenaud, D. K., H. Nickisch, and C. Rasmussen. 2011. "Additive Gaussian Processes". *Advances in Neural Information Processing Systems* 24:226–234.

Eriksson, D., M. Pearce, J. Gardner, R. D. Turner, and M. Poloczek. 2019. "Scalable Global Optimization via Local Bayesian Optimization". *Advances in Neural Information Processing Systems* 32:5497–5508.

Fan, J., and R. Li. 2006. "Statistical Challenges with High Dimensionality". In *Proceedings of the international Congress of Mathematicians*. August 22nd-30th, Madrid, Spain.

Fan, J., and J. Lv. 2008. "Sure Independence Screening for Ultrahigh Dimensional Feature Space". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5):849–911.

Fan, J., and J. Lv. 2010. "A Selective Overview of Variable Selection in High Dimensional Feature Space". *Statistica Sinica* 20(1):101.

Fan, J., R. Samworth, and Y. Wu. 2009. "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model". *The Journal of Machine Learning Research* 10:2013–2038.

Gardner, J., C. Guo, K. Weinberger, R. Garnett, and R. Grosse. 2017. "Discovering and Exploiting Additive Structure for Bayesian Optimization". In *Artificial Intelligence and Statistics*. April 20th-22th, Florida, USA, 1311–1319.

Gardner, J., G. Pleiss, R. Wu, K. Weinberger, and A. Wilson. 2018. "Product Kernel Interpolation for Scalable Gaussian Processes". In *International Conference on Artificial Intelligence and Statistics*. April 9th-11th, Playa Blanca, Lanzarote, 1407–1416.

Greenshtein, E. 2006. "Best Subset Selection, Persistence in High-dimensional Statistical Learning and Optimization under L1 Constraint". *The Annals of Statistics* 34(5):2367–2386.

Greenshtein, E., and Y. Ritov. 2004. "Persistence in High-dimensional Linear Predictor Selection and the Virtue of Over-parametrization". *Bernoulli* 10(6):971 – 988.

Griewank, A. O. 1981. "Generalized Descent for Global Optimization". *Journal of Optimization Theory and Applications* 34(1):11–39.

Johnstone, I. M., and D. M. Titterington. 2009. "Statistical Challenges of High-dimensional Data". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367(1906):4237–4253.

Kandasamy, K., J. Schneider, and B. Póczos. 2015. "High Dimensional Bayesian Optimisation and Bandits via Additive Models". In *International Conference on Machine Learning*. July 6th-11th, Lille, France, 295–304.

Lee, Y. K., E. R. Lee, and B. U. Park. 2012. "Principal Component Analysis in Very High-dimensional Spaces". *Statistica Sinica*:933–956.

Letham, B., R. Calandra, A. Rai, and E. Bakshy. 2020. "Re-examining Linear Embeddings for High-dimensional Bayesian Optimization". *Advances in Neural Information Processing Systems* 33:1546–1558.

Liu, H., Y.-S. Ong, X. Shen, and J. Cai. 2020. "When Gaussian Process Meets Big Data: A Review of Scalable GPs". *IEEE Transactions on Neural Networks and Learning Systems* 31(11):4405–4423.

Mathesen, L., G. Pedrielli, S. H. Ng, and Z. B. Zabinsky. 2021. "Stochastic Optimization with Adaptive Restart: A Framework for Integrated Local and Global Learning". *Journal of Global Optimization* 79(1):87–110.

Meinshausen, N., and P. Bühlmann. 2006. "High-dimensional Graphs and Variable Selection with the Lasso". *The Annals of Statistics* 34(3):1436–1462.

Meng, Q., and S. H. Ng. 2015. "An Additive Global and Local Gaussian Process Model for Large Data Sets". In *In Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 505–516. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Molga, Marcin and Smutnicki, Czesaw. "Test Functions for Optimization Needs, 2005". URLhttp://www.zsd.ict.pwr.wroc.pl/files/docs/functions.pdf, accessed 20nd April 2022.

Nayebi, A., A. Munteanu, and M. Poloczek. 2019. "A Framework for Bayesian Optimization in Embedded Subspaces". In *International Conference on Machine Learning*. June 9th-15th, California, USA, 4752–4761.

Pang, Y., Y. Yuan, and X. Li. 2008. "Effective Feature Extraction in High-dimensional Space". *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38(6):1652–1656.

Reddy, G. T., M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker. 2020. "Analysis of Dimensionality Reduction Techniques on Big Data". *IEEE Access* 8:54776–54788.

Roweis, S. T., and L. K. Saul. 2000. "Nonlinear Dimensionality Reduction by Locally Linear Embedding". *Science* 290(5500):2323–2326.

Santner, T. J., B. J. Williams, and W. I. Notz. 2013. *The Design and Analysis of Computer Experiments*. Berlin, Springer Science & Business Media.

Snelson, E., and Z. Ghahramani. 2006. "Sparse Gaussian processes using pseudo-inputs". *Advances in Neural Information Processing Systems* 18:1259–1266.

Sorzano, C. O. S., J. Vargas, and A. P. Montano. 2014. "A Survey of Dimensionality Reduction Techniques". *arXiv preprint arXiv:1403.2877*.

Styblinski, M., and T.-S. Tang. 1990. "Experiments in Nonconvex Optimization: Stochastic Approximation with Function Smoothing and Simulated Annealing". *Neural Networks* 3(4):467–483.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.

Titsias, M. 2009. "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In *Artificial Intelligence and Statistics*. April 16th-18th, Florida, USA, 567–574.

Van Der Maaten, L., E. Postma, J. Van den Herik et al. 2009. "Dimensionality Reduction: a Comparative". *Journal of Machine Learning Research* 10(66-71):13.

Wainwright, M. J. 2006. "Sharp Thresholds for High-dimensional and Noisy Recovery of Sparsity". *arXiv preprint math/0605740*.

Wang, Z., F. Hutter, M. Zoghi, D. Matheson, and N. de Feitas. 2016. "Bayesian Optimization in a Billion Dimensions via Random Embeddings". *Journal of Artificial Intelligence Research* 55:361–387.

Wasserman, L., and K. Roeder. 2009. "High Dimensional Variable Selection". *Annals of Statistics* 37(5A):2178.

Wilson, A., and H. Nickisch. 2015. "Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP)". In *International Conference on Machine Learning*. July 6th-11th, Lille, France, 1775–1784.

Xuereb, M., T. M. Huo, and S. H. Ng. 2019. "Principal Component Analysis for High Dimension Stochastic Gaussian Process Model Fitting". In *2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 632–636. IEEE.

Xuereb, M., S. H. Ng, and G. Pedrielli. 2020. "Stochastic Gaussian Process Model Averaging for High-dimensional Inputs". In *In Proceedings of the 2020 Winter Simulation Conferenc*, edited by K.–H. Bae, B. Feng, S. Kim, S. Lazarova–Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 373–384. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Yin, J., S. H. Ng, and K. M. Ng. 2011. "Kriging Metamodel with Modified Nugget-effect: The Heteroscedastic Variance Case". *Computers & Industrial Engineering* 61(3):760–777.

Zhao, P., and B. Yu. 2006. "On Model Selection Consistency of Lasso". *The Journal of Machine Learning Research* 7:2541–2563.

Zhu, X., Z. Huang, H. T. Shen, J. Cheng, and C. Xu. 2012. "Dimensionality Reduction by Mixed Kernel Canonical Correlation Analysis". *Pattern Recognition* 45(8):3003–3016.

AUTHOR BIOGRAPHIES

MENGRUI (MINA) JIANG is a Ph.D. student in the School of Computing and Augmented Intelligence at Arizona State University. Her email address is mjiang42@asu.edu.

GIULIA PEDRIELLI is Assistant Professor in the School of Computing and Augmented Intelligence at Arizona State University. Her email address is giulia.pedrielli@asu.edu.

SZU HUI NG is Associate Professor and Department Head of the Department of Industrial Systems Engineering and Management at National University of Singapore. Her email address is isensh@nsu.edu.sg.