A Sequential Split-Conquer-Combine Approach for Gaussian Process Modeling in Computer Experiments

Chengrui Li
Department of Statistics and Biostatistics, Rutgers University
Joint work with Ying Hung and Min-ge Xie

2017 QPRC
JUNE 13, 2017
Outline

- Introduction
- A Unified Framework with Theoretical Supports
- Simulation Study
- Real Data Example
- Summary
Introduction
Motivating example: Data center thermal management

- A **data center** is an integrated facility housing multiple-unit servers, providing application services or management for data processing.
- **Goal:** Design a data center with an efficient heat removal mechanism.
- Computational Fluid Dynamics (CFD) simulation \((n = 26820, p = 9)\)

**Figure 1:** Heat map for IBM T. J. Watson Data Center
Gaussian process model

• Gaussian process (GP) model:

\[ y = X\beta + Z(x), \]

- **y**: \( n \times 1 \) vector of observations (e.g., room temperatures)
- **X**: \( n \times p \) design matrix
- **\beta**: \( p \times 1 \) unknown parameters
Gaussian process model

- Gaussian process (GP) model:

\[ y = X\beta + Z(x), \]

- \( y \): \( n \times 1 \) vector of observations (e.g., room temperatures)
- \( X \): \( n \times p \) design matrix
- \( \beta \): \( p \times 1 \) unknown parameters
- \( Z(x) \) is a GP process with mean 0 and covariance \( \sigma^2 \Sigma(\theta) \)
- \( \Sigma(\theta) \): \( n \)-by-\( n \) Correlation matrix with correlation parameters \( \theta \)
  The \( ij \)th element of \( \Sigma \) is defined by a power exponential function

\[
corr(Z(x_i), Z(x_j)) = \exp(-\theta^T|x_i - x_j|) = \exp(- \sum_{k=1}^{p} \theta_k |x_{ik} - x_{jk}|).
\]

- **Remark**: Assume \( \sigma \) is known for simplicity in this talk.
• Likelihood inference

\[ l(\beta, \theta, \sigma) = -\frac{1}{2\sigma^2} (y - X\beta)^\top \Sigma^{-1}(\theta) (y - X\beta) - \frac{1}{2} \log |\Sigma(\theta)| - \frac{n}{2} \log(\sigma^2) \]

So,

\[ \hat{\beta} | \theta = \arg \max_{\beta} \{ l(\beta | \theta, \sigma^2) \} = (X^\top \Sigma^{-1}(\theta)X)^{-1} X^\top \Sigma^{-1}(\theta)y \]

\[ \hat{\theta} | \hat{\beta}, \sigma^2 = \arg \max_{\theta} \{ l(\theta | \hat{\beta}, \sigma^2) \} \]
Estimation and prediction

- Likelihood inference

\[
l(\beta, \theta, \sigma) = -\frac{1}{2\sigma^2} (y - X\beta)^\top \Sigma^{-1}(\theta)(y - X\beta) - \frac{1}{2} \log |\Sigma(\theta)| - \frac{n}{2} \log(\sigma^2)
\]

So,

\[
\hat{\beta}|\theta = \arg\max_{\beta} \{l(\beta|\theta, \sigma^2)\} = (X^\top \Sigma^{-1}(\theta)X)^{-1} X^\top \Sigma^{-1}(\theta)y
\]

\[
\hat{\theta}|\hat{\beta}, \sigma^2 = \arg\max_{\theta} \{l(\theta|\hat{\beta}, \sigma^2)\}
\]

- GP prediction, say \(y_0\), at a new point \(x_0\), given parameters \((\beta, \theta)\), follows a normal distribution with mean \(p_0(\beta, \theta)\) and variance \(m_0(\beta, \theta)\), where,

\[
p_0(\beta, \theta) = x_0^\top \beta + \gamma(\theta)^\top \Sigma^{-1}(\theta)(y - X\beta)
\]

\[
m_0(\beta, \theta) = \sigma^2 (1 - \gamma(\theta)^\top \Sigma^{-1}(\theta)\gamma(\theta)),
\]

and \(\gamma(\theta)\) is a \(n \times 1\) vector of \(i\)th element equals to \(\phi(||x_i - x_0||; \theta)\).
Two challenges in GP modeling

Computational issue:
- Estimation and prediction involve $\Sigma^{-1}$ and $|\Sigma|$ with order of $O(n^3)$:
- Not feasible when $n$ is large
Two challenges in GP modeling

😊 Computational issue:
  • Estimation and prediction involve $\Sigma^{-1}$ and $|\Sigma|$ with order of $O(n^3)$:
  • Not feasible when $n$ is large

😊 Uncertainty quantification of GP predictor
  • Plug-in predictive distribution is widely used
  • It underestimates the uncertainty
Existing methods

For the computational issue:
- Change the model to one that is computationally convenient: Rue and Held (2005), Cressie and Johannesson (2008).
- **Not focus on uncertainty quantification and bring in addition uncertainty**

For uncertainty quantification of GP predictor:
- Bayesian predictive distribution
- Bootstrap approach (Luna and Young 2003)
- **Intensive computation**
Solve both problems by a unified framework?

• Yes!
A Unified Framework
Introduction to confidence distribution (CD)

Statistical inference (Parameter estimation):

• Point estimate
• Interval estimate
• Distribution estimate

Example: \(X_1, \ldots, X_n\) i.i.d. follows \(N(\mu, 1)\)

• Point estimate: \(\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i\)
• Interval estimate: \((\bar{x}_n - 1.96/\sqrt{n}, \bar{x}_n + 1.96/\sqrt{n})\)
• Distribution estimate: \(N(\bar{x}_n, \frac{1}{n})\)
Introduction to confidence distribution (CD)

Statistical inference (Parameter estimation):

• Point estimate
• Interval estimate
• Distribution estimate

Example: $X_1, \ldots, X_n$ i.i.d. follows $N(\mu, 1)$

• Point estimate: $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$
• Interval estimate: $(\bar{x}_n - 1.96/\sqrt{n}, \bar{x}_n + 1.96/\sqrt{n})$
• Distribution estimate: $N(\bar{x}_n, \frac{1}{n})$

The idea of the CD approach is to use a sample-dependent distribution (or density) function to estimate the parameter of interest.

• Wide range of examples: bootstrap distribution, (normalized) likelihood function, $p$-value functions, fiducial distributions, some informative priors and Bayesian posteriors, among others

(Xie and Singh 2013)
Overview: Sequential Split-Conquer-Combine

Figure 2: Sequential Split-Conquer-Combine Approach
Split the entire dataset into subsets (correlated) based on compact support correlation assumption for 1-D
Ingredients

- Split the entire dataset into subsets (correlated) based on compact support correlation assumption for 1-D
- Perform a sequential updating to create independent subsets and estimate on each updated subsets
Ingredients

- Split the entire dataset into subsets (correlated) based on compact support correlation assumption for 1-D
- Perform a sequential updating to create independent subsets and estimate on each updated subsets
- Combine estimators
Ingredients

- Split the entire dataset into subsets (correlated) based on compact support correlation assumption for 1-D
- Perform a sequential updating to create independent subsets and estimate on each updated subsets
- Combine estimators
- Quantify prediction uncertainty
Split the entire dataset into subsets $y = \{y_a\}$, $a = 1, ..., m$. Denote the size of $y_a$ by $n_a$, i.e. $\sum n_a = n$.

- Assumption: compactly supported correlation

$$
\Sigma_t = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} & \cdots & O & O \\
\Sigma_{21} & \Sigma_{22} & \ddots & \cdots & O \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & \Sigma_{(m-1)(m-1)} & \Sigma_{(m-1)m} \\
0 & 0 & \cdots & \Sigma_{m(m-1)} & \Sigma_{mm}
\end{pmatrix}_{n \times n},
$$

(after index sorting according to $X_1$ values)
Transform $y$ to $y^*$ by sequentially updating:

$$y_a^* = y_a - L_{a(a-1)}y_{a-1},$$

where $L_{(a+1)a} = \Sigma_{(a+1)a}D_{a}^{-1}$, $D_a = \Sigma_{aa} - L_{a(a-1)}D_{(a-1)}L_{a(a-1)}^T$.

• Sequential updates are **computationally efficient**.
• The updated block $y_a^*$’s are independent.
Estimation from each subset

Given $\theta$, we have

- MLE of the $a$th subset:

$$\hat{\beta}_a = \arg\max_{\beta|\theta} l_t^{(a)}(\beta|\theta) = (C_a^T D_a^{-1} C_a)^{-1} C_a^T D_a^{-1} y_a^*.$$  

- An individual CD for the $a$th updated subset is (cf., Xie and Singh 2013): $N_p(\hat{\beta}_a, \text{Cov}(\hat{\beta}_a))$.

Given $\beta$, we have

- MLE of the $a$th subset: $\hat{\theta}_a = \arg\max_{\theta} l_t^{(a)}(\theta|\beta)$.

- Given $\beta$, an individual CD for the $a$th updated subset is $N(\hat{\theta}_a, \text{Cov}(\hat{\theta}_a))$.

Significant computational reduction because $D_a$ is much smaller than the original covariance matrix.
CD combining

• Following Singh, Xie and Strawderman (2005), Liu, Liu and Xie (2014) and Yang et al. (2014), a combined CD is $N_p(\beta_c, S_c)$, where
  \[ \hat{\beta}_c = \left( \sum W_a \right)^{-1} \left( \sum W_a \hat{\beta}_a \right) \]
  with $W_a = (C_a^T D_a^{-1} C_a)^{-1}$ and $S_c = \text{Cov}(\hat{\beta}_c)$.

• Similar framework can be applied to all the parameters $(\beta, \theta)$. 
Following Singh, Xie and Strawderman (2005), Liu, Liu and Xie (2014) and Yang et al. (2014), a combined CD is $N_p(\beta_c, S_c)$, where

$$\hat{\beta}_c = \left( \sum W_a \right)^{-1}\left( \sum W_a \hat{\beta}_a \right)$$

with $W_a = \left( C_a^T D_a^{-1} C_a \right)^{-1}$ and $S_c = \text{Cov}(\hat{\beta}_c)$.

Similar framework can be applied to all the parameters $(\beta, \theta)$.

**Theorem 1**

Under some regularity assumptions, when $\tau > O_p(n^{1/2})$ and $n \to \infty$, the SSCC estimator $\hat{\lambda}_c = (\hat{\beta}_c, \hat{\theta}_c)$ is asymptotically as efficient as MLE $\hat{\lambda}_{mle} = (\hat{\beta}_{mle}, \hat{\theta}_{mle})$. 

GP predictive distribution

• GP predictor at a new point \( x_0 \), given parameters \((\beta, \theta)\), follows a normal distribution with mean \( p_0(\beta, \theta) \) and variance \( m_0(\beta, \theta) \), where,

\[
p_0(\beta, \theta) = x_0^\top \beta + \gamma(\theta)^\top \Sigma^{-1}(\theta) (y - X\beta)
\]

\[
m_0(\beta, \theta) = \sigma^2 (1 - \gamma(\theta)^\top \Sigma^{-1}(\theta) \gamma(\theta)),
\]

and \( \gamma(\theta) \) is a \( n \times 1 \) vector of ith element equals to \( \phi(||x_i - x_0||; \theta) \).
GP predictive distribution

• GP predictor at a new point $x_0$, given parameters $(\beta, \theta)$, follows a normal distribution with mean $p_0(\beta, \theta)$ and variance $m_0(\beta, \theta)$, where,

\[
p_0(\beta, \theta) = x_0^T \beta + \gamma(\theta)^T \Sigma^{-1}(\theta)(y - X\beta)
\]
\[
m_0(\beta, \theta) = \sigma^2 (1 - \gamma(\theta)^T \Sigma^{-1}(\theta)\gamma(\theta)),
\]

and $\gamma(\theta)$ is a $n \times 1$ vector of $i$th element equals to $\phi(||x_i - x_0||; \theta)$.

• Drawbacks:
  - ☹  Computational issue again!
  - 😄  Conventional plug-in predictive distribution underestimate the uncertainty
Quantify GP prediction uncertainty using SSCC and CD

An easy-to-compute GP predictor:

\[
p_1(\beta, \theta) = \mathbf{x}_0^T \beta + \sum_{a=1}^{m} \gamma_a^*(\theta)^T D_a^{-1}(\theta) \mathbf{y}_a^* + \sum_{a=1}^{m} \gamma_a^*(\theta)^T D_a^{-1}(\theta) C_a(\theta) \beta
\]

\[
m_1(\beta, \theta) = \sigma^2 (1 - \sum_{a=1}^{m} \gamma_a^*(\theta)^T D_a^{-1}(\theta) \gamma_a^*(\theta))
\]
Quantify GP prediction uncertainty using SSCC and CD

😊 An easy-to-compute GP predictor:

\[ p_1(\beta, \theta) = \mathbf{x}_0^T \beta + \sum_{a=1}^{m} \gamma^*_a(\theta)^T D_a^{-1}(\theta) \mathbf{y}_a^* + \sum_{a=1}^{m} \gamma^*_a(\theta)^T D_a^{-1}(\theta) C_a(\theta) \beta \]

\[ m_1(\beta, \theta) = \sigma^2 (1 - \sum_{a=1}^{m} \gamma^*_a(\theta)^T D_a^{-1}(\theta) \gamma^*_a(\theta)) \]

😊 A better way to quantify uncertainty: use a CD-based predictive distribution function:

\[ Q(y_0; \mathbf{y}) = \int_{\lambda \in \Theta} G_\lambda(y_0) dH_c(\lambda; \mathbf{y}), \]

where \( G_\lambda(\cdot) \) is the CDF of \( N(p_1(\lambda), m_1(\lambda)) \), and \( H_c(\lambda; \mathbf{y}) \) is joint CD of \( \lambda = (\beta, \theta) \) obtained from our SSCC approach
Theoretical support

**Theorem 2**
Under some regularity assumptions, when $\tau > O_p(n^{1/2})$ and $n \to \infty$,

$$p_1(\beta, \theta) \to p_0(\beta, \theta), \quad m_1(\beta, \theta) \to m_0(\beta, \theta)$$

- The easy-to-compute GP predictor is asymptotically equivalent to the original predictor
- Compared with the plug-in predictive distribution, the predictive confidence distribution has smaller average Kullback-Leibler distance to the true predictive distribution (Shen, Liu, and Xie 2017+).
Simulation Study
Suppose we have the following GP model,

$$y = X\beta + Z(X),$$

- Sample points: $n = 1000, 1500, 2000$; Four covariates $p = 4$; $x \in [0, 1]^4$
- True parameter values: $\beta = (2, 3, 1, 2, 1.5)$, $\theta = (15, 1.5, 2, 3)$ and $\sigma^2 = 1$
- The $ij$th element of $\Sigma(\theta)$ is computed by
  $$\sigma_{ij} = \exp\left(- \sum_{k=1}^{p} \theta_k |x_{ik} - x_{jk}| \right)$$
- The number of splits: $m = 5$
Simulation study: analysis results

<table>
<thead>
<tr>
<th></th>
<th>$n = 1000$</th>
<th></th>
<th>$n = 1500$</th>
<th></th>
<th>$n = 2000$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>Compact</td>
<td>SCC</td>
<td>MLE</td>
<td>Compact</td>
<td>SCC</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>1.96(0.33)</td>
<td>1.96(0.33)</td>
<td>1.96(0.33)</td>
<td>2.00(0.30)</td>
<td>2.00(0.30)</td>
<td>2.00(0.30)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>3.03(0.43)</td>
<td>3.02(0.43)</td>
<td>3.02(0.43)</td>
<td>3.01(0.36)</td>
<td>3.01(0.36)</td>
<td>3.01(0.36)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>1.00(0.23)</td>
<td>1.00(0.23)</td>
<td>1.00(0.23)</td>
<td>0.99(0.21)</td>
<td>0.99(0.21)</td>
<td>0.99(0.21)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>2.04(0.24)</td>
<td>2.04(0.24)</td>
<td>2.04(0.24)</td>
<td>1.97(0.25)</td>
<td>1.97(0.25)</td>
<td>1.97(0.25)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>1.53(0.25)</td>
<td>1.53(0.25)</td>
<td>1.53(0.25)</td>
<td>1.52(0.28)</td>
<td>1.52(0.28)</td>
<td>1.52(0.28)</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>14.72(0.42)</td>
<td>14.72(0.42)</td>
<td>14.80(0.43)</td>
<td>14.65(0.39)</td>
<td>14.65(0.39)</td>
<td>14.65(0.49)</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>1.49(0.06)</td>
<td>1.49(0.06)</td>
<td>1.51(0.06)</td>
<td>1.49(0.06)</td>
<td>1.49(0.06)</td>
<td>1.50(0.10)</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>2.00(0.06)</td>
<td>2.00(0.06)</td>
<td>2.01(0.06)</td>
<td>1.98(0.06)</td>
<td>1.98(0.06)</td>
<td>2.00(0.06)</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>3.01(0.06)</td>
<td>3.01(0.06)</td>
<td>2.99(0.06)</td>
<td>3.00(0.06)</td>
<td>3.00(0.06)</td>
<td>3.00(0.08)</td>
</tr>
<tr>
<td>CT</td>
<td>32.46(1.29)</td>
<td>30.61(1.34)</td>
<td>4.54(0.11)</td>
<td>99.66(3.83)</td>
<td>95.90(5.44)</td>
<td>10.39(0.53)</td>
</tr>
</tbody>
</table>

**Table 1:** Simulation results for $n = 1000, 1500, 2000$ with 100 replications
Remark: Reduction of computational complexity from $O(n^3)$ to $O(\sum n_a^3)$, where $n_a$ is the size of the $a$th subset and $\sum n_a = n$. 
Real data example
Real data example: IBM data center thermal management

<table>
<thead>
<tr>
<th>Variable</th>
<th>$n = 1800$</th>
<th>$n = 3600$</th>
<th>$n = 26820$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>Compact</td>
<td>SSCC</td>
</tr>
<tr>
<td>$x_1$</td>
<td>$\beta_1$</td>
<td>-8.28(0.10)</td>
<td>-8.29(0.10)</td>
</tr>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>0.84(0.01)</td>
<td>0.84(0.01)</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$\beta_2$</td>
<td>-9.00(0.10)</td>
<td>-8.99(0.10)</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>0.76(0.01)</td>
<td>0.76(0.01)</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$\beta_3$</td>
<td>-6.44(0.10)</td>
<td>-6.45(0.10)</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>1.20(0.01)</td>
<td>1.20(0.01)</td>
</tr>
<tr>
<td>$x_4$</td>
<td>$\beta_4$</td>
<td>-5.42(0.11)</td>
<td>-5.41(0.11)</td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td>1.90(0.01)</td>
<td>1.90(0.01)</td>
</tr>
<tr>
<td>$x_5$</td>
<td>$\beta_5$</td>
<td>-0.08(0.13)</td>
<td>-0.07(0.13)</td>
</tr>
<tr>
<td></td>
<td>$\theta_5$</td>
<td>3.50(0.01)</td>
<td>3.50(0.01)</td>
</tr>
<tr>
<td>$x_6$</td>
<td>$\beta_6$</td>
<td>-1.98(0.10)</td>
<td>-1.97(0.10)</td>
</tr>
<tr>
<td></td>
<td>$\theta_6$</td>
<td>1.29(0.01)</td>
<td>1.29(0.01)</td>
</tr>
<tr>
<td>$x_7$</td>
<td>$\beta_7$</td>
<td>-3.39(0.06)</td>
<td>-3.41(0.06)</td>
</tr>
<tr>
<td></td>
<td>$\theta_7$</td>
<td>0.20(0.01)</td>
<td>0.20(0.01)</td>
</tr>
<tr>
<td>$x_8$</td>
<td>$\beta_8$</td>
<td>2.80(0.08)</td>
<td>2.80(0.08)</td>
</tr>
<tr>
<td></td>
<td>$\theta_8$</td>
<td>0.60(0.01)</td>
<td>0.60(0.01)</td>
</tr>
<tr>
<td>$x_9$</td>
<td>$\beta_9$</td>
<td>22.33(0.18)</td>
<td>22.35(0.18)</td>
</tr>
<tr>
<td></td>
<td>$\theta_9$</td>
<td>21.90(0.03)</td>
<td>21.90(0.03)</td>
</tr>
<tr>
<td>CT (seconds)</td>
<td>2768.70</td>
<td>2753.91</td>
<td>55.07</td>
</tr>
</tbody>
</table>

**Table 2:** Mean, standard deviation and computing time of estimations by MLE, compact and SSCC methods with subsample size $n = 1800, 3600$ and the entire CFD data
Real data example: quantification of the prediction uncertainty

**Figure 3:** Comparison of CD predictive distribution and plug-in predictive distribution when $n = 1800$
Summary
Summary

- Introduced a unified framework for GP modeling that can dramatically reduce computation and also provide a better quantification of the prediction uncertainty
  - A Sequential Split-Conquer-Combine (SSCC) procedure
  - CD Combining
  - Predictive confidence distribution
- Provide asymptotically equivalent estimation
- Reduce computational complexity from $O(n^3)$ to $O(\sum n_a^3)$, where $n_a$ is the size of the $a$th subset and $\sum n_a = n$. 
Thank You!
A General Framework of CD Combining

• Univariate CD (Singh, Xie and Strawderman (2005)) Suppose $H_1(\theta), \ldots, H_k(\theta)$ are CDs for $k$ studies,
  • A continuous and monotonic function (in each coordinate)
    \[ g_c(u_1, \ldots, u_k) : [0, 1]^k \to R \]
    Define $G_c(t) = P(G_c(U_1, \ldots, U_k) \leq t)$, where $U_1, \ldots, U_k$ are i.i.d. $U(0, 1)$ random variables.
  • Combined CD function:
    \[ H_c(\theta) = G_c\{g_c(H_1(\theta), \ldots, H_k(\theta))\} \]
    is a CD function for the parameter $\theta$.

• Multivariate CD Combining (Liu, Liu and Xie (2014) and Yang et al (2014))
Introduction to computer experiments

• If either
  (i) the components of the process of interest and their interactions are adequately understood so it can be simulated or
  (ii) the physics of the process is sufficiently well understood so it can be described by a mathematical model relating the response to the potential factors that affect the outputs,

• then the computer code/simulation can be served as a proxy for the physical process.

• Features of computer experiments:
  (i) deterministic outputs,
  (ii) traditional principles used in designing physical experiments (eg randomization, blocking, replication) are irrelevant,
  (iii) high-dimensional input and time-consuming to run.