Parallel Gaussian process regression for big data: Low-rank representation meets Markov approximation
Parallel Gaussian Process Regression for Big Data: Low-Rank Representation Meets Markov Approximation

Kian Hsiang Low* and Jiangbo Yu* and Jie Chen§ and Patrick Jaillet†
Department of Computer Science, National University of Singapore, Republic of Singapore*
Singapore-MIT Alliance for Research and Technology, Republic of Singapore§
Massachusetts Institute of Technology, USA†
{lowkh, yujiang}@comp.nus.edu.sg*, chenjie@smart.mit.edu§, jaillet@mit.edu†

Abstract
The expressive power of a Gaussian process (GP) model comes at a cost of poor scalability in the data size. To improve its scalability, this paper presents a low-rank-cum-Markov approximation (LMA) of the GP model that is novel in leveraging the dual computational advantages stemming from complementing a low-rank approximate representation of the full-rank GP based on a support set of inputs with a Markov approximation of the resulting residual process; the latter approximation is guaranteed to be closest in the Kullback-Leibler distance criterion subject to some constraint and is considerably more refined than that of existing sparse GP models utilizing low-rank representations due to its more relaxed conditional independence assumption (especially with larger data). As a result, our LMA method can trade off between the size of the support set and the order of the Markov property to (a) incur lower computational cost than such sparse GP models while achieving predictive performance comparable to them and (b) accurately represent features/patterns of any scale. Interestingly, varying the Markov order produces a spectrum of LMAs with PIC approximation and full-rank GP at the two extremes. An advantage of our LMA method is that it is amenable to parallelization on multiple machines/cores, thereby gaining greater scalability. Empirical evaluation on three real-world datasets in clusters of up to 32 computing nodes shows that our centralized and parallel LMA methods are significantly more time-efficient and scalable than state-of-the-art sparse and full-rank GP regression methods while achieving comparable predictive performances.

1 Introduction
Gaussian process (GP) models are a rich class of Bayesian non-parametric models that can perform probabilistic regression by providing Gaussian predictive distributions with formal measures of the predictive uncertainty. Unfortunately, a GP model is handicapped by its poor scalability in the size of the data, hence limiting its practical use to small data. To improve its scalability, two families of sparse GP regression methods have been proposed: (a) Low-rank approximate representations (Hensman, Fusi, and Lawrence 2013; Snelson and Ghahramani 2007; Quiñonero-Candela and Rasmussen 2005) that are well-suited for modeling slowly-varying functions with large correlation and can use all the data for predictions. But, they require a relatively high rank to capture small-scale features/patterns (i.e., of small correlation) with high fidelity, thus losing their computational advantage. (b) Recent sparse GP regression methods (Chen et al. 2013; Snæs et al. 2013) have unified approaches from the two families described above to harness their complementary modeling and predictive capabilities (hence, eliminating their deficiencies) while retaining their computational advantages. Specifically, after approximating the FGP (in particular, its covariance matrix) with a low-rank representation based on the notion of a support set of inputs, a sparse covariance matrix approximation of the resulting residual process is made. However, this sparse residual covariance matrix approximation imposes a fairly strong conditional independence assumption given the support set since the support set cannot be too large to preserve time efficiency (see Remark 2 after Proposition 1 in Section 3). In this paper, we argue that such a strong assumption is an overkill: It is in fact possible to construct a more refined, dense residual covariance matrix approximation by exploiting a Markov assumption and, perhaps surprisingly, still achieve scalability, which distinguishes our work here from existing sparse GP regression methods utilizing low-rank representations (i.e., including the unified approaches) described earlier. As a result, our proposed residual covariance matrix approximation can significantly relax the conditional independence assumption (especially with larger data; see Remark 1 after Proposition 1 in Section 3), hence potentially improving the predictive performance.

* Kian Hsiang Low and Jiangbo Yu are co-first authors.

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
This paper presents a low-rank-cum-Markov approximation (LMA) of the FGP model (Section 3) that is novel in leveraging the dual computational advantages stemming from complementing the reduced-rank covariance matrix approximation based on the support set with the residual covariance matrix approximation due to the Markov assumption; the latter approximation is guaranteed to be closest in the Kullback-Leibler distance criterion subject to some constraint. Consequently, our proposed LMA method can trade off between the size of the support set and the order of the Markov property to (a) incur lower computational cost than sparse GP regression methods utilizing low-rank representations with only the support set size (e.g., (Chen et al. 2013; Snelson and Ghahramani 2007)) or number of spectral points (Lázaro-Gredilla et al. 2010) as the varying parameter while achieving predictive performance comparable to them and (b) accurately represent features/patterns of any scale. Interestingly, varying the Markov order produces a spectrum of LMA with the partially independent conditional (PIC) approximation (Chen et al. 2013; Snelson and Ghahramani 2007) and FGP at the two extremes. An important advantage of LMA over most existing sparse GP regression methods is that it is amenable to parallelization on multiple machines/cores, thus gaining greater scalability for performing real-time predictions necessary in many time-critical applications and decision support systems (e.g., ocean sensing (Cao, Low, and Dolan 2013; Dolan et al. 2009; Low, Dolan, and Khosla 2008; 2009; 2011; Low et al. 2012; Podnar et al. 2010), traffic monitoring (Chen et al. 2012; Low, Dolan, and Khosla 2008; 2009; 2011; Low et al. 2012; Podnar et al. 2010),) and even into M disjoint subsets $D_1,\ldots,D_M$ ($\mathcal{U}_1,\ldots,\mathcal{U}_M$) such that the outputs $y_{D_m}$ and $Y_{\mathcal{U}_m}$ are as highly correlated as possible for $m = 1,\ldots,M$. Let $\mathcal{V}_m \triangleq D_m \cup \mathcal{U}_m$. Then, $\mathcal{V} = \bigcup_{m=1}^{M} \mathcal{V}_m$.

The key idea of our low-rank-cum-Markov approximation (LMA) method is to approximate the residual covariance matrix $R_{\mathcal{V}\mathcal{V}}$ by a block matrix $R_{\mathcal{V}\mathcal{V}}$ partitioned into $M \times M$ square blocks, that is, $R_{\mathcal{V}\mathcal{V}} \triangleq [R_{v_m,v_n}]_{m,n=1,\ldots,M}$ where

$$R_{v_m,v_n} \triangleq \begin{cases} R_{v_m,v_n} & \text{if } |m-n| \leq B, \\ R_{v_m,v_n} R_{v_n,v_n}^{-1} R_{v_m,v_n} & \text{if } |m-n| > B > 0, \\ 0 & \text{if } |m-n| > B = 0; \end{cases}$$

such that $B \in \{0,\ldots,M-1\}$ denotes the order of the Markov property imposed on the residual process $\{Y_x\}_{x \in \mathcal{D}}$ to be detailed later, $D_B \triangleq \bigcup_{k=m+1}^{\min(m+B,M)} D_k$, and 0 denotes a square block comprising components of value 0.

To understand the intuition underlying the approximation in (1), Fig. 1a illustrates a simple case of $R_{\mathcal{V}\mathcal{V}}$ with $B = 1$ and $M = 4$ for ease of exposition: It can be observed that only the blocks $R_{v_m,v_n}$ outside the $B$-block band of $R_{\mathcal{V}\mathcal{V}}$ (i.e., $|m-n| > B$) are approximated, specifically, by unshaded blocks $R_{v_m,v_n}$ being defined as a recursive series of $|m-n| - B$ reduced-rank residual covariance matrix approximations (1). So, when an unshaded block $R_{v_m,v_n}$ is
Figure 1: \( \overline{R}_{VV} \) and \( \overline{R}_{DD}^{-1} \) with \( B = 1 \) and \( M = 4 \). (a) Shaded blocks (i.e., \( |m - n| \leq B \)) form the \( B \)-block band while unshaded blocks (i.e., \( |m - n| > B \)) fall outside the band. Each arrow denotes a recursive call. (b) Unshaded blocks outside \( B \)-block band of \( \overline{R}_{DD}^{-1} \) (i.e., \( |m - n| > B \)) are \( \emptyset \).

Further from the diagonal of \( \overline{R}_{VV} \) (i.e., larger \( |m - n| \)), it is derived using more reduced-rank residual covariance matrix approximations. For example, \( R_{VV} \) is approximated by an unshaded block \( \overline{R}_{VV} \), being defined as a recursive series of 2 reduced-rank residual covariance matrix approximations, namely, approximating \( R_{VV} \) by \( \overline{R}_{VV} \), where \( R_{VV} = R_{11}R_{12}R_{21}^{-1}R_{22} \) is a matrix-based on the support set \( \mathcal{D}_1 = \mathcal{D}_2 \) of inputs and in turn approximating \( R_{VV} \) by a submatrix \( R_{VV}^{-1} = R_{11}R_{12}R_{21}^{-1} \) of unshaded block \( \overline{R}_{VV} \), on the support set \( \mathcal{D}_2 = \mathcal{D}_3 \) of inputs. As a result, \( R_{VV} = R_{11}R_{12}R_{21}^{-1} \), which is fully specified by five submatrices of the respective shaded blocks \( R_{VV}, R_{Vv}, R_{Vv}, R_{vV}, \) and \( R_{vV} \) within the \( B \)-block band of \( \overline{R}_{VV} \) (i.e., \( |m - n| \leq B \)). In general, any unshaded block \( \overline{R}_{VV} \), outside the \( B \)-block band of \( \overline{R}_{VV} \) (i.e., \( |m - n| > B \)) is fully specified by submatrices of the shaded blocks within the \( B \)-block band of \( \overline{R}_{VV} \) (i.e., \( |m - n| \leq B \)) due to its recursive series of \( |m - n| - B \) reduced-rank residual covariance matrix approximations (1). Though it may not be obvious now such an approximation would entail scalability, (1) interestingly offers an alternative interpretation of imposing a \( B \)-th order Markov property on residual process \( \{Y_x\}_{x \in \mathcal{D}} \), which reveals a further insight on the structural assumption of LMA to be exploited for achieving scalability, as detailed later.

The covariance matrix \( \Sigma_{VV} \) is thus approximated by a block matrix \( \Sigma_{VV} = \Sigma_{VV} + \Sigma_{VV}^{-1} \) partitioned into \( M \times M \) square blocks, that is, \( \Sigma_{VV} = \Sigma_{VV}^{-1} \), where

\[
\Sigma_{VV} = \begin{bmatrix} \Sigma_{v_n,v_n} & \cdots & \Sigma_{v_1,v_n} \\ \vdots & \ddots & \vdots \\ \Sigma_{v_1,v_1} & \cdots & \Sigma_{v_1,v_1} \end{bmatrix}_{|m,n| = 1,\ldots,M}.
\]

So, within the \( B \)-block band of \( \Sigma_{VV} \) (i.e., \( |m - n| \leq B \)), \( \Sigma_{v_n,v_n} = \Sigma_{v_1,v_1} \), by (1) and (2). Note that when \( B = 0 \), \( \Sigma_{v_n,v_n} = \Sigma_{v_1,v_1} \) for \( |m - n| > B \), thus yielding the prior covariance matrix \( \Sigma_{VV} \) of the partially independent conditional (PIC) approximation (Chen et al. 2013; Snelson and Ghahramani 2007). When \( B = M - 1 \), \( \Sigma_{VV} \) is the prior covariance matrix of FGP model. So, LMA generalizes PIC (i.e., if \( B = 0 \)) and becomes FGP if \( B = M - 1 \). Varying Markov order \( B \) from 0 to \( M - 1 \) produces a spectrum of LMAs with PIC and FGP at the two extremes.

By approximating \( \Sigma_{YY} \) with \( \Sigma_{YY}^{-1} \), our LMA method utilizes the data \( (\mathcal{D}, u_D) \) to predict the unobserved outputs for any set \( \mathcal{U} \subseteq \mathcal{X} \setminus \mathcal{D} \) of inputs and provide their corresponding predictive uncertainties using the following predictive mean and covariance matrix, respectively:

\[
\mu_{\mathcal{U}}^{\text{LMA}} = \mu_{\mathcal{U}} + \Sigma_{\mathcal{U}\mathcal{D}}^{-1} \Sigma_{\mathcal{D}\mathcal{D}}^{-1} (y_D - \mu_D)
\]

\[
\Sigma_{\mathcal{U}\mathcal{U}}^{\text{LMA}} = \Sigma_{\mathcal{U}\mathcal{U}} - \Sigma_{\mathcal{U}\mathcal{D}}^{-1} \Sigma_{\mathcal{D}\mathcal{D}}^{-1} \Sigma_{\mathcal{D}\mathcal{U}}
\]

where \( \Sigma_{\mathcal{U}\mathcal{D}}, \Sigma_{\mathcal{D}\mathcal{D}}, \) and \( \Sigma_{\mathcal{D}\mathcal{D}} \) are obtained using (2), and \( \Sigma_{\mathcal{D}\mathcal{D}} = \Sigma_{\mathcal{D}\mathcal{D}}^{-1} \Sigma_{\mathcal{D}\mathcal{D}} \). If \( \Sigma_{\mathcal{D}\mathcal{D}} \) in (3) and (4) is inverted directly, then it would still incur the same \( O(|\mathcal{D}|^3) \) time as inverting \( \Sigma_{\mathcal{D}\mathcal{D}} \) in the FGP regression method (Section 2). In the rest of this section, we will show how this scalability issue can be resolved by leveraging the computational advantages associated with both the reduced-rank covariance matrix approximation \( Q_{\mathcal{D}\mathcal{D}} \) based on the support set \( \mathcal{S} \) and our proposed residual covariance matrix approximation \( \Sigma_{\mathcal{D}\mathcal{D}} \) due to \( B \)-th order Markov assumption after decomposing \( \Sigma_{\mathcal{D}\mathcal{D}} \).

It can be observed from \( \overline{R}_{VV} \) (1) that \( \overline{R}_{DD} \) is approximated by a block matrix \( \overline{R}_{DD} = \overline{R}_{DD}, \overline{R}_{DD}, \overline{R}_{DD} \), which is a submatrix of \( \overline{R}_{VV} \), obtained using (1).

**Proposition 1** Block matrix \( \overline{R}_{DD} \) is \( B \)-block-banded, that is, any block outside its \( B \)-block band is \( \emptyset \) (e.g., Fig. 1b).

Its proof follows directly from a block-banded matrix result of Asif and Moura (2005) (specifically, Theorem 3).

**Remark 1.** In the same spirit as a Gaussian Markov random process, imposing a \( B \)-th order Markov property on residual process \( \{Y_x\}_{x \in \mathcal{D}} \) is equivalent to approximating \( R_{DD} \) by \( \overline{R}_{DD} \), whose inverse is \( B \)-block-banded (Fig. 1b). That is, if \( |m - n| > B \), \( Y_{\mathcal{D}m} \) and \( Y_{\mathcal{D}n} \) are conditionally independent given \( Y_{\mathcal{D}D(\mathcal{D}m \cup \mathcal{D}n)} \). Such a conditional independence assumption thus becomes more relaxed with larger data. More importantly, this \( B \)-th order Markov assumption or, equivalently, sparsity of \( B \)-block-banded \( \overline{R}_{DD} \) is the key to achieving scalability, as shown in the proof of Theorem 2 later.

**Remark 2.** Though \( \overline{R}_{DD} \) is sparse, \( \overline{R}_{DD} \) is a dense residual covariance matrix approximation if \( B > 0 \). In contrast, the sparse GP regression methods utilizing low-rank representations (i.e., including unified approaches) utilize a sparse residual covariance matrix approximation (Section 1), hence imposing a significantly stronger conditional independence assumption than LMA. For example, PIC (Chen et al. 2013; Snelson and Ghahramani 2007) assumes \( Y_{\mathcal{D}m} \) and \( Y_{\mathcal{D}n} \) to be conditionally independent given only \( Y_{\mathcal{D}D} \) if \( |m - n| > B \).

The next result reveals that, among all \( |\mathcal{D}| \times |\mathcal{D}| \) matrices whose inverse is \( B \)-block-banded, \( \overline{R}_{DD} \) approximates \( R_{DD} \) most closely in the Kullback-Leibler (KL) distance criterion, that is, \( \overline{R}_{DD} \) has the minimum KL distance from \( R_{DD} \):

**Theorem 1** Let KL distance \( D_{KL}(R, \overline{R}) \) \( \triangleq 0.5(\text{tr}(R \overline{R}^{-1}) - \log |R| + \log |\overline{R}| - |\mathcal{D}|) \) between two \( |\mathcal{D}| \times |\mathcal{D}| \) positive definite matrices \( R \) and \( \overline{R} \) measure the error of approximating \( R \) with \( \overline{R} \). Then, for any matrix \( \overline{R} \) whose inverse is \( B \)-block-banded, \( D_{KL}(R_{DD}, \overline{R}) \geq D_{KL}(\overline{R}_{DD}, \overline{R}_{DD}) \).

Its proof is in (Low et al. 2014c). Our main result in Theo-
rem 2 below exploits the sparsity of $R_{DD}^{-1}$ (Proposition 1) for deriving an efficient formulation of LMA, which is amenable to parallelization on multiple machines/cores by constructing and communicating the following summary information:

**Definition 1 (Local Summary)** The $m$-th local summary is defined as a tuple $(\hat{y}_m, \hat{R}_m, \hat{\Sigma}_S, \hat{\Sigma}_{UU})$ where

$$\hat{y}_m \triangleq y_{Dm} - \mu_{Dm} - R_{DmDm}^{-1}(y_{Dm} - \mu_{Dm}),$$
$$\hat{R}_m \triangleq (R_{DmDm} - R_{DmDm}^{-1} R_{DmDm})^{-1},$$
$$\hat{\Sigma}_S \triangleq \Sigma_{DS} + \sum_{m=1}^M (\hat{\Sigma}_m^{UU})^\top \hat{R}_m \hat{\Sigma}_S,$$
$$\hat{\Sigma}_{UU} \triangleq \Sigma_{UU} + \sum_{m=1}^M (\hat{\Sigma}_m^{UU})^\top \hat{R}_m \hat{\Sigma}_m^{UU}.$$  

such that $R_{DmDm}^{-1} \triangleq R_{DmDm} - R_{DmDm}^{-1} R_{DmDm}$.  

**Definition 2 (Global Summary)** The global summary is defined as a tuple $(\hat{y}_S, \hat{y}_U, \hat{\Sigma}_{SS}, \hat{\Sigma}_{UU}, \hat{\Sigma}_{UUUU})$ where

$$\hat{y}_S \triangleq \sum_{m=1}^M (\hat{\Sigma}_m^{SS})^\top \hat{R}_m \hat{y}_m; \quad \hat{y}_U \triangleq \sum_{m=1}^M (\hat{\Sigma}_m^{UU})^\top \hat{R}_m \hat{y}_m;$$
$$\hat{\Sigma}_{SS} \triangleq \Sigma_{SS} + \sum_{m=1}^M (\hat{\Sigma}_m^{SS})^\top \hat{R}_m \hat{\Sigma}_S,$$
$$\hat{\Sigma}_{UU} \triangleq \Sigma_{UU} + \sum_{m=1}^M (\hat{\Sigma}_m^{UU})^\top \hat{R}_m \hat{\Sigma}_m^{UU}.$$

**Theorem 2** For $B > 0$, $\mu_{AA}^\text{LMA}$ (3) and $\Sigma_{AA}^\text{LMA}$ (4) can be reduced to $\mu_{AA} = \mu_U + \hat{y}_S - \hat{\Sigma}_{SS}^{-1} \hat{\Sigma}_{UU} \hat{y}_S$ and $\Sigma_{AA} = \Sigma_{UU} - \hat{\Sigma}_{UUU} - \hat{\Sigma}_{SS}^{-1} \hat{\Sigma}_{UU}$.  

Its proof in (Low et al. 2014c) essentially relies on the sparsity of $R_{DD}^{-1}$ and the matrix inversion lemma.

**Remark 1.** To parallelize LMA, each machine/core $m$ constructs and uses the $m$-th local summary to compute the $m$-th summation terms in the global summary, which are then communicated to a master node. The master node constructs and communicates the global summary to the $M$ machines/cores, specifically, by sending the tuple $(\hat{y}_S, \hat{y}_U, \hat{\Sigma}_{SS}, \hat{\Sigma}_{UU}, \hat{\Sigma}_{UUUU})$ to each machine/core $m$. Finally, each machine/core $m$ uses this received tuple to predict the unobserved outputs for the set $U_m$ of inputs and provide their corresponding predictive uncertainties using $\mu_{lm}^\text{LMA}$ (3) and $\Sigma_{lm}^\text{LMA} (4)$, respectively. Computing $\hat{\Sigma}_{DD}^\top \hat{R}_m \hat{\Sigma}$ terms in the local summary can also be parallelized due to their recursive definition (i.e., (1) and (2)), as discussed in (Low et al. 2014c). This parallelization capability of LMA shows another key advantage over existing sparse GP regression methods in gaining scalability.

**Remark 2.** Supposing $M, |U|, |S| \leq |D|$, LMA can compute $\mu_{AA}^\text{LMA}$ and $\Sigma_{AA}^\text{LMA}$ distributedly in $O(|S|^3 + (B|M|)^2 + |U||(D|M) + (S + B|M)|)$ time using $M$ parallel machines/cores and sequentially in $O(|D||S|^2 + B|M|)$ time on a single centralized machine. So, our LMA method incurs cubic time in support set size $|S|$ and Markov order $B$. Increasing the number $M$ of parallel machines/cores and blocks reduces the incurred time of our parallel and centralized LMA methods, respectively. Without considering communication latency, the speedup of our parallel LMA method grows with increasing $M$ and training data size $|D|$; to explain the latter, unlike the additional $O(|D||S|^2)$ time of our centralized method that increases with more data, parallel LMA does not have a corresponding $O(|D|M)$ term.

**Remark 3.** Predictive performance of LMA is improved by increasing the support set size $|S|$ and/or Markov order $B$ at the cost of greater time overhead. From Remark 2, since LMA incurs cubic time in $|S|$ as well as in $B$, one should trade off between $|S|$ and $B$ to reduce the computational cost while achieving the desired predictive performance. In contrast, PIC (Chen et al. 2013; Snelson and Ghahramani 2007) (sparse spectrum GP (Lázaro-Gredilla et al. 2010)) can only vary support set size (number of spectral points) to obtain the desired predictive performance.

**Remark 4.** We have illustrated through a simple toy example in (Low et al. 2014c) that, unlike the local GP approaches, LMA does not exhibit any discontinuity in its predictions despite data partitioning.

### 4 Experiments and Discussion

This section first empirically evaluates the predictive performance and scalability of our proposed centralized and parallel LMA methods against that of the state-of-the-art centralized PIC (Snelson and Ghahramani 2007), parallel PIC (Chen et al. 2013), sparse spectrum GP (SSGP) (Lázaro-Gredilla et al. 2010), and FGP on two real-world datasets: (a) The SARCOS dataset (Vijayakumar, D’Souza, and Schaal 2005) of size 48933 is obtained from an inverse dynamics problem for a 7 degrees-of-freedom SARCOS robot arm. Each input is specified by a 21D feature vector of joint positions, velocities, and accelerations. The output corresponds to one of the 7 joint torques. (b) The AIMPEAK dataset (Chen et al. 2013) of size 41850 comprised traffic speeds (km/h) along 775 road segments of an urban road network during morning peak hours on April 20, 2011. Each input (i.e., road segment) denotes a 5D feature vector of length, number of lanes, speed limit, direction, and time. The time dimension comprises 54 five-minute time slots. This traffic dataset was modeled using a relational GP (Chen et al. 2012) whose correlation structure can exploit the road segment features and road network topology information. The outputs correspond to the traffic speeds.

Both datasets are modeled using GPs whose prior covariance $\sigma_{xx'}^2$ is defined by the squared exponential covariance function

$$\sigma_{xx'}^2 \triangleq \sigma_n^2 \exp(-0.5 \sum_{i=1}^d (x_i - x'_i)^2/\ell_i^2) + \sigma_d^2 \delta_{xx'},$$

due to the $i$-th component of input feature vector $x (x')$, the hyperparameters $\sigma_n^2, \sigma_d^2, \ell_1, \ldots, \ell_d$.

---

2A notable exception is the work of Chen et al. (2013) that parallelizes PIC. As mentioned earlier, our LMA generalizes PIC.

3Speedup is the incurred time of a sequential/centralized algorithm divided by that of its parallel counterpart.

4For AIMPEAK dataset, multi-dimensional scaling is used to map the input domain (i.e., of road segments) onto the Euclidean space (Chen et al. 2012) before applying the covariance function.
are, respectively, signal variance, noise variance, and length-scales, and $\delta_{x'x}$ is a Kronecker delta that is 1 if $x = x'$ and 0 otherwise. The hyperparameters are learned using randomly selected data of size 10000 via maximum likelihood estimation. Test data of size $|U| = 3000$ are randomly selected from each dataset for predictions. From remaining data, training data of varying $|D|$ are randomly selected. Support sets for LMA and PIC and the set $S$ of spectral points for SSGP are selected randomly from both datasets.

The experimental platform is a cluster of 32 computing nodes connected via gigabit links: Each node runs a Linux system with Intel® Xeon® E5620 at 2.4 GHz with 24 GB memory and 16 cores. Our parallel LMA method and parallel PIC are tested with different numbers $M = 32$, 48, and 64 of cores; all 32 computing nodes with 1, 1-2, and 2 cores each are used, respectively. For parallel LMA and parallel PIC, each computing node will be storing, respectively, a subset of the training data ($D_m \cup D_B$ and $D_m \cup D_B$) and $(D_m, y_{D_m})$ associated with its own core $m$.

Three performance metrics are used to evaluate the tested methods: (a) Root mean square error (RMSE) ($|U|^{-1/2} \sum_{x \in U}(y_x - \mu_x)^2$)$^{1/2}$, (b) incurred time, and (c) speedup. For RMSE metric, each tested method has to plug its predictive mean into $\mu_x$.

Table 1 shows results of RMSEs and incurred times of parallel LMA, parallel PIC, SSGP, and FGP averaged over 5 random instances with varying data sizes $|D|$ and cores $M$ for both datasets. The observations are as follows:

(a) Predictive performances of all tested methods improve with more data, which is expected. For SARCOS dataset, parallel LMA, parallel PIC, and SSGP achieve predictive performances comparable to that of FGP. For AIMPEAK dataset, parallel LMA does likewise and outperforms parallel PIC and SSGP with more data ($|D| \geq 24000$), which may be due to its more relaxed conditional independence assumption with larger data (Remark 1 after Proposition 1).

(b) The incurred times of all tested methods increase with more data, which is also expected. FGP scales very poorly with larger data such that it incurs $> 4$ hours for $|D| = 32000$. In contrast, parallel LMA incurs only 1-5 minutes for both datasets when $|D| = 32000$. Parallel LMA incurs much less time than parallel PIC and SSGP while achieving a comparable or better predictive performance because it requires a significantly smaller $|S|$ than parallel PIC and SSGP simply by imposing a 1-order Markov property ($B = 1$) on the residual process (Remark 3 after Theorem 2). Though $B$ is only set to 1, the dense residual covariance matrix approximation provided by LMA (as opposed to sparse approximation of PIC) is good enough to achieve its predictive performances reported in Table 1. From Table 1b, when training data is small ($|D| = 8000$) for AIMPEAK dataset, parallel PIC incurs more time than FGP due to its huge $|S| = 5120$, which causes communication latency to dominate the incurred time (Chen et al. 2013). When $|D| \leq 24000$, SSGP also incurs more time than FGP due to its large $|S| = 4096$.

(c) Predictive performances of parallel LMA and PIC generally remain stable with more cores, thus justifying the practicality of their structural assumptions to gain time efficiency.

Table 2 shows results of speedups of parallel LMA and parallel PIC as well as incurred times of their centralized counterparts averaged over 5 random instances with varying data sizes $|D|$ and numbers $M$ of cores for AIMPEAK dataset. The observations are as follows:

(a) The incurred times of centralized LMA and centralized PIC increase with more data, which is expected. When $|D| \geq 32000$, centralized LMA incurs only 16-30 minutes (as compared to FGP incurring $> 4$ hours) while centralized PIC and SSGP incur, respectively, more than 3.5 and 2 hours due to their huge $|S|$. In fact, Table 2 shows that centralized PIC incurs even more time than FGP for almost all possible settings of $|D|$ and $M$ due to its huge support set.

(b) The speedups of parallel LMA and parallel PIC generally increase with more data, as explained in Remark 2 after Theorem 2, except for that of parallel LMA being slightly higher than expected when $|D| = 16000$.

(c) The incurred time of centralized LMA decreases with more blocks (i.e., larger $M$), as explained in Remark 2 after Theorem 2. This is also expected of centralized PIC, but its incurred time increases with more blocks instead due to its huge support set, which entails large-scale matrix operations causing a huge number of cache misses. This highlights the need to use a sufficiently small support set on a single centralized machine so that cache misses will contribute less to incurred time, as compared to data processing.

(d) The speedup of parallel LMA increases with more cores, as explained in Remark 2 after Theorem 2. Though the speedup of parallel PIC appears to increase considerably

<table>
<thead>
<tr>
<th>$M$</th>
<th>8000</th>
<th>16000</th>
<th>24000</th>
<th>32000</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMA</td>
<td>2.4(0.056)</td>
<td>2.4(0.077)</td>
<td>2.1(0.572)</td>
<td>2.0(0.251)</td>
</tr>
<tr>
<td>PIC</td>
<td>2.4(0.254)</td>
<td>2.2(0.284)</td>
<td>2.1(0.321)</td>
<td>2.0(0.663)</td>
</tr>
<tr>
<td>SSGP</td>
<td>2.4(0.202)</td>
<td>2.2(0.176)</td>
<td>2.1(0.106)</td>
<td>2.0(0.312)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M$</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMA</td>
<td>8.9(0.217)</td>
</tr>
<tr>
<td>PIC</td>
<td>8.1(0.209)</td>
</tr>
<tr>
<td>SSGP</td>
<td>8.1(0.189)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M$</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMA</td>
<td>8.8(0.178)</td>
</tr>
<tr>
<td>PIC</td>
<td>8.1(0.145)</td>
</tr>
<tr>
<td>SSGP</td>
<td>8.1(0.152)</td>
</tr>
</tbody>
</table>

Table 1: RMSEs and incurred times (seconds) reported in brackets of parallel LMA, parallel PIC, SSGP, and FGP with varying data sizes $|D|$ and numbers $M$ of cores for (a) SARCOS and (b) AIMPEAK datasets.

---

5Varying the set $S$ of spectral points over 50 random instances hardly changes the predictive performance of SSGP in our experiments because a very large set of spectral points ($|S| = 4096$) is used in order to achieve predictive performance as close as possible to FGP and our LMA method (see Table 1).

6A cache miss causes the processor to access the data from main memory, which costs $10 \times$ more time than a cache memory access.
### Table 2: Speedups of parallel LMA \((B = 1, |S| = 1024)\) and parallel PIC \((|S| = 5120)\) and incurred times \((\text{seconds})\) reported in brackets of their centralized counterparts with varying data sizes \(|D|\) and cores \(M\) for AIMPEAK dataset.

| \(|D|\) | 8000 | 16000 | 24000 | 32000 |
|--------|------|-------|-------|-------|
| FGP    | \(-271\) | \(-137\) | \(-323\) | \(-186\) |
| SSGP   | \(-3029\) | \(-3798\) | \(-5552\) | \(-7309\) |
| LMA    | \(6.9(139)\) | \(9.4(414)\) | \(8.0(894)\) | \(8.2(1764)\) |
| PIC    | \(19.4(3412)\) | \(18.8(10105)\) | \(19.3(11584)\) | \(21.6(12954)\) |
| \(M = 32\) | | | | |

### Figure 2: RMSEs and incurred times \((\text{seconds})\) of parallel LMA with varying support set sizes \(|S|\) and Markov orders \(B\), \(|D| = 8000\), and \(M = 32\) obtained using 8 computing nodes \((\text{each using 4 cores})\) for AIMPEAK dataset. Darker gray implies longer incurred time \((\text{larger RMSE})\) for the left \((\text{right})\) plot.

with more cores, it is primarily due to the substantial number of cache misses \((\text{see observation c above})\) that inflates the incurred time of centralized PIC excessively.

Fig. 2 shows results of RMSEs and incurred times of parallel LMA averaged over 5 random instances with varying support set sizes \(|S|\) and Markov orders \(B\), \(|D| = 8000\), and \(M = 32\) obtained using 8 computing nodes \((\text{each using 4 cores})\) for AIMPEAK dataset. Observations are as follows:

(a) To achieve RMSEs of 8.1 and 8.0 with least incurred times, one should trade off a larger support set size \(|S|\) for a larger Markov order \(B\) \((\text{or vice versa})\) to arrive at the respective settings of \(|S| = 1024, B = 5\) \((34\text{ seconds})\) and \(|S| = 1024, B = 9\) \((68\text{ seconds})\), which agrees with Remark 3 after Theorem 2. However, to achieve the same RMSE of 7.9 as FGP, the setting of \(|S| = 128, B = 21\) incurs the least time \((\text{i.e.}, 205\text{ seconds})\), which seems to indicate that, with small data \((|D| = 8000)\), we should instead focus on increasing Markov order \(B\) for LMA to achieve the same predictive performance as FGP; recall that when \(B = M - 1\), LMA becomes FGP. This provides an empirically cheaper and more reliable alternative to increasing \(|S|\) for achieving predictive performance comparable to FGP; the latter of which, in our experiments, causes Cholesky factorization failure easily when \(|S|\) becomes excessively large.

(b) When \(|S| = 1024, B = 1, \text{and} \ M = 32\), parallel LMA using 8 computing nodes incurs less time \((\text{i.e.}, 10\text{ seconds})\) than that using 32 nodes \((\text{i.e.}, 20\text{ seconds})\); see Table 1b because the communication latency between cores within a machine is significantly less than that between machines.

Next, the predictive performance and scalability of our parallel LMA method are empirically compared with that of parallel PIC using the large EMULATE mean sea level pres-

### Table 3: RMSEs and incurred times \((\text{seconds})\) reported in brackets of parallel LMA \((B = 1, |S| = 512)\) and parallel PIC \((|S| = 3400)\) with \(M = 512\) cores and varying data sizes \(|D|\) for EMSLP dataset.

| \(|D|\) | 128000 | 256000 | 384000 | 512000 | 1000000 |
|--------|--------|--------|--------|--------|---------|
| LMA    | \(823(150)\) | \(774(814)\) | \(728(3125)\) | \(682(1154)\) | \(506(78384)\) |
| PIC    | \(836(948)\) | \(-\) | \(-\) | \(-\) | \(-\) |

**Summary of Experimental Results.** LMA is significantly more scalable than FGP in the data size while achieving a comparable predictive performance for SARCOS and AIMPEAK datasets. For example, when \(|D| = 32000\) and \(M \geq 48\), our centralized and parallel LMA methods are, respectively, at least 1 and 2 orders of magnitude faster than FGP while achieving comparable predictive performances for AIMPEAK dataset. Our centralized (parallel) LMA method also incurs much less time than centralized PIC (parallel PIC) and SSGP while achieving comparable or better predictive performance because LMA requires a considerably smaller support set size \(|S|\) than PIC and SSGP simply by setting Markov order \(B = 1\), as explained earlier. Trading off between support set size and Markov order of LMA results in less incurred time while achieving the desired predictive performance. LMA gives a more reliable alternative of increasing the Markov order \((\text{i.e.}, \text{to increasing support set size})\) for achieving predictive performance similar to FGP; in practice, a huge support set causes Cholesky factorization failure and insufficient shared memory between cores easily. Finally, parallel LMA can scale up to work for EMSLP dataset of more than a million in size.

### 5 Conclusion

This paper describes a LMA method that leverages the dual computational advantages stemming from complementing the low-rank covariance matrix approximation based on support set with the dense residual covariance matrix approximation due to Markov assumption. As a result, LMA can make a more relaxed conditional independence assumption \((\text{especially with larger data})\) than many existing sparse GP regression methods utilizing low-rank representations,
the latter of which utilize a sparse residual covariance matrix approximation. Empirical results have shown that our centralized (parallel) LMA method is much more scalable than FGP and time-efficient than centralized PIC (parallel PIC) and SSGP while achieving comparable predictive performance. In our future work, we plan to develop a technique to automatically determine the "optimal" support set size and Markov order and devise an "anytime" variant of LMA using stochastic variational inference like (Hensman, Fusi, and Lawrence 2013) so that it can train with a small subset of data in each iteration instead of learning using all the data. We also plan to release the source code at http://code.google.com/p/pgpr/.

Acknowledgments. This work was supported by Singapore-MIT Alliance for Research and Technology Subaward No. 52 R-252-000-550-592.

References


