Supplementary materials to Fast Nonseparable Gaussian Stochastic Process with Application to Methylation Level Interpolation

Mengyang Gu and Yanxun Xu

All the formulas in this supplementary materials are cross-referenced in the main body of the article.

1 Closed form quantities of the continuous state space model

We give the quantities of continuous state space model representation in (11) in the main body of the article.

For $1 \leq i \leq K$, the SDE is

$$\frac{d\boldsymbol{\theta}_i(s)}{ds} = \mathbf{J}_i\boldsymbol{\theta}_i(s) + \mathbf{L}z(s),$$

where

$$\mathbf{J}_i = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\lambda_i^3 & -\lambda_i^2 & -3\lambda_i \end{pmatrix},$$

and $\mathbf{L} = (0, 0, 1)^T$.

Denote $d_j = |s_{j+1} - s_j|$ for $j = 1, ..., N-1$. We have the following expressions for the solution of the SDE in (11) in the main body of the article,

$$\mathbf{G}_i(s_j) = e^{\mathbf{J}_i d_j} = \frac{e^{-\lambda_i d_j}}{2} \begin{pmatrix} \lambda_i^2 d_j^2 + 2\lambda_i + 2 & 2(\lambda_i d_j^2 + d_j) & d_j^2 \\ -\lambda_i^3 d_j^2 & -2(\lambda_i^2 d_j^2 - \lambda_i d_j - 1) & 2d_j - \lambda_i d_j^2 \\ \lambda_i^4 d_j^2 - 2\lambda_i^3 d_j & 2(\lambda_i^3 d_j^2 - 3\lambda_i^2 d_j) & \lambda_i^2 d_j^2 - 4\lambda_i d_j + 2 \end{pmatrix}$$

$$\mathbf{W}_i(s_j) = \frac{4\sigma_i^2 \lambda_i^5}{3} \begin{pmatrix} W_{1,i}(s_j) & W_{2,i}(s_j) & W_{3,i}(s_j) \\ W_{4,i}(s_j) & W_{5,i}(s_j) & W_{6,i}(s_j) \\ W_{7,i}(s_j) & W_{8,i}(s_j) & W_{9,i}(s_j) \end{pmatrix},$$

with

$$\begin{aligned}
W_{1,i}(s_j) &= \frac{e^{-2\lambda_i d_j}(3 + 6\lambda_i d_j + 6\lambda_i^2 d_j^2 + 4\lambda_i^3 d_j^3 + 2\lambda_i^4 d_j^4) - 3}{-4\lambda_i^5}, \\
W_{2,i}(s_j) &= W_{4,i}(s_j) = \frac{e^{-2\lambda_i d_j} t^4}{2}, \\
W_{3,i}(s_j) &= W_{7,i}(s_j) = \frac{e^{-2\lambda_i d_j}(1 + 2\lambda_i d_j + 2\lambda_i^2 d_j^2 + 4\lambda_i^3 d_j^3 - 2\lambda_i^4 d_j^4) - 1}{4\lambda_i^3}, \\
W_{5,i}(s_j) &= \frac{e^{-2\lambda_i d_j}(1 + 2\lambda_i d_j + 2\lambda_i^2 d_j^2 - 4\lambda_i^3 d_j^3 + 2\lambda_i^4 d_j^4) - 1}{-4\lambda_i^3}, \\
W_{6,i}(s_j) &= W_{8,i}(s_j) = \frac{e^{-2\lambda_i d_j} d_j^2 (4 - 4\lambda_i d_j + \lambda_i^2 d_j^2)}{2}, \\
W_{9,i}(s_j) &= \frac{e^{-2\lambda_i d_j}(-3 + 10\lambda_i^2 d_j^2 - 22\lambda_i^2 d_j^2 + 12\lambda_i^2 d_j^2 - 2\lambda_i^4 d_j^4) + 3}{4\lambda_i},
\end{aligned}$$

and

$$\mathbf{W}_i(s_0) = \begin{pmatrix} \sigma_i^2 & 1 & -\sigma_i^2 \lambda_i^2/3 \\ 0 & \sigma_i^2 \lambda_i^2/3 & 1 \\ -\sigma_i^2 \lambda_i^2/3 & 0 & \sigma_i^2 \lambda_i^4 \end{pmatrix},$$

for $j = 1, ..., N$ and $i = 1, ..., K$.

For $j = 1, ..., N$, the joint distribution of $\boldsymbol{\theta}_i(s_{0:n})$ is given below

$$\begin{pmatrix} \boldsymbol{\theta}_i(s_0) \\ \boldsymbol{\theta}_i(s_1) \\ \boldsymbol{\theta}_i(s_2) \\ ... \\ \boldsymbol{\theta}_i(s_N) \end{pmatrix} \sim \text{MN} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ ... \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{W}_i(s_0)^{-1} & -\mathbf{G}_i^T(s_1)\mathbf{W}_i(s_1)^{-1}\mathbf{G}_i(s_1) & & & \\ -\mathbf{G}_i^T(s_1)\mathbf{W}_i(s_1)^{-1}\mathbf{G}_i(s_1) & \mathbf{W}_i^{-1}(s_1) & & & \\ & -\mathbf{G}_i^T(s_2)\mathbf{W}_i(s_2)^{-1}\mathbf{G}_i(s_2) & ... & & \\ & ... & ... & ... & \\ & & ... & \mathbf{W}_i^{-1}(s_N) & \end{pmatrix}^{-1} \right).$$

2 Combing feature data into the nonseparable model

To impute the methylation levels, some site-specific features such as genomic position, DNA sequence properties, cis-regulatory element, can be used as covariates in a regression model. Incorporating regressors/covariates is less studied in the nonseparable GaSP model. In this section, we discuss a way to jointly model the site-specific features and output.

Let $\mathbf{X}(s)_{[q\times 1]}$ be features at site s (including the intercept). Consider an extended model

$$\mathbf{Y}^e(s) = \mathbf{A}^e\tilde{\mathbf{v}}(s) + \boldsymbol{\epsilon}_0, \tag{S1}$$

for every $s \in \mathscr{S}$, where $\mathbf{Y}^e(s) = (\mathbf{X}^T(s); \mathbf{Y}^T(s))^T$, the weight $\tilde{\mathbf{v}}(\cdot)$ is defined the same as in (2) in the main body of the article, and $\boldsymbol{\epsilon}_0 \sim \text{MN}(0, \sigma_0^2\mathbf{I}_{K+q})$. Let $\mathbf{X}(\mathbf{s}^{\mathscr{D}})_{[q\times n]}$ be the features at sites $\mathbf{s}^{\mathscr{D}}$ and $\mathbf{Y}^e(\mathbf{s}^{\mathscr{D}}) = (\mathbf{X}^T(\mathbf{s}^{\mathscr{D}}); \mathbf{Y}^T(\mathbf{s}^{\mathscr{D}}))^T$. The extended basis matrix $\mathbf{A}^e = \mathbf{U}^e\mathbf{D}^e/\sqrt{n}$, where $\mathbf{U}^e$ and $\mathbf{D}^e$ are still defined through the SVD decomposition $\mathbf{Y}^e(\mathbf{s}^{\mathscr{D}}) = \mathbf{U}^e\mathbf{D}^e\mathbf{V}^e$. The predictive distribution of the model (S1) can be obtained similarly to Lemma 2.

The connection among regression model, the separable GaSP model, and the nonseparable GaSP model shown in the previous section still holds. Let $\mathbf{Z}^e$ follow a matrix normal distribution with mean zero and covariance $\boldsymbol{\Sigma}^e \otimes \boldsymbol{\Lambda}$, where $\boldsymbol{\Sigma}^e = \begin{pmatrix} \boldsymbol{\Sigma}^e_{00} & \boldsymbol{\Sigma}^e_{0*} \\ \boldsymbol{\Sigma}^e_{*0} & \boldsymbol{\Sigma}^e_{**} \end{pmatrix}$ is a $(q+K)\times(q+K)$ covariance matrix. The separable model is a special case of the nonseparable GaSP model (S1) specified in Remark 2. The prediction by the regression model (19) with

25% held-out CpG sites	RMSE	$P_{CI}(95\%)$	$L_{CI}(95\%)$	Accuracy
Nonseparable GaSP full model	0.083	0.956	0.278	0.969
Nonseparable GaSP by batch	0.091	0.923	0.258	0.966
75% held-out CpG sites	RMSE	$P_{CI}(95\%)$	$L_{CI}(95\%)$	Accuracy
Nonseparable GaSP full model	0.087	0.963	0.305	0.968
Nonseparable GaSP by batch	0.096	0.934	0.283	0.966

Table S1: Comparison of different methods in terms of out of sample prediction for WGBS data with 25% and 75% CpG sites are held out for testing.

covariates $\mathbf{H}_i(s_j) = (\mathbf{X}^T(s_j); \mathbf{y}^T(s_j))^T$ is also a special case of the extended nonseparable GaSP model specified in Remark 1.

3 Comparison to approximation method by blocks

In this subsection, we compare our exact and fast computation of the nonseparable GaSP model with a straightforward approximation, in which the long sequence is divided into small blocks and GaSP models are built independently in each block. Assume the data are divided into M blocks, each with n_0 inputs (where $n = Mn_0$), the computational operations of which are then $O(Mn_0^3)$ instead of $O(n^3)$ for the inversion of the covariance matrix.

For illustration purposes, the data are divided into 100 batches and 200 CpG sites are used as the training data in each batch. We consider two scenarios, with 600 CpG sites and 66 CpG sites between these 200 CpG sites being selected as the test CpG sites in each batch, respectively. That means that roughly 75% and 25% of the data are held out. We still assume the methylation levels for the first 20 samples are available at all CpG sites and 4 randomly selected samples are only partially observed. The total number of the test CpG sites is $240,000$ and $26,400$, respectively.

As shown in Table S1, the prediction by the nonseparable GaSP with the full model is about 10% better in terms of RMSE in both scenarios. This is further justified by Figure S1. One possible reason is that the boundary effect is large in the approximation method when

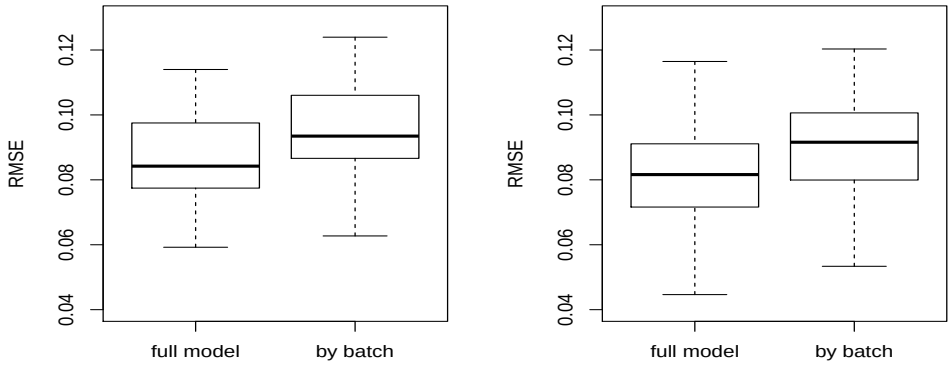

Figure S1: Boxplots of RMSE_j for the test samples in batch j with 75% of CpG sites (left panel) and 25% of CpG sites (right panel) being held-out, respectively, $j = 1, \ldots, 100$.

we divide the data into batches. All these results suggest that simply approximating the likelihood by batch yields inferior predictive results than the nonseparable GaSP model with the full likelihood. Again, the implementation of the full model relies on the FFBS algorithm discussed in Section 3 in the main body of the article.