Bayesian Gaussian Process Latent Variable Model for Pseudotime Inference of Cancer Progression

Melissa Slaughter and Larry Zhang
December 11, 2016

1. Background

Cancer is a worldwide problem with deadly consequences. Approximately 10 million people are diagnosed with cancer yearly, of which 6 million people die [1]. In many cancer-related cases, the disease is poorly prognosed, and so treatment cannot occur until the cancer is already at a lethal stage. Unfortunately, current cancer stage classifications are coarsely differentiated and thus do not accurately reflect the continuous and variable progression of cancer. Overall, this makes treatment selection very difficult. Thus, it is important to accurately determine a finer level of cancer progression, and even to predict treatment responses.

One possible way to artificially measure cancer progression is the idea of pseudotime. Pseudotime ordering assigns a score between 0 and 1 to each observation, where values near 0 indicate the start of a biological process, and values near 1 indicate the end of the process. For example, pseudotime ordering has been used fairly extensively in determining a cell's progression through differentiation or apoptosis [2].

Current attempts at pseudotime ordering algorithms take a variety of different approaches: *Monocle* [2] and *TSCAN* [3] use Independent Component Analysis with Minimum Spanning Trees; *Wanderlust* [4] uses *k*-nearest-neighbors graphs; *embeddr* [5] uses Laplacian Eigenmaps and non-parametric curve fitting. While all of these methods have been shown to provide pseudotime estimates, they cannot evaluate the uncertainty of their estimates. One proposed method for determining pseudotime estimates with their corresponding uncertainty is by using a Gaussian Process Latent Variable Model (GPLVM) [6] to perform Bayesian inference on the pseudotimes [7].

Thus, we propose developing a GPLVM, combined with a repulsive prior [8], that utilizes Bayesian inference to predict posterior distributions of cancer progression pseudotimes. For computational reasons, we will only be looking at the RNA-seq data of a single cohort of cancers: skin cutaneous melanoma. In the future, we would like to expand our model to extend beyond just RNA-seq analysis, incorporating features from clinical, miR-seq, and mutation profile data. A Python implementation of our algorithm can be found at https://github.com/lzhang124/gplvm_pseudotime.

2. Methods

2.1 Gaussian Process Latent Variable Models

A Gaussian process (GP) is a distribution of sample functions μ_t, $t \in T$, for which all finite dimensional marginal distributions are multivariate Gaussian distributions. That is, for any distinct values $t_1,...,t_n$, we have

$$(\mu_{t_1}, \ldots, \mu_{t_n}) \sim \mathcal{N}(m, K)$$

for mean function m and covariance function K.

GPLVMs are models that define a distribution over latent functions μ that relate a latent variable t to an output variable x, such as

$$x = \mu(t) + \epsilon, \epsilon \in \mathcal{N}(0, \sigma^2)$$

If the prior distribution on μ is modeled by a Gaussian process, then the N observations

$$\mathbf{x} = \{x_{t_1}, \ldots, x_{t_N}\}$$

and corresponding latent points

$$\mathbf{t} = \{t_1, \ldots, t_N\}$$

give rise to the likelihood

$$P(\mathbf{x}|\mathbf{t}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, K(\mathbf{t}, \mathbf{t}) + \sigma^2 I_N)$$

where $K(\mathbf{t}, \mathbf{t})$ is the covariance function over $\mathbf{t}$ and I_N is the $N \times N$ identity matrix [9]. Specifications of the covariance matrix K allow us to control certain features of our latent functions, such as their characteristic lengthscale (see section 2.2).

2.2 GPLVM Specifications

Our data consists of N observed samples $\mathbf{X} = \{x_1,...,x_N\}$, each with P features. Each of the N samples are assumed to be conditionally independent. For each of the N samples we define a latent, unobserved pseudotime $t_i \in [0,1]$. For each of the P features we have an observed variance over the samples, which we represent as the diagonals of a $P \times P$ observed covariance matrix $\mathbf{\Sigma}$. We model the priors of each of the P variances as an inverse gamma distribution with shape α and rate β.

The sample function $\boldsymbol{\mu}(t)$ used to relate latent pseudotimes $\mathbf{t}$ with observed samples $\mathbf{X}$ has P dimensions, one for each feature. Each dimension, $\mu_j(t)$, is given an independent Gaussian process prior with mean $\mathbf{0}$ and covariance function K_j.

We define our covariance function K_j using the squared exponential covariance function, which is the most widely-used kernel function [10]:

$$K_j(\lambda_j, t, t') = \exp\left(-\frac{(t-t')^2}{2\lambda_j^2}\right)$$

Here, λ_j represents the characteristic lengthscale of our sample function μ_j, which describes how "smooth" μ_j is. Small λs represent function values that can change quickly, while large λs represent functions that change slowly. Alternatively, it represents "how close" t and t' have to be in order to influence each other. As the values of λ_j vary across the P features, we can effectively regularize λ by modeling its prior as an exponential distribution with scale γ.

We model the prior of pseudotimes using the Coulomb repulsive process (Corp) [8], which models repulsion between points using a process inspired by electrostatic potentials. In our model, the Corp prior favors pseudotime orderings that are evenly distributed over the interval [0,1]. For pseudotimes $t_1,...,t_N$, the Corp prior is defined as:

$$P(t_1, \ldots, t_N) \propto \prod_{i=1}^{N} \prod_{j=i+1}^{N} \sin^{2r}\left(\pi |t_i - t_j|\right)$$

where r is the repulsion parameter. With this prior, as $|t_i - t_j|$ gets smaller, the probability decreases to zero (i.e. two pseudotimes will never coincide). [8] demonstrates that using the Corp prior in GPLVMs gives superior results compared to traditionally used uniform or normal priors.

Thus, our model can be summarized as the following:

$$\begin{aligned}
\mathbf{x}_i &\sim \mathcal{N}\left(\boldsymbol{\mu}(t_i), \boldsymbol{\Sigma}\right), \ i = 1, \ldots, N \\
\mu_j &\sim \mathrm{GP}(0, K_j), \ j = 1, \ldots, P \\
K_j(\lambda_j, t, t') &= \exp\left(-\frac{(t-t')^2}{2\lambda_j^2}\right), \ j = 1, \ldots, P \\
\boldsymbol{\Sigma} &= \mathrm{diag}(\sigma_1^2, \ldots, \sigma_P^2) \\
\lambda_j &\sim \mathrm{Exp}(\gamma), \ j = 1, \ldots, P \\
\sigma_j^2 &\sim \mathrm{InvGamma}(\alpha, \beta) \\
\mathbf{t} &\sim \mathrm{Corp}(r)
\end{aligned}$$

We assume that the likelihood of **X** given latent pseudotimes **t** is conditionally independent across its *P* features:

$$P(\mathbf{X}|\mathbf{t},\boldsymbol{\lambda},\boldsymbol{\sigma}^2) = \prod_{j=1}^{P} P(\mathbf{x}_j|\mathbf{t},\lambda_j,\sigma_j^2)$$

As mentioned in section 2.1, the use of GPLVMs gives rise to the following:

$$P(\mathbf{x}_j|\mathbf{t},\lambda_j,\sigma_j^2) = \mathcal{N}\left(\mathbf{x}_j|\mathbf{0},K_j(\lambda_j,\mathbf{t}) + \sigma_j^2 I_N\right)$$

Thus, the posterior likelihood can be written as

$$\begin{aligned} P(\mathbf{t},\boldsymbol{\lambda},\boldsymbol{\sigma}^2|\mathbf{X}) \propto & \prod_{j=1}^{P} \mathcal{N}(\mathbf{x}_j|\mathbf{0},K_j(\lambda_j,\mathbf{t}) + \sigma_j^2 I_N) \\ & \times \prod_{i=1}^{N}\prod_{j=i+1}^{N} \sin^{2r}\left(\pi|t_i - t_j|\right) P(\boldsymbol{\lambda})P(\boldsymbol{\sigma}^2) \end{aligned}$$

2.3 Metropolis-Hastings

Traditionally, methods for sampling the posterior distribution of GPLVMs involve using maximum *a posteriori* (MAP) estimates [8]. However, we adopted a Markov Chain Monte Carlo approach [11] since this allows us to determine posterior uncertainty in our pseudotime orderings.

Specifically, we implemented a Metropolis-Hastings random walk to sample our posterior distribution. In each iteration of the algorithm, we propose new values for the parameters **t**, **λ**, and **σ**^{2} of our model using a truncated Gaussian distribution:

$$\begin{aligned} \mathbf{t}_{i+1} &\sim \mathcal{N}_{[0,1]}\left(\mathbf{t}_i,\sigma_t^2 I_N\right) \\ \boldsymbol{\lambda}_{i+1} &\sim \mathcal{N}_{[0,\infty)}\left(\boldsymbol{\lambda}_i,\sigma_\lambda^2 I_N\right) \\ \boldsymbol{\sigma}_{i+1}^2 &\sim \mathcal{N}_{[0,\infty)}\left(\boldsymbol{\sigma}_i^2,\sigma_\sigma^2 I_N\right) \end{aligned}$$

With our proposed model, the posterior distribution is extremely multi-modal. In the Corp prior alone, there are many local maxima since the prior probability is zero whenever $t_i = t_j$. To avoid these local maxima, we can utilize the repulsive nature of the Corp prior, since it favors pseudotimes that are evenly distributed over [0,1]. Thus, it is reasonable to initialize our pseudotimes as

$$\mathbf{t}_0 \sim U\left(\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\right)$$

for arbitrarily small ϵ.

3. Results

We applied our algorithm to two datasets: (1) a synthetic dataset generated by our GPLVM and (2) RNA-seq data from the TCGA skin cutaneous melanoma dataset.

3.1 Synthetic Data

We generated our synthetic dataset using the model described in section 2.2. We randomly chose *N* = 50 'true' pseudotimes from U(0,1) to create *N* sample data points, each with *P* = 2 features. Then, we performed the Metropolis-Hastings algorithm to learn the 'true' pseudotimes of the synthetic data. We performed 100,000 iterations with a burn-in period of 50,000. The parameters of our GPLVM were as follows:

$$\epsilon = 1 \times 10^{-3} \qquad \sigma_t^2 = 5 \times 10^{-4} \qquad r = 1$$

$$\lambda = \frac{1}{\sqrt{2}} \qquad \sigma_\lambda^2 = 5 \times 10^{-6} \qquad \gamma = 1$$

$$\sigma^2 = 1 \times 10^{-3} \qquad \sigma_\sigma^2 = 5 \times 10^{-11} \qquad \alpha, \beta = 1$$

The resulting pseudotime inference of the synthetic data with these parameters can be found in Figure 1. We clearly see that even though we initialized our pseudotimes at around 0.5, the repulsive Corp prior causes them to be evenly spaced on the interval [0,1] after around 20,000 iterations. Furthermore, the results show that the model converges on pseudotime estimates that are similar to their 'true' pseudotimes. We can see that the regression of MAP pseudotimes generated by our model and the 'true' pseudotime values are highly correlated.

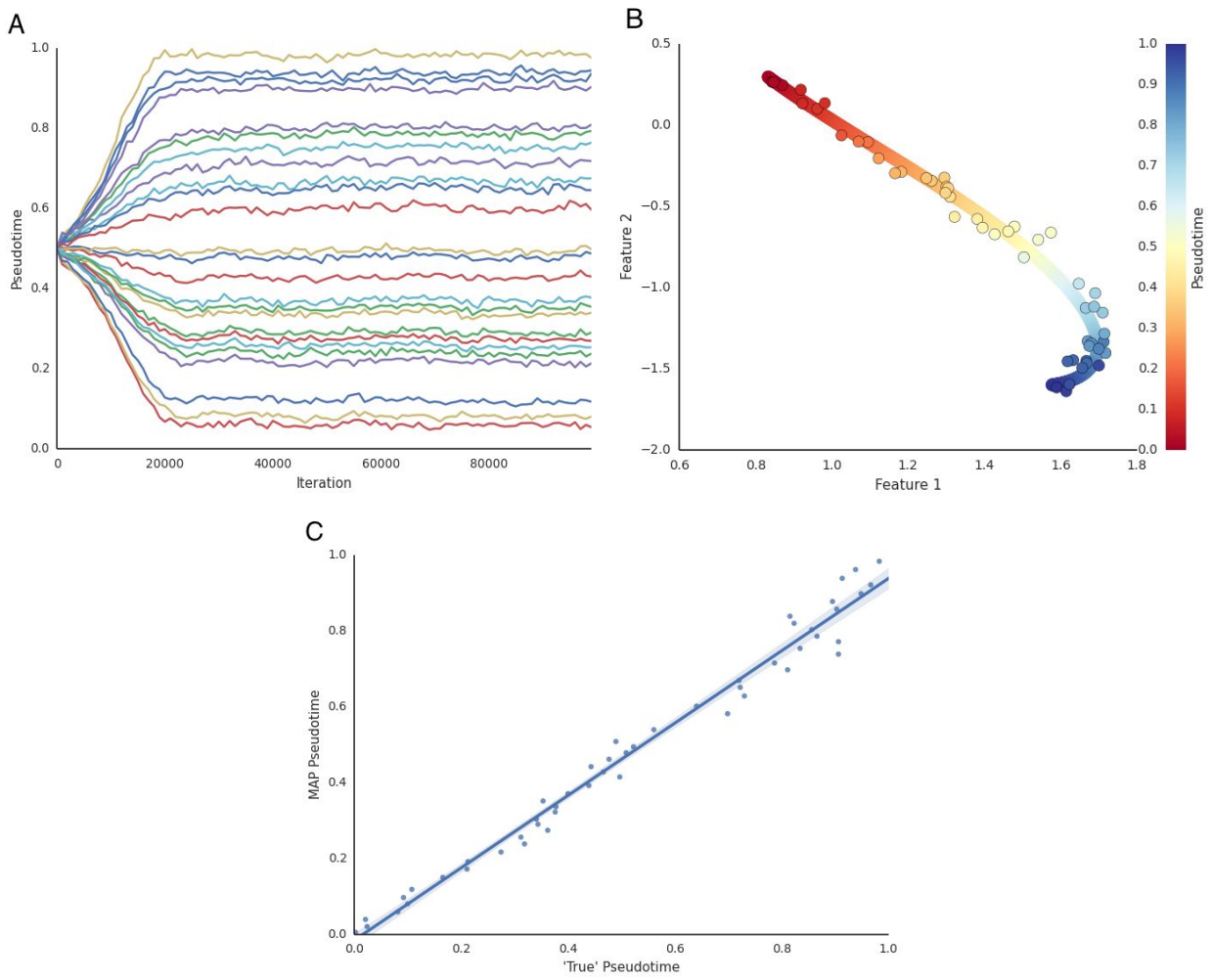

Figure 1: **Pseudotime inference of synthetic data. A.** Thinned traces for 25 randomly chosen samples and their pseudotimes over the course of the Metropolis-Hastings algorithm. **B.** The synthetic data points along with the posterior estimate of the sample function $\boldsymbol{\mu}(t)$. Points are colored with their 'true' pseudotimes while the curve is colored with MAP pseudotimes. **C.** Linear regression of MAP pseudotimes and 'true' pseudotimes. The shaded region represents the 95% confidence interval of the regression.

3.2 Skin Cutaneous Melanoma Data

Next, we applied our algorithm to mRNA-seq data from the TCGA skin cutaneous melanoma dataset. We chose to specifically focus on the 959 representative genes discovered in [12]. Unfortunately, GPLVM computations are $O(n^3)$ with respect to the dimensions of our data. Thus, due to computational limitations, we first performed PCA on these genes, and then used the first $P = 2$ principal components for our analysis.

We also center-scaled the gene expression values to have mean 0 and standard deviation 1.

In order to assess the validity of our pseudotime estimates, we need to compare them to a clinical feature that can function as a cancer progression indicator. For melanomas, one such feature is Breslow depth, which represents how deeply tumor cells have invaded the skin: larger Breslow depth values roughly correspond with later cancer progression levels. We thus used the *N* = 358 samples from the TCGA database with Breslow depth data.

We again performed 100,000 iterations with a burn-in period of 50,000. The parameters of our GPLVM were as follows:

$$\epsilon = 1 \times 10^{-3} \qquad \sigma_t^2 = 5 \times 10^{-4} \qquad r = 1$$
$$\lambda = \frac{1}{\sqrt{50}} \qquad \sigma_\lambda^2 = 5 \times 10^{-6} \qquad \gamma = 1$$
$$\sigma_\sigma^2 = 5 \times 10^{-11} \qquad \alpha, \beta = 1$$

The resulting pseudotime inference of the melanoma data with these parameters can be found in Figure 2. Again, we see that the repulsive Corp prior causes the pseudotimes to be evenly spaced on the interval [0,1] after around 40,000 iterations. However, unlike the synthetic data, our model converges on pseudotime estimates that do not seem to accurately reflect Breslow depth. We can see that the regression of MAP pseudotimes does not significantly correlate with either Breslow depth or the first principal component of the data.

This lack of correlation is likely due to the dimensional reduction we imposed on the data prior to running our algorithm. By only using the first two principal components in our analysis, we are dramatically reducing the amount of variation in the data that can be effectively modeled with our GPLVM. Furthermore, we make the assumption that the likelihood of our data is conditionally independent across principal components, which may not be true.

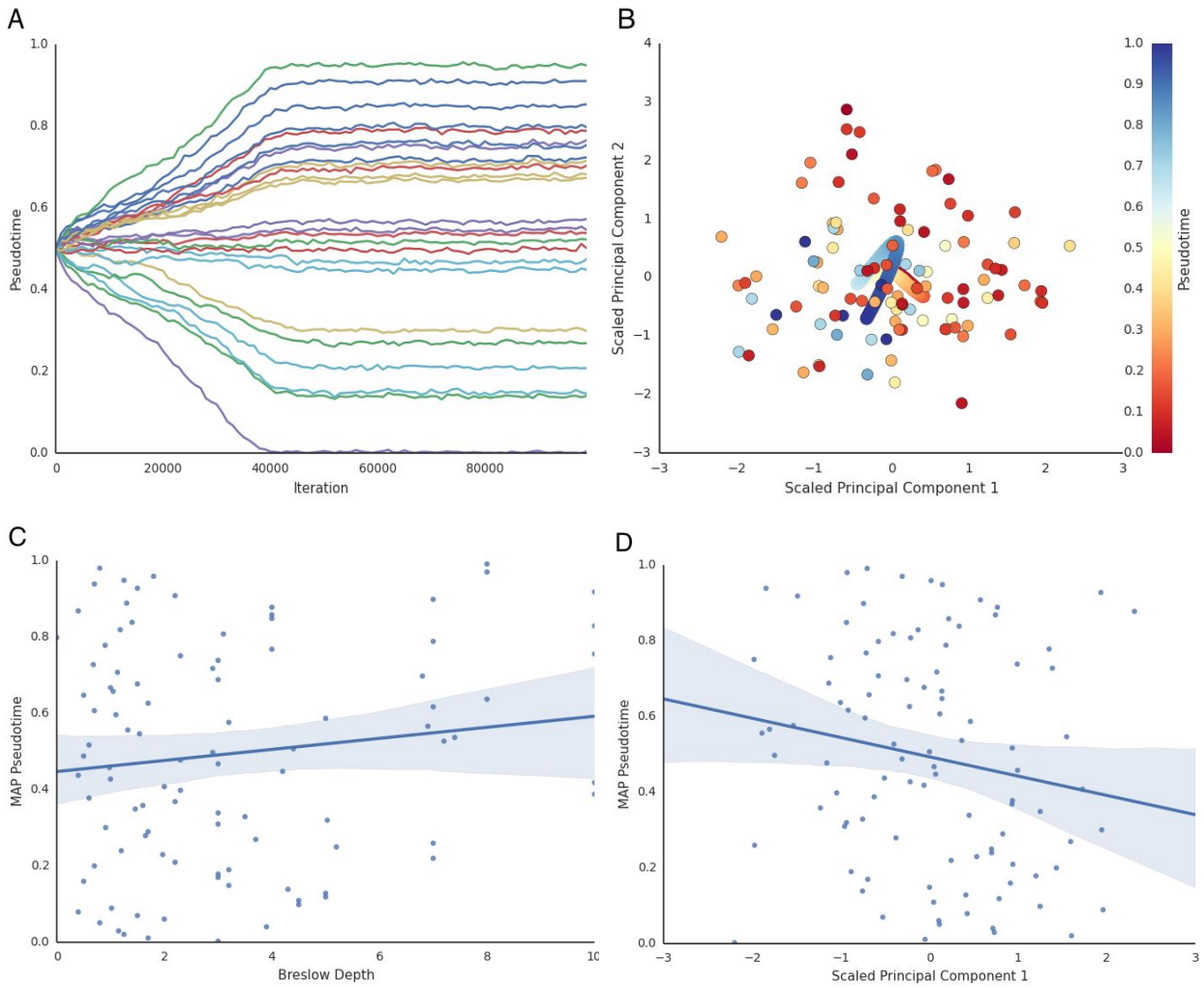

Figure 2: **Pseudotime inference of mRNA-seq data of skin cutaneous melanomas. A.** Thinned traces for 25 randomly chosen samples and their pseudotimes over the course of the Metropolis-Hastings algorithm. **B.** 100 randomly chosen samples plotted on their first two principal components along with the posterior estimate of the sample function $\boldsymbol{\mu}(t)$. Points are colored with their 'true' pseudotimes (based on Breslow depth) while the curve is colored with MAP pseudotimes. **C.** Linear regression of the MAP pseudotimes and their corresponding Breslow depths for the same 100 samples. The shaded region represents the 95% confidence interval of the regression. **D.** Linear regression of the MAP pseudotimes and their center-scaled first principal component for the same 100 samples. The shaded region represents the 95% confidence interval of the regression.

4. Conclusions

We have presented a Bayesian inference method for inferring pseudotime from RNA-seq data from skin cutaneous melanomas. Our model utilizes Gaussian Process Latent Variable Modelling, a Coulomb repulsive process prior, and a Metropolis-Hastings random walk to determine a probabilistic posterior distribution of pseudotimes.

By applying our method to synthetic data, we have shown that our algorithm can effectively and accurately recover pseudotime orderings. However, applying our method to real melanoma mRNA-seq data was not nearly as conclusive. Unfortunately, we had to dramatically reduce the dimensionality of our data due to computational limitations, and thus our results are not very meaningful. Ideally we would run our algorithm over all 959 target genes in the dataset, however it is often more practical to reduce this dimension first. Using PCA may not be the most effective form of dimensional reduction, and we hope to investigate other methods, like Laplacian Eigenmaps, in the future.

In additional future extensions of our work we would also like to expand our pseudotime ordering algorithm to incorporate other cancer features beyond mRNA-seq data. Due to our inconclusive results, we will need to refine our methodology before expanding our input data to include these features. Eventually, we would like to also explore how effective a GPLVM-based pseudotime ordering algorithm is in predicting treatment response.

We would like to thank Yue Li and Alvin Shi from the Kellis Lab in CSAIL at MIT for their guidance in statistical machine learning models and cancer genomics.

5. References

1. Steward, B. W. and Kleihues, P. “World Cancer Report.” *IARCPress*. Lyon. 2003.
2. Trapnell, C. *et al*. “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.” *Nature Biotechnology* **32**, 381-386. 2014.
3. Ji, Z. and Ji, H. “TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis.” *Nucleic Acids Research* **44** (13), e117. 2016.

4. Bendall, S. C. *et al*. “Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.” *Cell* **157**, 714–725. 2014.
5. Campbell, K., Ponting, C., and Webber, C. “Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles.” 2015.
6. Lawrence, N. D. “Gaussian process latent variable models for visualisation of high dimensional data.” *Advances in neural information processing systems* **16**, 329–336. 2004.
7. Campbell, K. and Yau, C. “Bayesian Gaussian Process Latent Variable Models for pseudotime inference in single-cell RNA-seq data.” (submitted). 2015.
8. Wang, Y. and Dunson, D. B. “Probabilistic Curve Learning: Coulomb Repulsion and the Electrostatic Gaussian Process.” *Advances in Neural Information Processing Systems*. 2015.
9. Titsias, M. K. and Lawrence, N. D. “Bayesian Gaussian Process Latent Variable Model.” *Artificial Intelligence* **9**, 844-851. 2010.
10. Rasmussen, C. E. and Williams, C. K. I. “Gaussian Processes for Machine Learning.” *The MIT Press*. 2006.
11. Titsias, M. K., Lawrence, N. and Rattray, M. “Markov chain Monte Carlo algorithms for Gaussian processes.” *Inference and Estimation in Probabilistic Time-Series Models* **9**. 2008.
12. Akbani, *et al*. “Genomic Classification of Cutaneous Melanoma.” *Cell* **161**. 2015.