Fast Out-of-Sample Predictions for Bayesian Hierarchical Models of Latent Health States, using Importance Sampling

Aaron J Fisher, R Yates Coley, Scott L Zeger

October 19, 2015

Abstract

Hierarchical Bayesian models can be especially useful in precision medicine settings, where clinicians are interested in estimating the patient-level variables associated with a specific person's risk. Such models are often fit using batch Markov Chain Monte Carlo (MCMC). However, the slow speed of batch MCMC computation makes it difficult to implement in clinical settings, where immediate risk estimates are often desired in response to new patient data. For models fit on protected clinical data from multiple hospitals, this computation is exacerbated by the algorithm's need to communicate between firewalled servers. In this report, we discuss how importance sampling (IS) can be used to obtain fast, in-clinic risk estimates, while naturally avoiding the issue of server communication. We apply IS to the hierarchical model proposed in Coley et al. (2015) for predicting risk of aggressive prostate cancer. Risk estimates via IS can typically be obtained in 1-10 seconds per person, and have high agreement with estimates coming from longer-running batch MCMC methods. Alternative options for out-of-sample fitting and online updating are also discussed.

Introduction

Hierarchical Bayesian models can be especially useful in the context of precision medicine, when scientists are interested in estimating not only treatment effects, but also the risk for any individual patient. In the context of hierarchical models, treatment effects can be estimated from population-level parameters, and the risk for any specific patient can be estimated from patient-level variables. For example, Coley et al. (2015) use a patient-level latent class to categorize patients as having either indolent or aggressive cancer. In a Bayesian setting, estimation of both levels of the model – the population level and the patient level – can be attained from the posterior distribution. When such models are fit on a training dataset using Markov Chain Monte Carlo (MCMC), risk estimates are immediately available for any patient in the training dataset.

A computational challenge arises though when new patients enter the clinic, or when existing patients accrue new measurements. Here, clinicians may wish to give patients a fast, in-visit estimate of their risk. However, batch MCMC approaches for new risk estimates that require refitting the entire model can take hours to complete. Instead, sampling algorithms tailored for out-of-sample fitting can be used to get fast risk estimates in response to new patient data. In this report, we specifically describe how Importance Sampling (IS) (Bishop et al., 2006) can be used to such an end. We apply IS to the prostate cancer model proposed by Coley et al. (2015) to get fast risk estimates for new, simulated patients. In this case, the procedure typically takes only 1-10 seconds per

patient. This approach can be combined with periodic refitting of the entire model via MCMC, in order to update the posteriors for the the population-level parameters (Lee and Chia, 2002).

This IS approach is related to online (or streaming) learning methods, which aim to continuously update population-level parameters with a constant computational cost over time. We avoid a fully online approach here though, due to additional known challenges in online learning. Specifically, our use of IS can be viewed as a 1-step version of a sequential importance sampler (SIS), also known as particle filter. Employing a standard particle filter to update estimates of the population parameters would seem to be a natural extension. However, particle filters are known to suffer from the problem of degeneracy, which makes it difficult to estimate posteriors for "static" parameters that do not change as more data is acquired (Kantas et al. (2014), see section II of Andrieu et al. (2005) for an intuitive explanation). This applies in our case, as our population-level parameters are assumed to be static. Instead, we combine IS with periodic MCMC (Lee and Chia, 2002) to update the posteriors for all parameters and latent variables at all levels. Note that this is not a fully online method, as the computational cost of MCMC increases as more data is acquired. [1]

Online model fitting has also been explored in the literature on topic modeling for corpuses of texts. Text corpuses are often too large to fit an entire model on at once, making online fitting a more feasible option. Hoffman et al. (2010) propose a online variational Bayes approach for topic modeling. Canini et al. (2009) propose a particle filter approach, in a context where the static parameters can be integrated out. However, in the setting of Canini et al. (2009), even the best performing online methods were outperformed by batch (non-online) MCMC, and generally did not improve in accuracy as more data was incorporated.

Our specific context within precision medicine is different that of topic modeling in that while the model is complex and contains several layers, the data can be fully stored in memory at once. Thus, while the approach of combining IS with periodic MCMC is not fully online and not feasible for text analysis, it is still a feasible option for the limited sample sizes in our problem. Relative to variational Bayes approaches, the formulas required to apply IS are simple to derive, and can be easily ported to other applications within precision medicine.

The remainder of this document is organized as follows. In Section 1 we give an overview of our motivating data example of prostate cancer risk estimation. In Section 2 we detail our approach for applying IS in hierarchical models. We use an abbreviated notation that can be readily generalized to other precision medicine settings. In this section we also conceptually compare our approach with Rejection Sampling (RS), Gibbs Sampling approaches, and with the out-of-sample fitting approach of Wu et al. (2015). However, with the exception of RS, do not explore the performance of these alternate methods here. In Section 3 we apply IS to simulated data, and compare the results to risk estimates obtained from batch MCMC.

1 Clinical Application & Motivation

Our application is based on the clinical framework of Coley et al. (2015), who develop a latent class model to predict underlying prostate cancer state in men participating in an active surveillance program for low risk disease. The latent cancer state is defined as being indolent or aggressive, corresponding to the Gleason score (Gleason, 1977, 1992) that would be assigned if a patient's entire prostate were to be removed and analyzed. Gleason scores < 6 are classified as indolent, and Gleason scores ≥ 7 are classified as aggressive. The model is used to estimate probabilities of latent class membership, or, in other words, the risk of having an aggressive cancer with the potential to metastasize. Risk predictions can then be used by clinicians and patients to make decisions about

[1] See Kantas et al. (2014) for a recent literature review of particle methods in the context of static parameters.

future treatment or biopsies. This prediction tool addresses a pressing need in prostate cancer care as the most common treatments for prostate cancer have a high risk of persistent side effects including erectile dysfunction and urinary incontience, while prostate biopsies are painful and pose a risk of infection Chou et al. (2011a,b).

The hierarchical model of Coley et al. (2015) includes sub-models for longitudinal prostate specific antigen (PSA) measurements, and for longitudinal biopsy results. Both of these sub-models incorporate information about the patients latent state. The (log-transformed) PSA measurements are modeled as multivariate normal, with a mean defined by a linear predictor that includes subject-specific random effects. The distribution of these random effects is modeled to depend on latent class membership. Biopsy results are coded as binary outcomes denoting *grade reclassification* on a biopsy, that is the biopsied tissue was aside a Gleason score of 7 or higher. The log-odds of reclassification is also modeled with a linear predictor whose value depends on a patient's latent state, reflecting the imperfect sensitivity and specificity of the biopsy procedure. Each patient's latent class is assumed constant over the surveillance period. As a patient continues in active surveillance, additional PSA and biopsy measurements are accrued and the accuracy of latent class predictions improves. Sub-models are also included for informative observation processes associated with biopsies and surgeries.

In our context, the subject-level variables refer to the latent classes, and the random effects used in the sub-model for PSA. The population-level parameters refer to the coefficients in each sub-model, and the variance parameters for the subject-level variables. See Coley et al. (2015) for a full model description.

2 Methods for Fast Prediction Updates

In this section we detail an IS algorithm that enables rapid estimates of subject-level variables, such as latent classes. This method is meant to be applied to out-of-sample data, after MCMC has been applied to get a posterior sample based on current training data. We present the algorithm in a simple, abbreviated notation that is applicable in many clinical settings.

Let the joint posterior based on training data from n subjects be denoted as

$$p(\theta, b_{1:n}|y_{1:n}) \propto \prod_{i=1}^{n}[f(y_i|b_i,\theta)g(b_i|\theta)]\pi(\theta) \tag{1}$$

where y_i is the vector of clinical measurements (here, PSA and biopsy measurements) for patient i, $y_{1:n}$ is the list of measurements for the first n patients, b_i is a vector of latent variables (here, latent class and random effects) for patient i, $b_{1:n}$ is a list of latent variables for the first n patients, θ contains the population-level parameters, π is the prior for θ, and f and g are multivariate distributions, which will depend on the application and context. Estimation of b_i is of primary interest in this report.

Let $\mathscr{J}_n = \{\theta^{(j)}, b_{1:n}^{(j)}\}_{j=1}^{J}$ be a set of J draws from the posterior distribution in Eq 1, obtained via methods such as MCMC.

2.1 IS Algorithm

After posterior samples from the joint model are obtained for current data, importance sampling to update these estimates given new data requires three steps: (1) generating proposal values for the latent variables to be updated, (2) calculating weights for proposed values, and (3) weighting proposed values to estimate an updated posterior. We first illustrate how this process can be used to quickly estimate latent variables for a new patient,

and then show how similar calculations can be done to incorporate newly measured data on existing patients in real-time.

For a new patient (indexed by $i = n+1$), obtaining posterior predictions of latent variables requires calculating expectations with respect to the posterior distribution based on all $n+1$ patients (i.e. $p(\theta, b_{1:(n+1)}|y_{1:(n+1)})$). While we cannot immediately draw from this distribution, we can evaluate a function that is proportional to its density (based on Eq 1). We can also use the posterior distribution based on the first n patients as a proposal distribution (denoted by q) from which to generate candidate values of $(\theta, b_{1:(n+1)})$. Let

$$q(\theta, b_{1:(n+1)}) \quad := \quad g(b_{n+1}|\theta)p(\theta, b_{1:n}|y_{1:n}) \tag{2}$$

Practically, this proposal step is achieved by conditioning on each $\theta^{(j)}$ in $\mathscr{J}_n$, and then of drawing $b_{n+1}^{(j)}$ from the distribution $g(b_{n+1}^{(j)}|\theta^{(j)})$. This results in the augmented set $\mathscr{J}_{n+1} := \{\theta^{(j)}, b_{1:(n+1)}^{(j)}\}_{j=1}^{J}$. The importance weights $w^{(j)}$ are then proportional to

$$\begin{aligned} w^{(j)} &\propto \frac{p(\theta^{(j)}, b_{1:(n+1)}^{(j)}|y_{1:(n+1)})}{q(\theta^{(j)}, b_{1:(n+1)}^{(j)})} \\ &\propto \frac{\prod_{i=1}^{n+1}[f(y_i|b_i^{(j)}, \theta^{(j)})g(b_i^{(j)}|\theta^{(j)})]\pi(\theta^{(j)})}{g(b_{n+1}^{(j)}|\theta^{(j)})\prod_{i=1}^{n}[f(y_i|b_i^{(j)}, \theta^{(j)})g(b_i|\theta^{(j)})]\pi(\theta^{(j)})} \\ &= f(y_{n+1}|b_{n+1}^{(j)}, \theta^{(j)}) \end{aligned} \tag{3}$$

The final weights $w^{(j)}$ are standardized to sum to 1. The new posterior for $(\theta, b_{1:(n+1)})$ can then be represented as the mixture distribution satisfying $P(\theta = \theta^{(j)}, b_{1:(n+1)} = b_{1:(n+1)}^{(j)}) = w^{(j)}$. Posterior means for $b_{(n+1)}$ can be calculated as $\sum_{j=1}^{J} w^{(j)} b_{(n+1)}^{(j)}$.

The approach is similar when we wish to incorporate new measurement data for a patient who's previous data has already informed the posterior sample in Eq 1. For a patient k with existing data (i.e., $k \leq n$), our set $\mathscr{J}_n$ already contains the proposals $\{b_k^{(j)}\}_{j=1}^{J}$ for the subject k's latent variable values. Thus, we can draws from $\mathscr{J}_n$ as our proposal distribution $q(\theta^{(j)}, b_{1:n}^{(j)})$. Our goal then is to re-weight this set of points based on new data. Let $y_{1:n}^*$ refer to the data set after incorporating new data on patient k, such that $y_i^* = y_i$ if and only if $k \neq i$. The importance weights in Equation 3 then simplify to

$$\begin{aligned} w^{(j)} &\propto \frac{p(\theta^{(j)}, b_{1:n}^{(j)}|y_{1:n}^*)}{q(\theta^{(j)}, b_{1:n}^{(j)})} \\ &\propto \frac{\prod_{i=1}^{n}[f(y_i^*|b_i^{(j)}, \theta^{(j)})g(b_i^{(j)}|\theta^{(j)})]\pi(\theta^{(j)})}{\prod_{i=1}^{n}[f(y_i|b_i^{(j)}, \theta^{(j)})g(b_i^{(j)}|\theta^{(j)})]\pi(\theta^{(j)})} \\ &= \frac{f(y_k^*|b_k^{(j)}, \theta^{(j)})}{f(y_k|b_k^{(j)}, \theta^{(j)})} \end{aligned} \tag{4}$$

If the repeated measures for each patient are independent conditional on b_i, as is the case in the proposed model from Coley et al. (2015), then the ratio in Eq 4 reduces to the likelihood of only the new data, conditional on $b_k^{(j)}$ and $\theta^{(j)}$.

2.2 Efficient Implementation

For implementation in clinical practice, proposals for new patients can be generated prior to actually observing new data, so that only weight calculation and re-weighting of the proposal distribution needs to be done in real-time.

By random chance, some new patients may have data such that very few of the pre-generated, proposed latent variables values receive high weights. This can cause their posterior mean estimates to be less stable. This problem is due to higher Monte Carlo error, and is thus more likely when the number of pre-generated latent variables (J) is low. However, we can flag patients who might have high error by monitoring the effective size of the posterior sample, also known as the effective number of particles ($1/\sum_{j=1}^{J}\left[\left(w^{(j)}\right)^2\right]$). When this number drops below a given threshold (e.g. 1000), we can repeat our procedure with a larger set of pre-generated proposals. If limited computing is available for MCMC, we can also approximate a larger sample from Eq 1 by drawing multiple b_{n+1} values for each $\theta^{(j)}$, rather than drawing just one (see Eq 2).

2.3 Alternative Out-of-Sample Posterior Estimations

Our IS approach functions similarly to the out-of-sample estimation approach of Wu et al. (2015). Their approach can be generalized to estimate the updated posterior probability that $P(b_{n+1} = x|y_{1:(n+1)})$, using the estimator $\hat{P}(b_{n+1} = x|y_{1:(n+1)}) = \left(\frac{1}{P}\right)\sum_{j=1}^{P}\left\{\frac{f(y_{n+1}|b_{n+1}=x,\theta^{(j)})g(x|\theta^{(j)})}{\int f(y_{n+1}|b_{n+1}=x',\theta^{(j)})g(x'|\theta^{(j)})dx'}\right\}$. This approach is especially practical when the subject-specific variables b_{n+1} are discrete, and the integral in the denominator can be replaced with a summation. For cases with both continuous and discrete subject-specific variables, the approach can be combined with a proposal generation method based on Eq 2.

Rejection sampling can also be applied using the unstandardized weights in Eq 3, although we found this approach to be less computationally efficient than IS for our scenario (see Section 3.1).

Another approach out-of-sample estimation approach is the use of Gibbs Sampling to update only parameters associated with new patient data. One simple implementation is to run separate MCMC chains, each initialized on a different element of $\mathscr{J}_n$. An alternate approximate implementation that combines these chains is to treat the set $\{\theta_n^{(j)}\}_{j=1}^{J}$ as fixed, and to create a proxy categorical parameter z according to the following heirarchical distribution

$$\begin{aligned}
\mathbf{p} &\sim \text{Dirichlet}(\alpha = \mathbf{1}_J) \\
z &\sim \text{Categorical}(\mathbf{p}) \\
b_{n+1} &\sim g(\theta^{(z)}) \\
y_{n+1} &\sim f(b_{n+1}, \theta^{(z)})
\end{aligned}$$

Where $\mathbf{p}$ is a J-length vector of probabilities, $\mathbf{1}_J$ is a J-length vector of ones, and z is a scalar such that $P(z = j) = \mathbf{p}_j$ for $j = 1, 2, ...J$. The above model can then be fit with traditional Gibbs Sampling, and the resulting posterior estimates for $\mathbf{p}$ are analogous to the weights in Eq 3. We do not explore the performance of these alternate approaches in this report.

3 Application

We applied the proposed IS approach to simulated data based on the Johns Hopkins Active Surveillance (JHAS) cohort. 1,298 men with very low or low risk prostate cancer diagnoses were enrolled in JHAS from January 1995 to June 2014. Results of all PSA tests and biopsies performed prior to enrollment and during active surveillance were collected. Patients were followed until grade reclassification, elective treatment, or loss to follow-up. Patients still active in the program were administratively censored at the time of data collection for this analysis (October 2014). The Gleason score determination based on pathologic analysis of the entire prostate specimen was also recored for patients who underwent prostatectomy. Details on the dataset are available in Coley et al. (2015).

We applied IS to a simulated dataset of 1,000 patients. The model proposed in Coley et al. (2015) was used as the data generating model, with parameter values set equal to their corresponding posterior mean estimates from fitting the model to the JHAS data. Covariates to the model (age and date of diagnosis) were each generated from a normal distribution with mean and variance equal to that observed in JHAS patients. See Coley et al. (2015) for more details on model specification and covariates.

Using this data as our initial sample ($y_{1:n}$), we generate 500,000 draws ($\mathscr{J}_n$) from the posterior for the population-level and subject-level variables (see Eq 1). Averaging over $\mathscr{J}_n$, we estimate the risk of having aggressive cancer for each subject who's latent class is unknown. The task of generating $\mathscr{J}_n$ was run across 400 parallel jobs on a x86-based linux cluster, with as many as 200 jobs allowed to run simultaneously. The total elapsed computation time was 33 hours. Within each job, MCMC was implemented using the R2jags software package (Su and Yajima, 2015).

We then re-estimate each patient's risk using IS, taking as input only the population-level parameter posteriors from the MCMC step. When generating values for the subject-level variables b_i we further increase the diversity of the proposal set by we drawing 10 values from $g(b_i|\theta^{(j)})$ for each posterior draw $\theta^{(j)}$, for a total of 5 million proposals. We experimented with approaches of using only 50,000 proposals, using all 5 million proposals, or starting with 50,000 and increasing number of proposals until the effective sample size exceeds 1000. We refer to these approaches respectively as "small," "big," and "dynamic". Within each approach, the final set of proposals are then weighted to obtain IS risk estimates. These simulation steps are meant to approximate the procedure of using IS to get risk estimates for a new patient, under the assumption that any individual patient has only a minor affect on the population-level parameter posteriors. In section 3.1, we assess coherence between IS risk estimates and MCMC risk estimates.

3.1 Results

We find a high degree of coherence between estimated risk of aggressive cancer from IS and from MCMC, as shown if Figure 1. With the "dynamic" proposal approach, the root mean square of the difference (rMSD) between these two sets of risk estimates was 1.05% (on the probability scale, from 0% to 100%). The maximum absolute difference was 5.6%, with 95% of patients having a difference less than 2.4%. Estimation time per patient ranged from approximately 1-24 seconds, with an interquartile range of 2.1-4.2 seconds. We also considered a rejection sampling approach using the unstandardized weights in Eq 3, but found the results to have a greater deviation from the MCMC estimates (rMSD = 1.82%).

Figure 3 illustrates the roughly inverse relationship between the effective sample size used for IS, and the difference between IS and MCMC risk estimates. In general, the "dynamic" approach had an accuracy comparable to the "big" approach, but with a substancially lower computation time. Computation time for the "big" approach ranged from 9-36 seconds, with an interquartile range of 21-26 seconds.

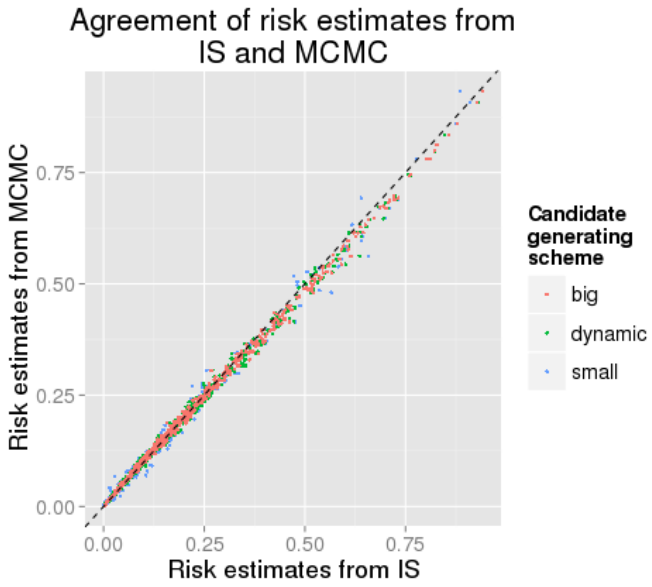

Figure 1: Agreement between IS and MCMC estimates for the posterior predictions of aggressive prostate cancer state in a new patient - Color of the points refers to the number of candidate points used, either 50,000 (small), 5 million (big), or dynamic. The dashed line indicates the axis of equality (i.e., perfect agreement).

These findings suggest that the proposed IS algorithm can be an appropriate substitute for full MCMC runs in order to provide real-time updates in a clinical setting.

4 Discussion

The joint model of Coley et al. (2015) is among a growing number of statistical models for making individualized health predictions and recommendations. Development of such precision medicine methods must occur within a framework for clinical implementation. Specifically, concerns about convenience, security, and effective communication must be addressed alongside statistical considerations. In this technical report, we present a fast implementation of the of latent health state model proposed in Coley et al. (2015), using importance sampling to generate in-clinic predictions. This approach informs decision-making by enabling doctors and patients to access updated predictions in real-time in a clinical setting.

Supplemental Code

Code for simulating data, obtaining IS estimates, and comparing the results against MCMC estimates, is available at: `https://github.com/aaronjfisher/prostate_surveillance/tree/master/IS-demo`

References

Andrieu, C., Doucet, A., and Tadic, V. B. (2005). On-line parameter estimation in general state-space models. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference*

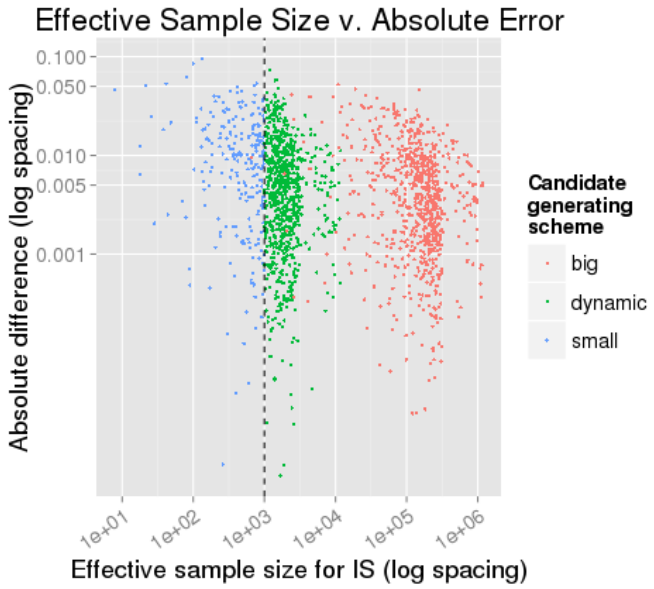

Figure 2: Difference between IS and MCMC risk estimates, as a function of effective sample size for IS - Color of the points refers to the number of candidate points used, either 50,000 (small), 5 million (big), or dynamic. The dotted verticle line shows the threshold used for dynamic proposal generation, at 1000. Both axes are shown with log scale spacing.

on, pages 332–337. IEEE.

Bishop, C. M. et al. (2006). *Pattern recognition and machine learning*, volume 4. springer New York.

Canini, K. R., Shi, L., and Griffiths, T. L. (2009). Online inference of topics with latent dirichlet allocation. In *International conference on artificial intelligence and statistics*, pages 65–72.

Chou, R., Croswell, J. M., Tracy, D., Bougatsos, C., Blazina, I., Fu, R., Gleitsmann, K., Koenig, H. C., Lam, C., Maltz, A., Rugge, J. B., and Lin, K. (2011a). Screening for prostate cancer: a review of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 155:762–771.

Chou, R., Dana, T., Bougatsos, C., Fu, R., Blazina, I., Gleitsmann, K., and Rugge, J. B. (2011b). Treatments for Localized Prostate Cancer: Systematic Review to Update the 2002 U.S. Preventive Services Task Force. Evidence Synthesis No. 91. ARHQ Publication No. 12-0516-EF-2. Rockville, MD: Agency for Healthcare Research and Quality.

Coley, R. Y., Fisher, A. J., Mamawala, M., Carter, H. B., Pienta, K. J., Zeger, and L, S. (2015). Bayesian joint hierarchical model for prediction of latent health states with application to active surveillance of prostate cancer.

Gleason, D. (1977). The Veteran's Administration Cooperative Urologic Research Group: Histologic grading and clinical staging of prostatic carcinoma. In Tannenbaum, M., editor, *Urologic Pathology: The Prostate*, pages 171–198. Lea and Febiger, Philadelphia.

Gleason, D. F. (1992). Histologic grading of prostate cancer: a perspective. *Human pathology*, 23(3):273–279.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 856–864.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J. M., and Chopin, N. (2014). On particle methods for parameter estimation in state-space models. *arXiv preprint arXiv:1412.8695*.

Lee, D. S. and Chia, N. K. (2002). A particle algorithm for sequential bayesian parameter estimation and model selection. *Signal Processing, IEEE Transactions on*, 50(2):326–336.

Su, Y.-S. and Yajima, M. (2015). *R2jags: A Package for Running jags from R*. R package version 0.05-01.

Wu, Z., Deloria-Knoll, M., Hammitt, L. L., and Zeger, S. L. (2015). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.