Fast Predictions for Bayesian Hierarchical Models of Latent Health States (Technical Report)

Aaron J Fisher, R Yates Coley, Scott L Zeger

August 28, 2015

Abstract

This technical report discusses the use of importance sampling in precision medicine settings, to obtain fast, in-clinic risk estimates for out-of-sample patients with new data. We apply importance sampling to the hierarchical model proposed in Coley et al. (2015) for predicting risk of aggressive prostate cancer. Estimates can be obtained in 2-5 seconds per person, and have high agreement with estimates coming from longer-running Markov Chain Monte Carlo (MCMC) methods. Alternative options for online updating of the model are also discussed.

Introduction

Hierarchical bayes models can be especially useful in the context of precision medicine, when scientists are interested not only in estimating treatment effects, but also in estimating 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 described by a patient-level latent class. For example, Coley et al. (2015) use a patient specific latent class to identify patients as having either benign cancer, or aggressive cancer. In a bayesian setting, estimation at both levels of the model, the population level and the patient level, can be done by averaging over the posterior. 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, traditional MCMC approaches for new risk estimates require refitting the entire model, which can take hours to complete.

In this technical report we describe how importance sampling (IS) can instead be used in precision medicine settings to get fast risk estimates in response to new patient data. We apply this to a version of the prostate cancer model proposed by Coley et al. (2015) to get fast risk estimates for new, simulated patients. Here, the procedure only takes approximately 2-5 seconds per patient. This approach can be combined with periodic refitting of the entire model via MCMC, in order to update estimates of the population-level parameters (Lee and Chia, 2002).

This IS approach is related to online learning methods, which aim to continuously update population-level parameters at a constant computational cost over time. We avoid a fully online approach 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 generic particle filter to get updated estimates of the population parameters would seem to be a natural extension, but particle filters are known to suffer from the problem of degeneracy in the presence of "static" parameters (see section II of Andrieu et al. (2005) for an intuitive explanation). This applies in our case, where the population-level parameters are modeled as static, and changing as more data is acquired. A combination of IS and periodic MCMC (Lee and Chia, 2002) can be used to update population-level posteriors, but is not fully online as the computational cost of each MCMC iteration increases as more data is acquired. See Kantas et al. (2014) for a recent literature review of online methods relating to particle filtering.

Online, or streaming, model fitting has 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. Our specific context within personalized 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 topic modeling, it is still a feasible option for limited sample sizes in our problem. An additional benefit of IS is that, 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 a clinical overview of our motivating data example of prostate cancer risk estimation. In Section 2 we detail the general approach for applying IS in hierarchical models. We use an abbreviated notation that can be readily generalized to other precision medicine settings. In Section 3 we apply IS to simulated data, and compare the results to risk estimates obtained from 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 whether a patient's prostate cancer is indolent or aggressive. This latent cancer state can be interpreted as corresponding to the Gleason score (Gleason, 1977, 1992) that would be assigned to the patient if his entire prostate were to be removed and a pathologic analysis was performed. Gleason scores < 6 are classified as indolent, and Gleason scores ≥ 7 are classified as aggressive. This model is used to estimate probabilities of latent class membership, or, in other words, the risk of having aggressive cancer. Risk predictions can be used by clinicians to decide whether further biopsies or surgeries are necessary. This is an important issue within precision medicine, as unneeded biopsies have side effects of... .

Coley et al. (2015) derive a hierarchical model, which 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. Biopsy results are coded as binary outcomes, denoting whether the clinician reclassified the patient's cancer as aggressive. The log-odds of reclassification is also modeled with a linear predictor.[1] The final version of the hierarchical model in (Coley et al., 2015) also includes sub-models for informative observation processes associated with biopsies and surgeries. As a first step though, we look here only at the model version that treats the data as missing at

[1] Patients are modeled to have the same latent class throughout their participation in the study. As the study progresses, clinicians reclassify patients with the goal of better aligning all patients with their true latent status.

random.

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 Importance Sampling Algorithm for Fast Prediction Updates

In this section we detail an importance sampling algorithm that enables rapid updates of joint latent class model predictions. This method is meant to be applied to out-of-sample data, after MCMC has been applied to get a posterior sample based on all 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.

After posterior samples from the joint model are obtained for current data, importance sampling to update these estimates given new data requires three steps: first, generating proposal values for the latent variables to be updated, second, calculating weights for proposed values, and, third, 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 (Equation 1). The posterior distribution based on the first n patients provides an appropriate proposal distribution (q) from which to generate candidate values of $(\theta, b_{1:(n+1)})$:

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

Practically, this step consists of taking J draws of θ and $b_{1:n}$ from the previously fitted posterior in Equation 1. Then, conditional on θ, we draw b_{n+1} from the distribution g. We index each of the resulting draws as $(\theta^{(j)}, b_{1:(n+1)}^{(j)})$, with $j = 1, \ldots, J$. 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 $P(\theta = \theta^{(j)}, b_{1:(n+1)} = b^{(j)}_{1:(n+1)}) = w^{(j)}$. A posterior mean for $b_{(n+1)}$ can be calculated as $\sum_{j=1}^{J} w^{(j)} b^{(j)}_{(n+1)}$. [2]

For a patient k with existing data, where we already have a posterior sample for their latent variable values, we instead use this posterior as our proposal distribution $q(\theta^{(j)}, b^{(j)}_{1:n})$, with $i \leq n$. Let $y^{k+}_{1:n}$ refer to the data set after incorporating new data on patient k, where $y^+_i = y_i$ if $k \neq i$. The importance weights in Equation 3 then simplify to

$$\begin{aligned} w^{(j)} &\propto \frac{p(\theta^{(j)}, b^{(j)}_{1:n} | y^+_{1:n})}{q(\theta^{(j)}, b^{(j)}_{1:n})} \\ &\propto \frac{\prod_{i=1}^{n} [f(y^+_i | b^{(j)}_i, \theta^{(j)}) g(b^{(j)}_i | \theta^{(j)})] \pi(\theta^{(j)})}{\prod_{i=1}^{n} [f(y_i | b^{(j)}_i, \theta^{(j)}) g(b^{(j)}_i | \theta^{(j)})] \pi(\theta^{(j)})} \\ &= \frac{f(y^+_k | b^{(j)}_k, \theta^{(j)})}{f(y_k | b^{(j)}_k, \theta^{(j)})} \end{aligned}$$

Let L_k denote that number of measurements for which we've previously fit a posterior for b_k, and let N_k denote the number of new measurements we wish to incorporate into this posterior. Then, y^+_k can be expressed as the vector $y^+_k = (y_{k[1]}, y_{k[2]}, \ldots y_{k[L_k]}, y^+_{k[L_k+1]}, \ldots y^+_{k[L_k+N_k]})$, where $y^+_{k[l]}$ is the l^{th} measurement from patient k. If the repeated measures for each patient are independent conditional on b_i, as is the case in the proposed model, then the above ratio reduces to

$$\begin{aligned} w^{(j)} &\propto \frac{\prod_{l=1}^{L_k+N_k} f(y^+_{k[l]} | b^{(j)}_k, \theta^{(j)})}{\prod_{l=1}^{L_i} f(y_{k[l]} | b^{(j)}_k, \theta^{(j)})} \\ &= \prod_{l=L_k+1}^{L_k+N_k} f(y^+_{k[l]} | b^{(j)}_k, \theta^{(j)}) \end{aligned}$$

We then proceed as above to get a re-weighted posterior for the latent variables of patient k.

By random chance, some 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. When this number drops below a given threshold (e.g. 500), 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).

For implementation in clinical practice, proposals for new patients can be generated prior to actually observing new data, so that only weight calculating and re-weighting of the proposal distribution needs to be done in real-time. Then, predictions for each patient can be obtained in a matter of seconds.

[2] The unstandardized weights can also be used in a rejection sampling procedure, although we found this approach to be less computationally efficient than IS for our scenario (see Section 3.1).

3 Simulation & Performance

We assessed performance of the proposed importance sampling approach in a simulated dataset....

Using this data as our initial sample $y_{1:n}$, we generate 62,500 draws from the posterior for the populational-level and subject-level variables (see Eq 1). Let $\mathcal{P}$ denote this specific posterior sample. Averaging over $\mathcal{P}$, we estimate the risk of having aggressive cancer for each subject who's latent class is unknown. This task was run across 50 parallel jobs on a x86-based linux cluster. All jobs were run simultaneously, with a total computation time of 5 hours and 26 minutes. 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 $\theta^{(j)}$. Proposals are then weighted to obtain IS risk estimates. These 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. The root mean square of the difference (rMSD) between these two sets of risk estimates was 0.74% (on the probability scale, from 0% to 100%). The maximum absolute difference was 4.8%, with 95% of patients having a difference less than 1.6%. Estimation time per patient ranged from approximately 5-13 seconds, depending on the number of measurements to be incorporated for that patient. 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 = 1.1%). A small portion of these differences between these risk estimates is to be expected simply due to the nature of stochastic posterior sampling – for example, the rMSD between two separately run sets of MCMC risk estimates has a rMSD of 0.2%.

The bottom panel of Figure 2 illustrates the roughly inverse relationship between the effective sample size used for IS, and the difference between IS and MCMC risk estimates. Even with a large posterior sample $\mathcal{P}$, we still found it advantageous to draw 10 values of b_i for each draw $\theta^{(j)}$ from $\mathcal{P}$. The top panel of Figure 2 shows what happens if we instead draw only one value of b_i for each $\theta^{(j)}$. Here, effective sample sizes drop, and deviations between MCMC and IS estimates increase. Large deviations could still be flagged though, by setting a threshold for effective sample size anywhere between 50 and 500.

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

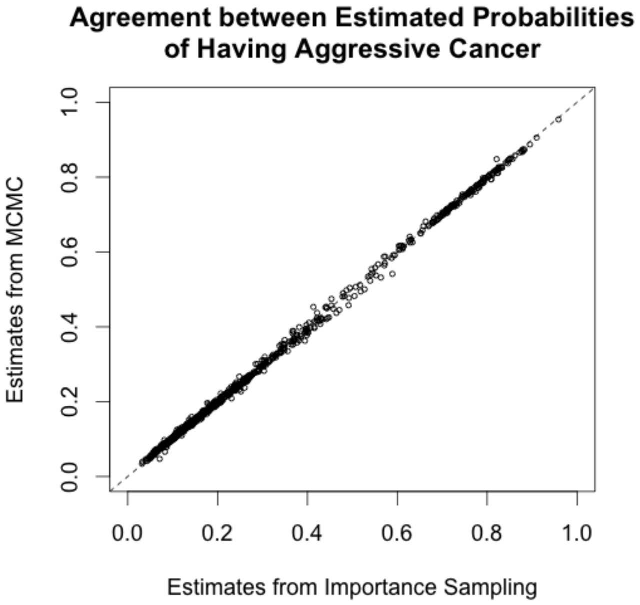

Figure 1: Agreement between IS and MCMC estimates for the posterior predictions of aggressive prostate cancer state in a new patient. (Dashed line indicates the axis of equality, i.e., perfect agreement.)

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 on*, pages 332–337. IEEE.

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.

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

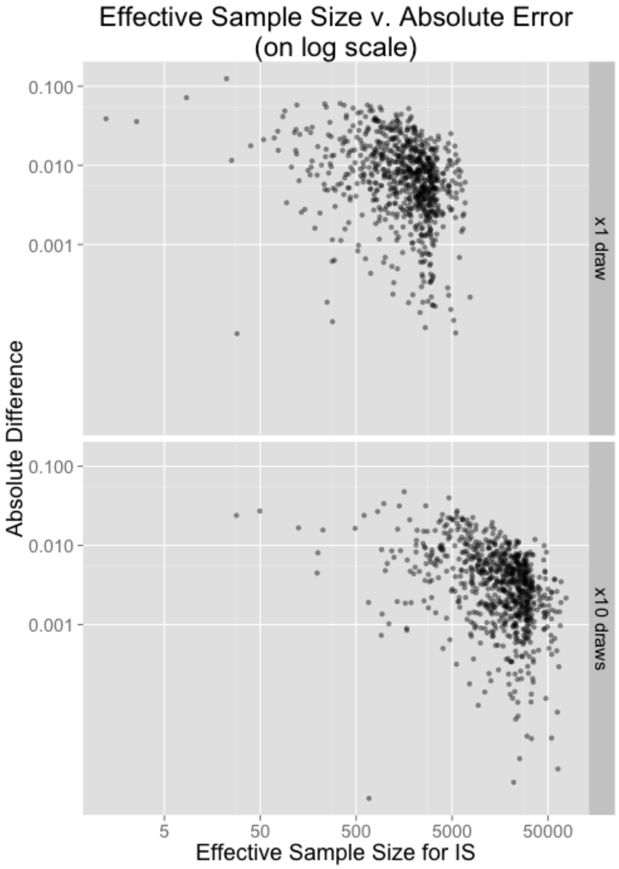
Effective Sample Size v. Absolute Error
(on log scale)
Absolute Difference
0.100
0.010
0.001
x1 draw
x10 draws
5
50
500
5000
50000
Effective Sample Size for IS

Figure 2

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.