1 Fast latent variable estimates for new patient data

Ideally, physicians would like to give patients fast, in-visit risk estimates whenever new lab results are acquired. A standard implementation of our approach would entail re-running MCMC to get updated posteriors for the subject's latent variables, which can take hours to complete. Instead, we use importance sampling (Bishop et al., 2006, chapter 11) to get fast latent variable posterior estimates. This can be combined with MCMC periodically (e.g. every two weeks) as more patients are acquired, to update the population-level parameter posteriors.[1] It may also be possible to attain a fast, "online" update of the population-level parameter posteriors, but there are known obstacles to this type of updating which push its solution beyond the scope of our current work.

In order to generate proposal values for importance sampling, we start with draws from the posterior of the population-level parameters, obtained by fitting model refXX on the previously observed data. For each draw, we use the conditional distributions in Equation refXX to generate proposed latent variable values for the subject with new data. The importance weights for these proposed latent variable values are then proportional to the likelihood of the newly acquired data, given the proposed parameters and latent variables. Such a procedure can be thought of a 1-step version of a sequential importance resampler, also known as a particle filter (Bishop et al., 2006). Note that proposals can be pre-generated before patients enter the clinic, so that only the weights need to be calculated in real time. Posterior means for each subject can then be computed in approximately 2 seconds.

By random chance, some patients will have data such that very few of the pre-generated, proposed latent variables values receive high weights. This causes their posterior mean estimates from importance sampling to be less stable. However, such cases can be detected by monitoring the effective size of the posterior sample, also known as the effective number of particles. When this number drops below 500, we repeat our procedure with a larger set of pre-generated proposals.

These fast estimates have a correlation of 0.9950 with the estimates from running MCMC to estimate all parameters. For reference, estimates from two different runs of the full MCMC have a correlation of 0.9993, due to the stochastic nature of the posterior sampling. Figure 1 shows the high degree of agreement between posterior risk estimates from MCMC, and from importance sampling. We give further details of the importance sampling calculations in the supplementary materials .

Our approach still requires the periodic use of MCMC to update all other parameters. It is tempting to try to use sequential importance resampling (SIRS)

[1] This approach is conceptually very similar to the approach of Lee and Chia (2002), who combine a sequential importane sampling with periodic MCMC to update all dynamic parameters. The dynamic parameters in their work are analogous to the subject-specific parameters in ours.

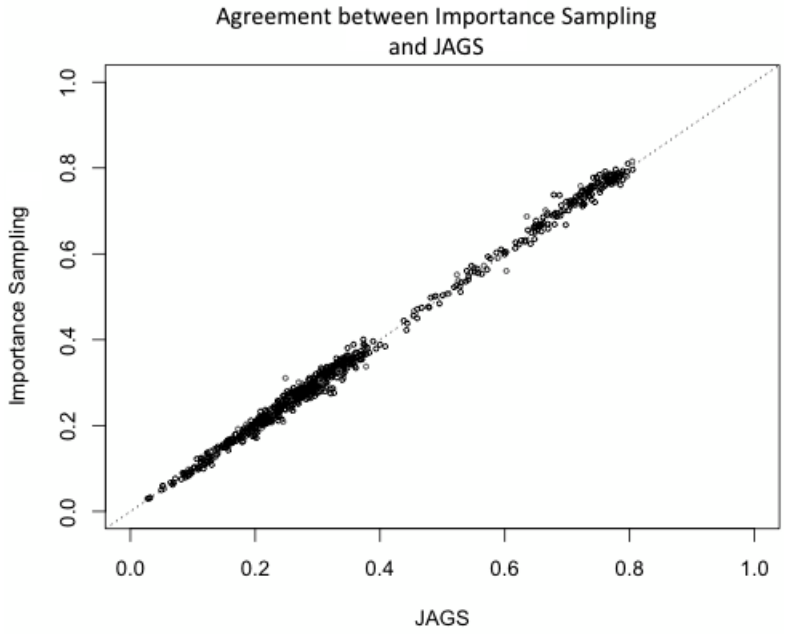

Figure 1: Agreement between MCMC (via JAGS) and Importance Sampling - (!! THIS IS A PLACEHOLDER FIGURE !!) for each subject, we plot the posterior mean value of η_i from importance sampling, and from full MCMC. The correlation between these two estimates is 0.9950.

to estimate posteriors in a fully online fashion, removing the computation costs of periodic MCMC. However, such online methods are known to suffer from the problem of "degeneracy" when the model includes static parameters (e.g. population-level parameters) in addition to dynamic parameters (e.g. subject-specific latent variables). For an intuitive discussion of this degeneracy, see section II of Andrieu et al. (2005). Some proposed alternatives include fixed lag updating methods (Polson et al., 2008); augementation of the static parameters (Kitagawa, 1998); and variational bayes approaches (Hoffman et al., 2010). See Kantas et al. (2014) for a recent literature review.

2 Supplement

2.1 Importance sampling procedure

For the purposes of this section, we introduce the following abbreviated form of the model in XX. Let the posterior for our model be

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

Where y_i is the vector of measurements for subject i, $y_{1:n}$ is the list of measurements for the first n subjects, b_i is a vector of latent variables for subject i, $b_{1:n}$ is a list of latent variables for the first n subjects, θ contains the population-level parameters, π is the prior for θ, and f and g are multivariate distributions coming from the likelihood in XX.

In section 2.1.1, we illustrate how importance weighting can be used to estimate latent variables for a new subject entering the study. In section 2.1.2, we show how similar calculations can be done to quickly incorporate newly measured data on existing patients.

2.1.1 Fast estimates for a new patient

Here, we focus on estimating latent variable posteriors for a new subject (indexed by $n+1$). Our ultimate goal is to calculate expectations with respect to the posterior distribution based on all $n+1$ subjects (i.e. $p(\theta, b_{1:(n+1)}|y_{1:(n+1)})$). Unfortunately, we cannot immediately draw from this distribution, but we can evaluate a function that is proportional to its density (Equation 1). To carry out importance sampling, we need choose a proposal distribution q from which to generate candidate values of $(\theta, b_{1:(n+1)})$. We use the posterior distribution based on the first n subjects as our proposal distribution. This approach is analagous to a 1-step particle filter (Bishop et al., 2006).

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

Practically, this consists of taking J draws of θ and $b_{1:n}$ from the previously fitted posterior in Eq 1. Then, conditional on θ, we draw b_{n+1} from the distribution g. We index each of the resulting draws as $(\theta^{(j)}, b^{(j)}_{1:(n+1)})$, with $j = 1, \ldots, J$. The importance weights w_j are then proportional to

$$\begin{aligned} w^{(j)} &\propto \frac{p(\theta^{(j)}, b^{(j)}_{1:(n+1)}|y_{1:(n+1)})}{q(\theta^{(j)}, b^{(j)}_{1:(n+1)})} \\ &\propto \frac{\prod_{i=1}^{n+1}[f(y_i|b_i^{(j)}, \theta^{(j)})g(b_i^{(j)}|\theta^{(j)})]\pi(\theta^{(j)})}{g(b^{(j)}_{n+1}|\theta^{(j)})\prod_{i=1}^{n}[f(y_i|b_i^{(j)}, \theta^{(j)})g(b_i^{(j)}|\theta^{(j)})]\pi(\theta^{(j)})} \\ &= f(y_i|b_i^{(j)}, \theta^{(j)}) \end{aligned} \tag{2}$$

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 $\sum_{j=1}^{J} w^{(j)} b^{(j)}_{(n+1)}$. The unstandardized weights can also be used in a rejection sampling procedure, although we found this approach to be less computationally efficient than importance sampling for our scenario.

2.1.2 Fast estimates for existing patients with new data

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 $q(\theta^{(j)}, b_{1:n}^{(j)})$, with $i \leq n$. Let $y_{1:n}^{k+}$ 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 2 then simplify to

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

Let L_k denote that number of measurements for which we've previously fit a posterior for b_k. 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 $y_k^+ = (y_{k[1]}, y_{k[2]}, \dots y_{k[L_k]}, y_{k[L_k+1]}^+, \dots y_{k[L_k+N_k]}^+)$, where $y_{k[l]}^+$ is the l^{th} measurement from subject k. If the repeated measures for each subject are independent conditional on b_i, as is the case in our model, then the above ratio reduces to

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

We then proceed as in section 2.1.1 to get a re-weighted posterior for the random effects of subject k.

2.2 Full model online updates

We generally propose that importance sampling be used for fast, in-visit estimates of patient's latent risk. This can be combined with periodic MCMC to update the other latent variables and population parameters. The issue with this approach is that the computational cost of each MCMC step increases as more patients are required, making the total computation take no less than

quadratic time. The task of updating a hierarchical model in constant time is an open problem.

Some initial work on online updates has been proposed in the field of text analysis. Hoffman et al. (2010) applied a variational Bayesian approach, but this has some problems [need to explore/talk to Beka]? Canini et al. (2009) consider online sampling methods for text analysis, and recommend a particle filter approach (also known as Sequential Importance Resampling)(Need canonical citation). However, all of the online methods considered by Canini et al. do not perform as well as refitting on the entire dataset, in a non-online fashion.

Our model also differs from Canini et al. (2009) in a way that further complicate the use of particle filters. Like Canini et al., we assume that our population distribution is *static* over time. In other words, we believe that the population-level parameters do not change as we acquire new data. The presence of such static parameters is known to cause particle filters to break down (see Andrieu et al. (2005), section II, for an intuitive illustration). Canini et al. mitigate this issue by analytically integrating out the population-level parameters, but this approach is not feasible in our case.

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.

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.

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*.

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association*, pages 1203–1215.

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.

Polson, N. G., Stroud, J. R., and Müller, P. (2008). Practical filtering with sequential parameter learning. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):413–428.