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 (CITE TEXTBOOK) to get fast latent variable posterior estimates. This can be combined with periodic MCMC for all parameters (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 type of updating which push its solution beyond the scope of our current work (see supplemental materials).

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. Note that proposals can be pre-generated before patients enter the clinic, so that only the weights need to be calculated in real time.

Using this approach, latent variable posterior mean estimates can be computed in approximately 2 seconds. These fast estimates have a correlation of 0.9950 with the estimates from running MCMC to estimate all parameters. For reference, risk estimates from two different runs of the full MCMC have a correlation of 0.9993, due to the stochastic nature of the posterior sampling. We give further details of the importance weighting, and their performance, calculations in the supplementary materials.

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

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

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. The goal of this section is to use importance weighting to estimate latent for a new subject (indexed by $n+1$) entering the study.

Our 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 propose the posterior distribution based on the first n subjects, an approach equivalent to a 1-step particle filter (CITE PARTICLE FILTERS).

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

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 $(\theta^{(j)}, b^{(j)}_{1:(n+1)})$ by $j = 1, \ldots, J$. The importance weights w_j are then proportional to

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

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 $\sum_{j=1}^{J} w^{(j)} b^{(j)}_{(n+1)}$.

2.2 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.

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.

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.