On random sampling of patient data from Bayesian models

In (Rørtveit et al., 2023), a method for estimating the random distribution of organ motion and deformation based on Bayesian theory is presented. The paper discusses two different approaches, based on different priors for the mean vector and covariance matrix of a patient distribution. The first approach uses a normal-inverse-Wishart prior, which is conjugate to the Gaussian likelihood, and therefore makes calculations relatively straightforward. The second approach uses a separate Gaussian distribution for the mean vector and an inverse-Wishart distribution for the covariance matrix. In this case, calculation of the posterior is intractable, therefore an approximation method called variational mean field estimation is used. The two methods are referred to as the NIW-model and the variational Bayes model. This technical note discusses how to sample new random data from these models.

The models assume that the data vectors $s_p$for a patient *p* is distributed according to a multivariate Gaussian distribution with mean μ_p and covariance matrix R_p.

$$s_p \sim N(\mu_p, R_p)$$

However, the parameters μ_p and R_p are not known. They can be estimated from patient-specific data, but even so, they are associated with a certain uncertainty. In the Bayesian models, this uncertainty is quantified by the *prior distribution*, which is a distribution of μ_p and R_p; in other words, μ_p and R_p are themselves stochastic variables. The prior distribution is the distribution of μ_p and R_p when nothing is known about the patient, i.e., when no data is observed. When data is observed, the uncertainty is reduced, and the distribution of μ_p and R_p changes. The uncertainty distribution after data has been observed is called the *posterior distribution*.

The paper describes how to derive the parameters of the prior from training data, and how to calculate the parameters of the posterior based on the prior parameters and patient-specific data. However, what we want is to sample new random data from a patient, not to sample these parameters.

From a Bayesian standpoint, we want to sample from the *posterior predictive distribution*, which is the best estimate of the distribution of new unknown data after some data has been measured. I.e.

$$p_{ppd}(s) = p(s_n|s_1, s_2, \dots, s_{n-1})$$

This is the same as the distribution of s with the posterior under the conditions that μ and R are distributed according to the posterior distribution, i.e.

$$s_n \sim N\big(\mu_p, R_p\big),$$

where

$$\mu_p, R_p = \mu, R|s_1, s_2, \dots, s_{n-1}.$$

We can draw from this distribution by first drawing μ_p, R_p from the posterior distribution, and then drawing s_n from the normal distribution with μ_p, R_p as parameters. For each new draw of s_n, μ_p, R_p must also be drawn anew.

In code, this would look like this:

```
[mu0_p, pcs_p, stddev_p] = normalInverseWishartPosteriorParameters(data, mu0, kappa,…
                                                                   psi_pcs, psi_stddev, nu);
for i = 1:n
        [mu, pcs_R, stddev_R] = normalInverseWishartSample(mu0_p, kappa+J, pcs_p, stddev_p, nu
        + J);
        s(:, i) = multivariateNormalSample(mu, pcs_R, stddev_R);
end
```

Here, *J* is the number of measured data points in the array "data" and *n* is the sample size. For the same using the variational Bayes model, you would replace `normalInverseWishartPosteriorParameters` and `normalInverseWishartSample` by `fastVariational` and `variationalBayesSample`.

Alternatively, we can use *point estimates* for μ_p, R_p, which was what was done in the referred paper. In this case, one would skip the first step in the loop above, and directly use `mu0_p` as the mean vector and `pcs_p` and `stddev_p` as the estimate of the covariance matrix to generate multivariate Gaussian samples from. This technique essentially ignores the uncertainty inherent in the model, but is simpler and may provide good results. The ideal value of the parameter nu might vary depending on whether one uses point estimates or samples from the posterior predictive distribution.

References

Rørtveit, Ø.L., Hysing, L.B., Stordal, A.S., Pilskog, S., 2023. An organ deformation model using Bayesian inference to combine population and patient-specific data. Phys. Med. Biol. 68, 055009. https://doi.org/10.1088/1361-6560/acb8fc