Summary Report of Bayesian Hierarchical Models Parameterization

Renat Sergazinov
Department of Statistics, Texas A&M University, College Station, TX

Abstract

In this report, I summarize the existing parameterization methodology and its applications in Bayesian hierarchical model. In particular, I discuss the centering and non-centering approaches as two competing and sometimes complementary techniques for constructing an efficient MCMC sampler. First, I highlight the general framework of centering and non-centering methods. I then give examples that are supposed to guide the judgment of when a particular methodology might be beneficial. The goal of the report is to help better understand the qualitative nature of choice involved between centering and non-centering techniques.

1 Introduction

When parameterizing a hierarchical model with a view of its later adaptation for a sampling algorithm, it is crucial to recognize that although mathematically equivalent different parameterizations can significantly affect computational costs. Theoretical results about the convergence rate and underlying assumptions of a given sampler can frequently guide the choice of the model parameterization. However, as is highlighted in [Papaspiliopoulos et al., 2007], choosing between different parameterization methodologies is as much art as it is science. In many cases, multiple diverse methods are iteratively tested for a given data at hand.

In this report, the goal is to slightly simplify the task of a researcher by offering a qualitative comparison of centering and non-centering approaches for the implementation of a Gibbs sampling algorithm. The hope is that this manuscript will help guide the appropriate choice of parameterization and highlight the inherent pitfalls of both methods. An important takeaway message from this report should be that the centering and non-centering methodologies are largely complementary. A researcher needs to make a choice between the two based on the model structure.

The paper starts with a theoretical exposition and definition of the appropriate quantities. I then demonstrate a couple of examples, the regression using the radon data set being the most prominent among them. The paper then concludes with a discussion and a summary of the results.

2 Theoretical View of Parameterization

In this section, I define a model consisting of an observed data Y and an unknown parameter of interest Θ. In the tradition of the classical Bayesian statistical inference, we would like to know the posterior distribution of $\Theta|Y$.

In many cases, however, we might be interested in a hierarchical reformulation of the above problem in terms of an introduced parameter X. The model of $P(Y|\Theta)$ is then rewritten as:

$$P(Y|\Theta) = \int P(Y, X|\Theta) d\mu(X) \tag{1}$$

$$= \int P(Y|X, \Theta) P(X|\Theta) d\mu(X), \tag{2}$$

where μ is the measure for which $P(X|\Theta)$ is defined. Such hierarchical reformulation is usually dictated by the problem at hand. It could also be introduced artificially when sampling from the posterior $\Theta|Y$ is difficult, but sampling from the joint posterior $\Theta, X|Y$ is easy. It is important to note that despite the introduction of X, we are still primarily interested in the marginal posterior $\Theta|Y$.

A parameterization of the model $(\tilde{X}, \Theta)$ is defined by a function h such that we have:

$$X = h(\tilde{X}, \Theta, Y) \tag{3}$$

In this sense, define the base augmented hierarchical model (X, Θ) to be the centered parameterization,

Preliminary work. Under review by AISTATS 2021. Do not distribute.

which is precisely summarized by Figure 1. The key in the centered parameterization is the dependence of X on the parameter Θ. On the other hand, define the non-centered parameterization $(\tilde{X}, \Theta)$ of the model as is summarized in Figure 2. In the non-centered parameterization, the dependence structure is such that the Θ and $\tilde{X}$ parameters are designed to be apriori independent.

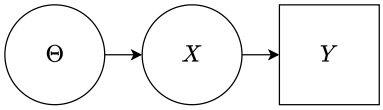

Figure 1: Diagram of centered parameterization where arrows indicate dependence.

The main difference between the centered and the non-centered parameterizations is pronounced when implementing the Gibbs sampler algorithm. The canonical Gibbs algorithm samples from the joint posterior $\Theta, X|Y$ by iteratively going through the steps:

1. Update X^{j+1} by sampling from $X|\theta^j, Y$;
2. Update Θ^{j+1} by sampling from $\Theta|X^j, Y$.

Ideally, we are looking for the case when the shape of the $\Theta|X, Y$ distribution is similar to the distribution of $\Theta|Y$. In such a case, the sampler would be highly efficient in exploring the posterior distribution of Θ.

The traditional problem of the Gibbs sampler is a serial correlation between the samples, which could undermine the quality of the posterior approximation. In the centered parameterization, X and Θ are already apriori dependent. Hence, to reduce their dependence, it must be that Y is informative about X. However, even the requirement of Y being informative about X may not be enough in some cases, as is highlighted by [Papaspiliopoulos et al., 2007]. On the other hand, in the case of the non-centered parameterization, we must require that Y is weakly informative about X. Otherwise, if Y is strongly informative about X, then $\tilde{X}$ and Θ would be aposteriori dependent.

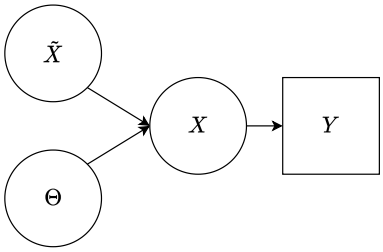

Figure 2: Diagram of non-centered parameterization where arrows indicate dependence.

Therefore, from the above analysis, we could immediately notice the complementary nature of the centering and non-centering approaches. Further discussion of the theoretical convergence rate for both methods is provided in [Papaspiliopoulos et al., 2007]. Nevertheless, it must also be noted here that centering and non-centering methods could be combined in some cases for an even more robust approach. The details of the last point are also provided in [Papaspiliopoulos et al., 2007].

3 Experiments

In this section, we demonstrate the centering and non-centering methodology on two examples. The first example is hypothetical and shows the case when the centering approach exhibits superior performance. The second example is based on the Radon data set provided in [Gelman and Hill, 2007].

3.1 Repeated Measures

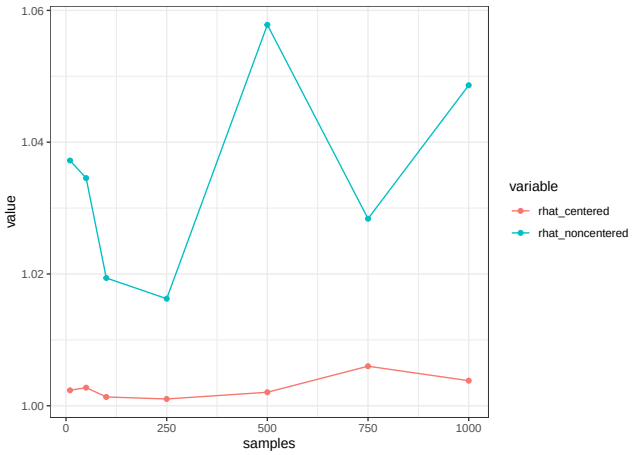

Figure 3: Maximum $\hat{R}$ across different parameters for centered and non-centered models fitted on the data of increasing sample size.

As is described in [Papaspiliopoulos et al., 2007], we consider a basic repeated measures defined hierarchically as:

$$\begin{aligned} Y_{ij} &\sim N(X_i, \sigma_y^2) \\ X_i|\Theta &\sim N(\Theta, \sigma_x^2) \\ \Theta &\propto 1, \end{aligned}$$

where we place an improper prior on Θ and assume variances (σ_x^2, σ_y^2) to be equal across groups and known. The non-centered parameterization of the

above model can then be formulated as:

$$
\begin{aligned}
Y_{ij} &\sim N(X_i, \sigma_y^2) \\
X_i &= \tilde{X}_i + \Theta \\
\tilde{X}_i &\sim N(0, \sigma_x^2) \\
\Theta &\propto 1,
\end{aligned}
$$

As is highlighted in [Papaspiliopoulos et al., 2007], for this model, the centered parameterization is preferred over the non-centered one. This fact is primarily due to the strong informativeness of Y about X. The centered parameterization of the model has $\tau = O(1/\log n)$ while the non-centered parameterization has $\tau = O(n)$, where τ is the mixing time of the MCMC indicating the time to achieve the desired accuracy. Hence, the performance of the non-centered parameterization deteriorates with the increase in the sample size, while the centered parameterization gets better.

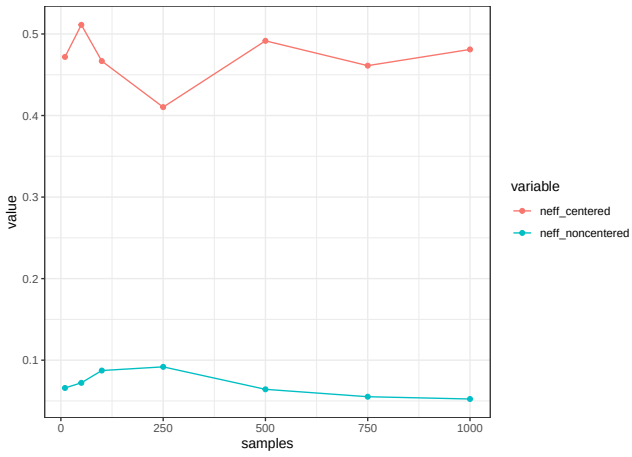

Figure 4: Maximum $\hat{R}$ across different parameters for centered and non-centered models fitted on the data of increasing sample size.

Indeed the dominance of the centered parameterization is also supported by the experimental implementation of the problem. In the practical implementation, the R package using NUT sampler is used to sample from the posterior [Stan Development Team, 2020]. The sampler is separately fitted using centered and non-centered parameterizations on the data of increasing sample size. In Figure 4, we can see that the minimum effective sample size ratio does not improve for the non-centered parameterization, while it slightly improves for the centered parameterization. The effective sample size ratio measures the amount by which autocorrelation in samples inflates the uncertainty relative to an independent sample. A higher effective sample size indicates better model performance. Also, in Figure 3, we can see that the maximum $\hat{R}$ deteriorates for the non-centered parameterization while it stays relatively constant for the centered parameterization. Potential scale reduction or $\hat{R}$ compares the distribution of the multiple Markov chains for a given parameter. Lower $\hat{R}$ values indicate better model performance, where the rule of thumb is that $\hat{R} > 1.1$ is bad.

3.2 Radon Data Regression

The radon data set taken from [Gelman and Hill, 2007] consists of the radon gas measurements taken across different counties. Additionally, the floor on which the radon concentration was measured is recorded. Radon is a poisonous gas believed to be associated with an increased risk of lung cancer. The concentration of radon depends on the type of soil as the gas primarily enters homes through the basement.

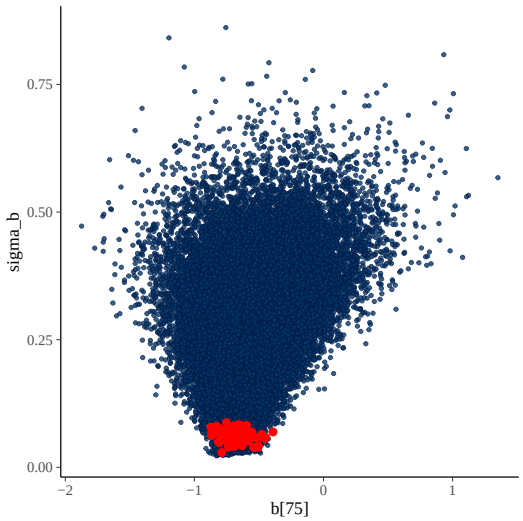

Figure 5: Scatter plot of σ_b^2 and b_{75} for the centered parameterization. The red dots indicate divergences during sampling with NUTS.

The natural (centered) parameterization for fitting a mixed model to the data looks as follows:

$$
\begin{aligned}
Y_{ij} &\sim N(a_i + b_i * X_{ij}, \sigma^2) \\
a|\mu_a, \sigma_a^2 &\sim N(\mu_a, \sigma_a^2) \\
b|\mu_b, \sigma_b^2 &\sim N(\mu_b, \sigma_b^2) \\
\mu_a &\sim N(0, 100) \\
\mu_b &\sim N(0, 100) \\
\sigma_a^2 &\sim Cauchy(0, 5) \\
\sigma_b^2 &\sim Cauchy(0, 5)
\end{aligned}
$$

where i indicates the county, Y_{ij} is the Radon mea-

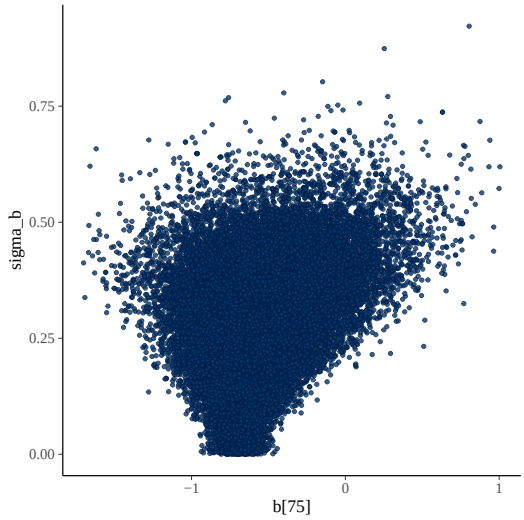

Figure 6: Scatter plot of σ_b^2 and b_{75} for the non-centered parameterization.

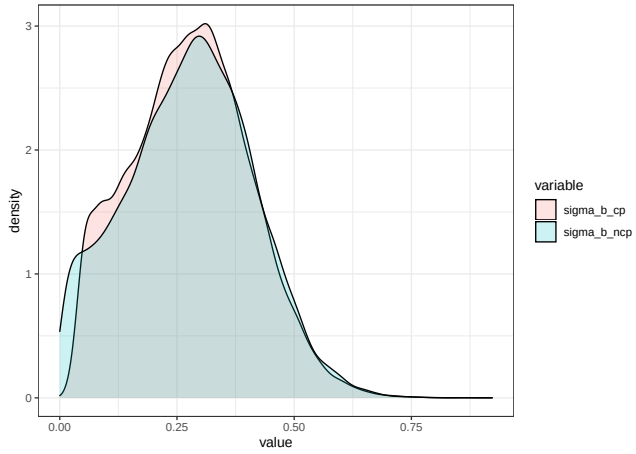

Figure 7: Scatter plot of σ_b^2 and b_{75} for the non-centered parameterization.

surement, and X_{ij} is the floor in which the concentration was measured [Thomas, 2017]. Following [Thomas, 2017], the non-centered parameterization can be written as:

$$
\begin{aligned}
Y_{ij} &\sim N(a_i + b_i * X_{ij}, \sigma^2) \\
a &= \tilde{a} * \sigma_a^2 + \mu_a \\
b &= \tilde{b} * \sigma_b^2 + \mu_b \\
\tilde{a} &\sim N(0, 1) \\
\tilde{b} &\sim N(0, 1) \\
\mu_a &\sim N(0, 100) \\
\mu_b &\sim N(0, 100) \\
\sigma_a^2 &\sim Cauchy(0, 5) \\
\sigma_b^2 &\sim Cauchy(0, 5)
\end{aligned}
$$

In these examples, one can argue that Y is only weakly informative about X and consequently regression slope b. Therefore, one might expect the non-centered parameterization to outperform the centered parameterization. This is indeed what happens in practice. We again fit both parameterizations using the R package of NUT sampler [Stan Development Team, 2020]. The scatter plot of b and σ_b^2 samples is shown in Figure 5 for the centered parameterization. As can be seen from the figure, the sampler would have a hard time exploring the posterior space of σ_b^2 for low values of b as space narrows down. This observation is indeed logical given the dependence structure of the centered method. On the other hand, looking at Figure 7, we can see that in the non-centered case, the sampler has no such problem. This is because, under the non-centered parameterization, we have $\tilde{b}$ and σ_b^2 mostly independent. The conclusion is further reinforced when we look at the posterior density of σ_b^2 for the two parameterizations in Figure 6. Figure 6 clearly illustrates that under the centered parameterizations, the sampler under-samples the low values of σ_b^2.

4 Conclusion

In this report, I have summarised and demonstrated the centered and non-centered methodologies for parameterizing a hierarchical model. In general, the two methods are quite complementary in practice. The particular choice has to be made based on the model and the available data. A simple heuristic guideline is to choose centered parameterization whenever the data is strongly informative about the parameter. On the other hand, if the data is only weakly informative, then non-centered parameterization should be used. In this report, I have also included two case studies: one in which the centered parameterization is better and one in which the non-centered parameterization should be chosen. The former example demonstrates the case when the non-centered parameterization shows a decreasing performance (longer time to achieve the desired accuracy) in the MCMC sampling scheme with the growing sample size. The latter example based on the radon data borrowed from [Gelman and Hill, 2007] shows the case when due to the weak informativeness of the data, the inherent dependence of the parameters in the centered parameterization deems the MCMC sampler inefficient.

References

[Gelman and Hill, 2007] Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical methods for social research. Cambridge University Press.

[Papaspiliopoulos et al., 2007] Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, 22(1):59–73.

[Stan Development Team, 2020] Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

[Thomas, 2017] Thomas, W. (2017). Why Hierarchical Models Are Awesome, Tricky, and Bayesian? https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/.