Bayesian Central Theorem: Bayesian Inference in the Presence of Large Number of Observational Data

Reza Madankan*[1]

[1]Department of Imaging Physics, University of Texas MD Anderson Cancer Center, Houston, TX, 77015 USA e-mail: rmadankan@mdanderson.org

Introduction

With increasing number of patient data, there is a crucial need to develop and apply more robust and computationally affordable algorithms for statistical analysis of these data sets. This manuscript concentrates on applying Bayesian Central Theorem for efficient statistical analysis and inference in presence of large data sets. Proposed technique in here, would have many applications in improving patient care through improving accuracy of applied statistical algorithms for data processing.

Problem Statement

Consider a physical process $\mathbf{x} \in \mathbb{R}^n$, which can be described in terms of a parameter $\boldsymbol{\Theta} \in \mathbb{R}^p$, using a mathematical model $\mathcal{M}$:

$$\mathbf{x} = \mathcal{M}(\boldsymbol{\Theta}) \tag{1}$$

For instance, in case of a heat diffusion process, $\mathbf{x}$ and $\boldsymbol{\Theta}$ represent the temperature map and conductivity, respectively; and the operator $\mathcal{M}$ denotes the heat diffusion process.

In clinical applications, it often happens that the mathematical model for each patient is different from the others. For instance in thermal ablation procedure, depending on the present tissue types in region of interest, different mathematical formulation is needed to describe the distribution of temperature over the tissue. Hence, given N number of patients, one can extend above equations to accurately describe the physical process of interest for each case, i.e. we will have:

$$\mathbf{x}_i = \mathcal{M}_i(\boldsymbol{\Theta}), \quad i = 1, 2, \cdots, N \tag{2}$$

On the other hand, assume that for each patient there exist some observations from a quantity of interest, denoted by $\mathbf{z}_i \in \mathbb{R}^m$, that can be acquired based on the following observation model:

$$\mathbf{z}_i = \mathcal{U}_i(\boldsymbol{\Theta}) + \nu_i, \quad i = 1, 2, \cdots, N \tag{3}$$

Note that observation data $\mathbf{z}_i$ is polluted with a white noise ν_i with covariance $\Sigma_i \in \mathbb{R}^{m \times m}$.

Given all these mathematical model predictions $\mathbf{x}_i$ and observations $\mathbf{z}_i$, our goal is to accurately estimate parameter $\boldsymbol{\Theta}$. As one can see, the major challenge over here, is the computational complexities raised due to presence of large data sets. In the following, we will discuss our technique to tackle this problem in more detail.

Bayesian Inference

To proceed with Bayesian inference framework, let's first concatenate model prediction and observation vectors as the following:

$$\mathcal{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T, \quad \mathcal{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N]^T$$

Note that $\mathcal{X} \in \mathbb{R}^{nN}$ and $\mathcal{Z} \in \mathbb{R}^{mN}$. In other words, all the models $\mathcal{M}_i$ can be concatenated into an augmented model as the following:

$$\mathcal{X} = \mathcal{M}_t(\boldsymbol{\Theta}) \tag{4}$$

and

$$\mathcal{Z} = \mathcal{U}(\boldsymbol{\Theta}) + \boldsymbol{\nu} \tag{5}$$

where, $\mathcal{M}_t(\boldsymbol{\Theta}), \mathcal{U}(\boldsymbol{\Theta}), \mathcal{Z}$, and $\boldsymbol{\nu}$ are defined as:

$$\mathcal{M}_t = [\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_N]^T, \quad \mathcal{U} = [\mathcal{U}_1, \mathcal{U}_2, \cdots, \mathcal{U}_N]^T, \quad \mathcal{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N]^T, \quad \boldsymbol{\nu} = [\nu_1, \nu_2, \cdots, \nu_N]^T$$

Now, given these observations, one can use Bayesian inference to find posterior statistics of the parameter $\boldsymbol{\Theta}$, as the following:

$$p(\boldsymbol{\Theta}|\mathcal{Z}) = \frac{p(\boldsymbol{\Theta})p(\mathcal{Z}|\boldsymbol{\Theta})}{p(\mathcal{Z})} \tag{6}$$

However, there is a major computational challenge in using the above formula for cases with large number of data observations. In detail, $p(\mathcal{Z})$ and $p(\mathcal{Z}|\boldsymbol{\Theta})$ go to zero whenever the dimension of observational data $\mathcal{Z}$ increases.

To avoid the computational complexities involved in Eq. (6), one can make use of Bayesian Central Limit Theorem (BCLT). Based on BCLT, for large number of data observations, posterior distribution of parameter $\boldsymbol{\Theta}$ can be approximated using the following normal distribution:

$$p(\boldsymbol{\Theta}|\mathcal{Z}) \simeq \mathcal{N}(\hat{\boldsymbol{\Theta}}, I_t^{-1}(\hat{\boldsymbol{\Theta}})), \quad I_t(\hat{\boldsymbol{\Theta}}) = mN.I(\hat{\boldsymbol{\Theta}}) \tag{7}$$

where, $\hat{\boldsymbol{\Theta}}$ is the *Maximum Likelihood Estimate* (MLE) of parameter $\boldsymbol{\Theta}$ and $I(\hat{\boldsymbol{\Theta}})$ denotes the Fisher information, as defined later on in Eq. (17). Note that mN in Eq. (7) corresponds with the number of available measurement data.

Derivation of Bayesian Central Limit Theorem:

To prove BCLT, we first rewrite logarithm of posterior pdf $p(\boldsymbol{\Theta}|\mathcal{Z})$ as:

$$\log p(\boldsymbol{\Theta}|\mathcal{Z}) = \log p(\boldsymbol{\Theta}) - \log p(\mathcal{Z}) + \sum_{i=1}^{mN} \log p(\mathcal{Z}_i|\boldsymbol{\Theta}) \tag{8}$$

Note that in presence of large samples (i.e. large mN), the last term dominates the above equation. Hence, $\log p(\boldsymbol{\Theta}|\mathcal{Z})$ can be approximated as:

$$\log p(\boldsymbol{\Theta}|\mathcal{Z}) \simeq \sum_{i=1}^{mN} \log p(\mathcal{Z}_i|\boldsymbol{\Theta}) \equiv l(\boldsymbol{\Theta}) \tag{9}$$

Now, by using Taylor series expansion of $l(\boldsymbol{\Theta})$ around maximum likelihood estimate of $\boldsymbol{\Theta}$, we will have:

$$l(\boldsymbol{\Theta}) \simeq l(\hat{\boldsymbol{\Theta}}) + (\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})\frac{\partial}{\partial \boldsymbol{\Theta}} l(\boldsymbol{\Theta})|_{\hat{\boldsymbol{\Theta}}} + \frac{1}{2}(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})^T(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})\frac{\partial^2 l(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta} \boldsymbol{\Theta}^T}|_{\hat{\boldsymbol{\Theta}}} \tag{10}$$

Note that $\frac{\partial l(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}|_{\hat{\boldsymbol{\Theta}}} = 0$ in above equation. Hence, Eq. (10) can be written as:

$$l(\boldsymbol{\Theta}) = \log p(\boldsymbol{\Theta}|\mathcal{Z}) \simeq l(\hat{\boldsymbol{\Theta}}) + \frac{1}{2}(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})^T(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})\frac{\partial^2 l(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta} \boldsymbol{\Theta}^T}|_{\hat{\boldsymbol{\Theta}}} \tag{11}$$

Or,

$$p(\boldsymbol{\Theta}|\mathcal{Z}) \simeq p(\hat{\boldsymbol{\Theta}}|\mathcal{Z})e^{\left(\frac{1}{2}(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})^T l''(\hat{\boldsymbol{\Theta}})(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})\right)} \tag{12}$$

where, $l''(\boldsymbol{\Theta}) = \frac{\partial^2 l(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta} \boldsymbol{\Theta}^T}|_{\hat{\boldsymbol{\Theta}}}$. On the other hand, for large value of mN, we have:

$$l''(\boldsymbol{\Theta}) = \sum_{i=1}^{mN} \frac{\partial^2 \log p(\mathcal{Z}_i|\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta} \boldsymbol{\Theta}^T}|_{\hat{\boldsymbol{\Theta}}} = \frac{mN}{mN} \sum_{i=1}^{mN} \frac{\partial^2 \log p(\mathcal{Z}_i|\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta} \boldsymbol{\Theta}^T}|_{\hat{\boldsymbol{\Theta}}}$$

$$\simeq mN.\mathcal{E}[\frac{\partial^2 \log p(\mathcal{Z}|\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta} \boldsymbol{\Theta}^T}] = -mN.I(\hat{\boldsymbol{\Theta}}) = -I_t(\hat{\boldsymbol{\Theta}}) \tag{13}$$

Hence, one can rewrite Eq. (12) as the following:

$$p(\boldsymbol{\Theta}|\mathcal{Z}) \simeq p(\hat{\boldsymbol{\Theta}}|\mathcal{Z})e^{\left(-\frac{1}{2}(\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}})^T I_t(\hat{\boldsymbol{\Theta}})(\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}})\right)} \tag{14}$$

Note that $p(\hat{\boldsymbol{\theta}}|\mathcal{Z})$ behaves as a normalization factor in above equation. $\qquad\square$

To make use of Eq. (7), one needs to evaluate MLE of $\boldsymbol{\Theta}$ and Fisher information $I(\hat{\boldsymbol{\Theta}})$. In the following, the mathematical formulation for finding each of these quantities is explained.

Calculation of MLE of $\boldsymbol{\Theta}$

From Eq. (3), we know that likelihood distribution $p(\mathcal{Z}|\boldsymbol{\Theta})$ can be written as:

$$p(\mathcal{Z}|\boldsymbol{\Theta}) = \mathcal{N}(\mathcal{U}(\boldsymbol{\Theta}), \Sigma) \tag{15}$$

where, Σ is the covariance matrix of white noise $\boldsymbol{\nu}$.

For the ease of computation, instead of maximizing the likelihood function, one can maximize the logarithm of the likelihood function. Hence, MLE can be formulated as:

$$\max_{\boldsymbol{\Theta}} \quad \log\left(p(\mathcal{Z}|\boldsymbol{\Theta})\right) = \min_{\boldsymbol{\Theta}} \quad \frac{1}{2} \left(\mathcal{Z} - \mathcal{U}(\boldsymbol{\Theta})\right)^T \Sigma^{-1} \left(\mathcal{Z} - \mathcal{U}(\boldsymbol{\Theta})\right) \tag{16}$$

Maximizing Eq. (16) results in maximum likelihood estimate of parameter $\boldsymbol{\Theta}$, denoted by $\hat{\boldsymbol{\Theta}}$. Note that in presence of a linear observation operator $\mathcal{U}$, the above cost function reduces to a quadratic function of $\boldsymbol{\Theta}$ that can be easily solved for $\hat{\boldsymbol{\Theta}}$. On the other hand, nonlinear observation operator results in a nonlinear optimization. However, since the dimension of $\boldsymbol{\Theta}$ is usually small, above optimization can be handled even in the presence of nonlinearities.

Evaluation of Fisher Information

Once we found the MLE estimate of $\boldsymbol{\Theta}$, we can proceed to calculate the Fisher information $I(\hat{\boldsymbol{\Theta}})$ using the following formula:

$$I(\hat{\boldsymbol{\Theta}}) = \mathcal{E}[\left(\frac{\partial}{\partial \boldsymbol{\Theta}} \log p(\mathcal{Z}|\boldsymbol{\Theta})\right)^2 |\hat{\boldsymbol{\Theta}}] = -\mathcal{E}[\left(\frac{\partial^2}{\partial \boldsymbol{\Theta} \boldsymbol{\Theta}^T} \log p(\mathcal{Z}|\boldsymbol{\Theta})\right) |\hat{\boldsymbol{\Theta}}] \tag{17}$$

where, $\hat{\boldsymbol{\Theta}}$ denotes the MLE of $\boldsymbol{\Theta}$, obtained from Eq. (16). Given the likelihood to be a normal distribution, one can simplify above equation as the following:

$$I(\hat{\boldsymbol{\Theta}}) = \frac{1}{4}\mathcal{E}[\left(\frac{\partial}{\partial \boldsymbol{\Theta}} \left\{(\mathcal{Z} - \mathcal{U}_t(\boldsymbol{\Theta}))^T \Sigma^{-1} \left(\mathcal{Z} - \mathcal{U}_t(\boldsymbol{\Theta})\right)\right\}\right)^2 |\hat{\boldsymbol{\Theta}}]$$

$$= \int_{\mathcal{Z}} \left(\frac{\partial}{\partial \boldsymbol{\Theta}} \left\{(\mathcal{Z} - \mathcal{U}_t(\boldsymbol{\Theta}))^T \Sigma^{-1} \left(\mathcal{Z} - \mathcal{U}_t(\boldsymbol{\Theta})\right)\right\}\right)^2 p(\mathcal{Z}|\boldsymbol{\Theta})d\mathcal{Z} \bigg|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} \tag{18}$$

Hence, using MLE of $\boldsymbol{\Theta}$, one can evaluate Fisher information $I(\hat{\boldsymbol{\Theta}})$ using Eq. (18). However, the integration involved in Eq. (18) can be computationally expensive to evaluate, depending on the nonlinearities involved in observation operator $\mathcal{U}$.

Note that in presence of a general $p-$dimensional parameter $\boldsymbol{\Theta}$, Fisher information $I(\hat{\boldsymbol{\Theta}})$ will be a $p \times p$ matrix, where each of its elements can be calculated as follows:

$$I_{i,j}(\hat{\boldsymbol{\Theta}}) = \int_{\mathcal{Z}} \left(\frac{\partial}{\partial \boldsymbol{\Theta}_i} \left\{(\mathcal{Z} - \mathcal{U}(\boldsymbol{\Theta}))^T \Sigma^{-1} \left(\mathcal{Z} - \mathcal{U}(\boldsymbol{\Theta})\right)\right\}\right)\left(\frac{\partial}{\partial \boldsymbol{\Theta}_j} \left\{(\mathcal{Z} - \mathcal{U}(\boldsymbol{\Theta}))^T \Sigma^{-1} \left(\mathcal{Z} - \mathcal{U}(\boldsymbol{\Theta})\right)\right\}\right) p(\mathcal{Z}|\boldsymbol{\Theta})d\mathcal{Z} \bigg|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} \tag{19}$$

Now, using matrix algebra one can rewrite $I_{ij}(\hat{\Theta})$ as follows:

$$I_{ij}(\hat{\Theta}) = \mathcal{U}_{\Theta_i}^T \Sigma^{-1}(M + mm^T) + Tr(\Sigma^{-1}\mathcal{U}_{\Theta_i}\mathcal{U}_{\Theta_j}^T \Sigma^{-1}M) + m^T(\Sigma^{-1}\mathcal{U}_{\Theta_i}\mathcal{U}_{\Theta_j}^T \Sigma^{-1})m$$

$$+ m^T \Sigma^{-1}\mathcal{U}_{\Theta_i} \frac{\partial}{\partial\Theta_j}(\mathcal{U}^T \Sigma^{-1}\mathcal{U}) + \mathcal{U}_{\Theta_i}^T \Sigma^{-1}(M + mm^T)\Sigma^{-1}\mathcal{U}_{\Theta_j}$$

$$+ \mathcal{U}_{\Theta_i}^T \Sigma^{-1}(M + mm^T)\Sigma^{-1}\mathcal{U}_{\Theta_j} + \mathcal{U}_{\Theta_i}^T \Sigma^{-1}m \frac{\partial}{\partial\Theta_j}(\mathcal{U}^T \Sigma^{-1}\mathcal{U})$$

$$+ \frac{\partial}{\partial\Theta_i}(\mathcal{U}^T \Sigma^{-1}\mathcal{U})m^T \Sigma^{-1}\mathcal{U}_{\Theta_j} + \frac{\partial}{\partial\Theta_i}(\mathcal{U}^T \Sigma^{-1}\mathcal{U})\mathcal{U}_{\Theta_j}^T \Sigma^{-1}m$$

$$+ \frac{\partial}{\partial\Theta_i}(\mathcal{U}^T \Sigma^{-1}\mathcal{U})\frac{\partial}{\partial\Theta_j}(\mathcal{U}^T \Sigma^{-1}\mathcal{U}) \qquad (20)$$

where, m and $M = \Sigma$ denote conditional mean and conditional covariance of observation $\mathcal{Z}$ in above equation, respectively. Note that Eq. (20) needs to be evaluated at $\Theta = \hat{\Theta}$.

A Simple Example

The effect of MrgLITT operation on human brain is simulated retrospectively. Fig. 1 illustrates the position of the tumor in the brain.
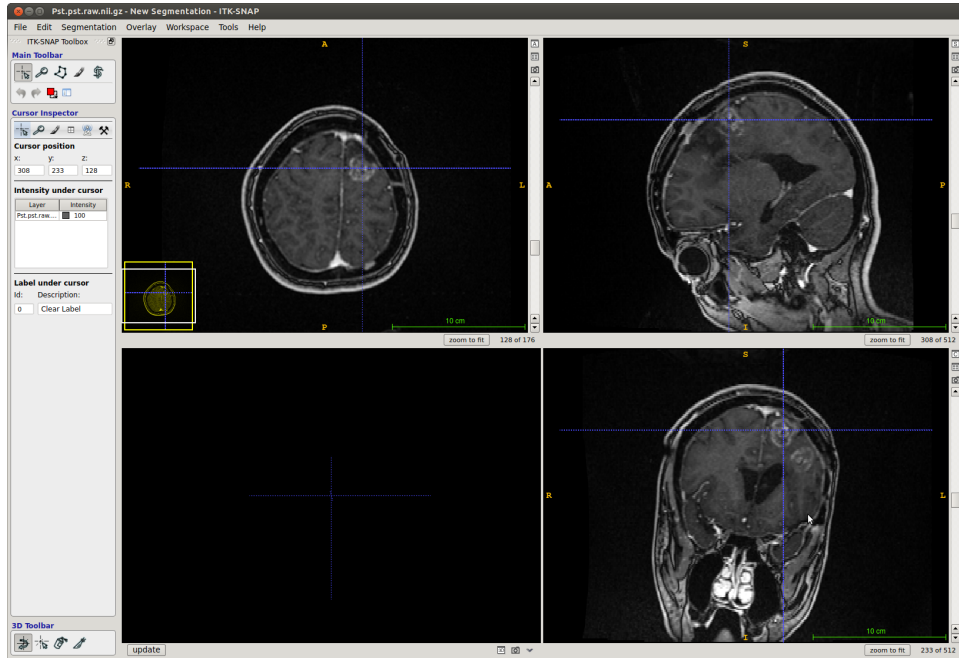

Figure 1: Position of the tumor in human brain.

As described before, Bayesian central limit theorem is used to find the optimal values of optical attenuation coefficients by comparing the temperature predictions and temperature measurements. Optimal values of optical attenuation coefficients are then utilized in Pennes bioheat equation to reconstruct the temperature map.

Fig. 2 represents the temperature map at a particular slice, given by direct measurement and temperature reconstruction through the Pennes bioheat equation. Similarly, reconstructed temperature map at another slice is being compared with direct temperature measurements in Fig. 3.
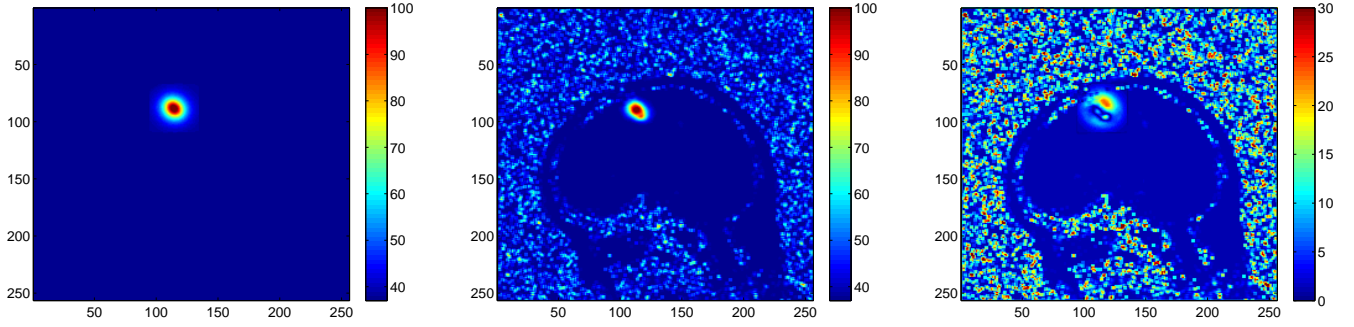
Figure 2: Temperature map over slice 1 a) Reconstructed temperature, b) observed temperature, c) the error between the reconstruction and measurement.
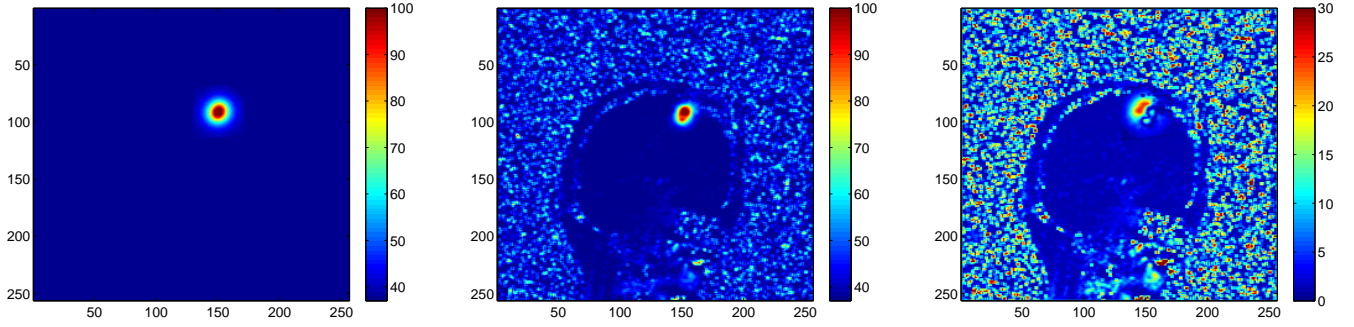

Figure 3: Temperature map over slice 2 a) Reconstructed temperature, b) observed temperature, c) the error between the reconstruction and measurement.

Treatment Planning and Connection with Tissue Segmentation

Clearly, accuracy of reconstructed temperature map crucially depends on precision of tissue segmentation in hand. Hence, to assess reliable temperature reconstruction, one needs to have accurate tissue segmentation. This can be achieved by using the optimal tissue segmentation algorithm. Fig. 4 illustrates the treatment planning for MRgLITT implementation that combines both BCLT and Optimal Tissue Segmentation to achieve a reliable estimation of temperature map over the tissue.

Conclusion

In this manuscript, we provided new mathematical formulations to address the problem of parameter estimation in presence of large number of data observations. The key point of these formulations is to make use of Bayesian Central Limit Theorem, which approximates posterior distribution of parameter of interest with a Gaussian distribution, whose mean and covariance are provided using maximum likelihood estimate of the parameter and Fisher information, respectively.

It should be noted here that the only possible challenge in using these formulations is to evaluate the Fisher information. As we discussed before, finding Fisher information can be challenging in general, depending on the nonlinearities involved in observation operator $\mathcal{U}(\boldsymbol{\Theta})$.
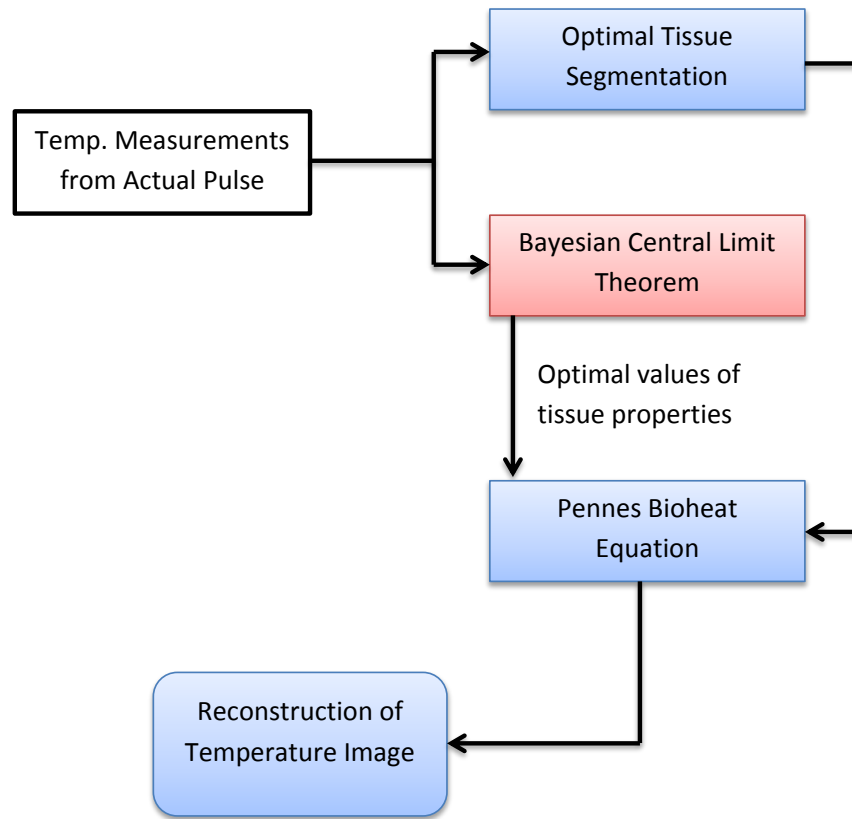
Figure 4: Implementation of MRgLITT by combining BCLT and Optimal Tissue Segmentation Algorithm.