Likelihood-Free Inference in GLMMs via Divergences

APTS week 2

David Huk

Introduction

1. From mixed effects models to GLMMs
2. Traditional GLMM estimation
3. Divergences
4. Proposed approach
5. Simulation study

Linear Mixed Effects - Model (1)

Used to model measurements of correlated observations:

- Repeated measurements of different patients.
- Data coming from multiple clusters.

Have a model component for population-wide effects and a second component for within-cluster effects.

Linear Mixed Effects - Model (2)

Consider $\boldsymbol{y} = (y_1, \ldots, y_n)$ as coming from $Y \sim N(\mu, \Sigma)$ where

$$\mu = X\boldsymbol{\beta} + Z\boldsymbol{u}, \quad \boldsymbol{u} \sim N(0, \Sigma_u), \tag{1}$$

In matrix form:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon} \tag{2}$$

- $\boldsymbol{y}$ are observations, with mean $E(\boldsymbol{y}) = X\boldsymbol{\beta}$.
- $\boldsymbol{\beta}$ is an unknown vector of fixed effects.
- $\boldsymbol{u}$ is an unknown vector of random effects, with mean $\mathbb{E}[\boldsymbol{u}] = \mathbf{0}$ and covariance matrix $\mathbb{V}(\boldsymbol{u}) = \Sigma_u$.
- $\boldsymbol{\epsilon}$ is an unknown vector of random errors, with mean $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and variance usually $\mathbb{V}(\boldsymbol{\epsilon}) = I_n$.
- X and Z are known design matrices.

Linear Mixed Effects - Inference (1)

We want to estimate $\boldsymbol{\beta}, \Sigma$ and Σ_u as well as (predict) $\boldsymbol{u}$.
Method 1: MLE

$$f\left(y \mid \boldsymbol{\beta}, \Sigma, \Sigma_u\right) = \int f\left(y \mid \boldsymbol{u}, \boldsymbol{\beta}, \Sigma, \Sigma_u\right) f\left(\boldsymbol{u} \mid \Sigma, \Sigma_u\right) d\boldsymbol{u}$$

$$\propto |V|^{-1/2} \exp\left(-\frac{1}{2}(y - X\boldsymbol{\beta})^T V^{-1}(y - X\boldsymbol{\beta})\right) \tag{3}$$

where $V = \Sigma + Z\Sigma_u Z^T$.
This can be optimised numerically, iterating steps for optimising (a) $\boldsymbol{\beta}, \Sigma$ given fixed Σ_u and (b) Σ_u given $\hat{\boldsymbol{\beta}}(\Sigma_u), \hat{\Sigma}(\Sigma_u)$.

Treating $\boldsymbol{\beta}$ as fixed quantities when finding the variance components introduces biases leading to lower variances.

Linear Mixed Effects - Inference (2)

Method 2: REML (REstricted Maximum Likelihood)
Estimate Σ, Σ_u by maximising the restricted likelihood (4) first and using it to recover estimates $\hat{\beta}$ in (5).

$$f\left((I_n - H)y \mid \Sigma, \Sigma_u\right) \propto f\left(y \mid \hat{\beta}, \Sigma, \Sigma_u\right) \left|X^T V X\right|^{1/2} \tag{4}$$

$$\hat{\beta} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} y \tag{5}$$

While this has no bias, its estimates cannot be used to compare models with different design matrices.

Linear Mixed Effects - Inference (3)

Can also do Bayesian inference using Gibbs with following conditional posteriors:

$$
\begin{aligned}
f(\beta \mid y, \text{ rest }) &\propto \mathcal{N}(y - Z\boldsymbol{u}; X\beta, V) f(\beta) \\
f(\boldsymbol{u} \mid y, \text{ rest }) &\propto \mathcal{N}(y - X\beta; Z\boldsymbol{u}, V) \mathcal{N}(\boldsymbol{u}; 0, \Sigma_u) \\
f(\Sigma \mid y, \text{ rest }) &\propto \mathcal{N}(y - X\beta - Z\boldsymbol{u}; 0, V) f(\Sigma) \\
f(\Sigma_u \mid y, \text{ rest }) &\propto \mathcal{N}(\boldsymbol{u}; 0, \Sigma_u) f(\Sigma_u)
\end{aligned}
$$

Linear Mixed Effects - Inference of $\mathbf{u}$

To get values for mixed effect $\boldsymbol{u}$, one resorts to predicting their level given known information:

$$\hat{\boldsymbol{u}} = \mathbb{E}[\boldsymbol{u} \mid y, \hat{\beta}, \hat{\Sigma}, \hat{\Sigma}_u] = (Z^T \Sigma^{-1} Z + \Sigma_u^{-1})^{-1} Z^T \Sigma^{-1} (y - X\beta) \qquad (6)$$

Generalised Linear Mixed Models

Generalise LMM with the common GLM framework:

$$Y_i \sim EDF\left(\cdot \mid \mu_i, \sigma^2\right)$$

$$g(\mu) \equiv \begin{pmatrix} g\left(\mu_1\right) \\ \vdots \\ g\left(\mu_n\right) \end{pmatrix} = X\beta + Z\boldsymbol{u}, \quad \boldsymbol{u} \sim N\left(0, \Sigma_u\right) \tag{7}$$

where EDF is an exponential family distribution with $\mathbb{E}(Y) = \mu$ and $\mathbb{V}(Y) = \frac{\sigma^2 \cdot V(\mu)}{m}$ for given m. A common assumption is $\sigma^2 = 1$.

Generalised Linear Mixed Models - Inference

Want to infer mean μ and covariance Σ_u. But the likelihood usually does not have a closed-form expression.

$$\begin{aligned} f(y \mid \beta, \Sigma_u) &= \int f(y \mid \beta, \boldsymbol{u}, \Sigma_u) f(\boldsymbol{u} \mid \beta, \Sigma_u) \, d\boldsymbol{u} \\ &= \int f(y \mid \beta, \boldsymbol{u}) f(\boldsymbol{u} \mid \Sigma_u) \, d\boldsymbol{u} \\ &= \int \prod_{i=1}^{n} f\left(y_i \mid g^{-1}\left([X\beta + Z\boldsymbol{u}]_i\right)\right) f(\boldsymbol{u} \mid \Sigma_u) \, d\boldsymbol{u} \end{aligned} \tag{8}$$

Generalised Linear Mixed Models - Traditional Estimation

- When the integral over $\boldsymbol{u}$ are of low dimension, or when a simplification is available, **Gaussian quadrature** approximations are used. For integrals of multiple dimensions, this can be applied recursively, but performance worsens with increases in dimension.
- An alternative is using a **Laplace approximation** which approximate the integrand by an un-normalised multivariate Gaussian.
- One can also resort to **penalised quasi-likelihood** (PQL) which is fast but often inaccurate.
- Finally, we can also perform Bayesian inference with **Gibbs sampling**.
- Research is still ongoing for inference of GLMMs.

Divergences - what are they?

Assume we observe $\mathbf{y} \sim \mathcal{P}^*$ and we have a model with parameters θ giving rise to a density $\mathcal{P}^\theta$ (explicit or not) from which we can sample.
A divergence $D(.||.)$ is then defined as a function of two distributions such that:

- $D(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) \geq 0$
- $D(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) = 0 \iff \mathcal{P}^* = \mathcal{P}^{\boldsymbol{\theta}}$

By minimising the divergence between $\mathcal{P}^*$ and $\mathcal{P}^\theta$ as a function of θ, one can recover the best possible model for the data $\mathbf{y}$.

Scoring Rules as Divergences

Scoring rule $S(\mathcal{P}^\theta, Y)$ is a function between a distribution $\mathcal{P}^\theta$ and random variable $Y \sim \mathcal{P}^*$.

- *expected scoring rule* is defined as $S(\mathcal{P}^\theta, \mathbf{y}) := \mathbb{E}_{Y \sim \mathcal{P}^*} S(\mathcal{P}^{\boldsymbol{\theta}}, Y)$
- *proper* SR (w.r.t. $\mathbf{P}$) if
  $S\left(\mathcal{P}^*, \mathcal{P}^*\right) \leq S\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^*\right) \ \forall\, \mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^* \in \mathbf{P}$.
- *strictly proper* if
  $S\left(\mathcal{P}^*, \mathcal{P}^*\right) < S\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^*\right) \ \forall\, \mathcal{P}^{\boldsymbol{\theta}}, \mathcal{P}^* \in \mathbf{P}$ s.t. $\mathcal{P}^* \neq \mathcal{P}^{\boldsymbol{\theta}}$

Scoring Rules as Divergences - Energy Score

By considering the quantity $D_{SR}(\mathcal{P}^*||\mathcal{P}^{\boldsymbol{\theta}}) := S\left(\mathcal{P}^*, \mathcal{P}^{\boldsymbol{\theta}}\right) - S\left(\mathcal{P}^*, \mathcal{P}^*\right)$ for a strictly proper SR, one can see that it defines a divergence.
The SR we will use is the Energy Score:

$$S_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{y}\right) = 2 \cdot \mathbb{E}_{\mathbf{X}' \sim \mathcal{P}^{\boldsymbol{\theta}}} \|\mathbf{X}' - \mathbf{y}\|_2^{\beta} - \mathbb{E}_{\mathbf{X}'_1, \mathbf{X}'_2 \sim \mathcal{P}^{\boldsymbol{\theta}}} \left\|\mathbf{X}'_1 - \mathbf{X}'_2\right\|_2^{\beta} \tag{9}$$

$$\hat{S}_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{y}\right) = \frac{2}{m} \sum_{j=1}^{m} \|\mathbf{x}_j - \mathbf{y}\|_2^{\beta} - \frac{1}{m(m-1)} \sum_{\substack{k=1 \\ k \neq j}}^{m} \|\mathbf{x}_j - \mathbf{x}_k\|_2^{\beta} \tag{10}$$

Solve $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \hat{S}_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{y}\right)$ since the first part of D_{SR} is constant in $\boldsymbol{\theta}$.

Proposed approach

As a GLM induces an EDF model, we can sample from it by conditioning on predictors and mixed effects with chosen parameters θ. We can then compare observations against simulations and compute $\hat{S}_{\mathrm{E}}\left(\mathcal{P}^{\boldsymbol{\theta}}, \mathbf{y}\right)$, choosing θ which minise it. For a given parameter vector $\theta = (\beta, \Sigma_u)$:

1. Sample random effects $\mathbf{u} \sim \mathcal{N}(0, \Sigma_u)$.
2. Compute $g(\mu) = X\beta + Z\mathbf{u}$.
3. Sample $\mathbf{y}' = (\mathbf{y}'_1, \ldots, \mathbf{y}'_m) \sim EDF(\mathbf{y}'|\mu, I_n)$ iid.
4. Compute $\hat{S}_{\mathrm{E}}\left(\mathbf{y}', \mathbf{y}\right)$ by comparing simulations $\mathbf{y}'$ to observations $\mathbf{y}$.

By performing this routine in an optimising algorithm, we can infer the correct values for θ.

Simulation Study

Consider the true model $\mathcal{P}^*$ where I have 100 observations in $\{0, 1\}$ coming from 3 distinct clusters as:

$$Y_{ij} \sim \text{Bernoulli}(\mu_{ij}), \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \beta_0 + b_{0j} + \beta_1 x_i, \quad b_{0i} \sim N(0, \sigma_b^2) \quad (11)$$

with

$$\theta = (\beta_0, \beta_1, \sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 2, 1, 2, 3)$$

Observations are given clusters as 1 if $i \in [1 - 30]$, 2 if $i \in [31 - 60]$ and 3 else. We simulate predictors $x_i \sim \mathcal{N}(0, 1)$ iid.

Simulation Study - Results

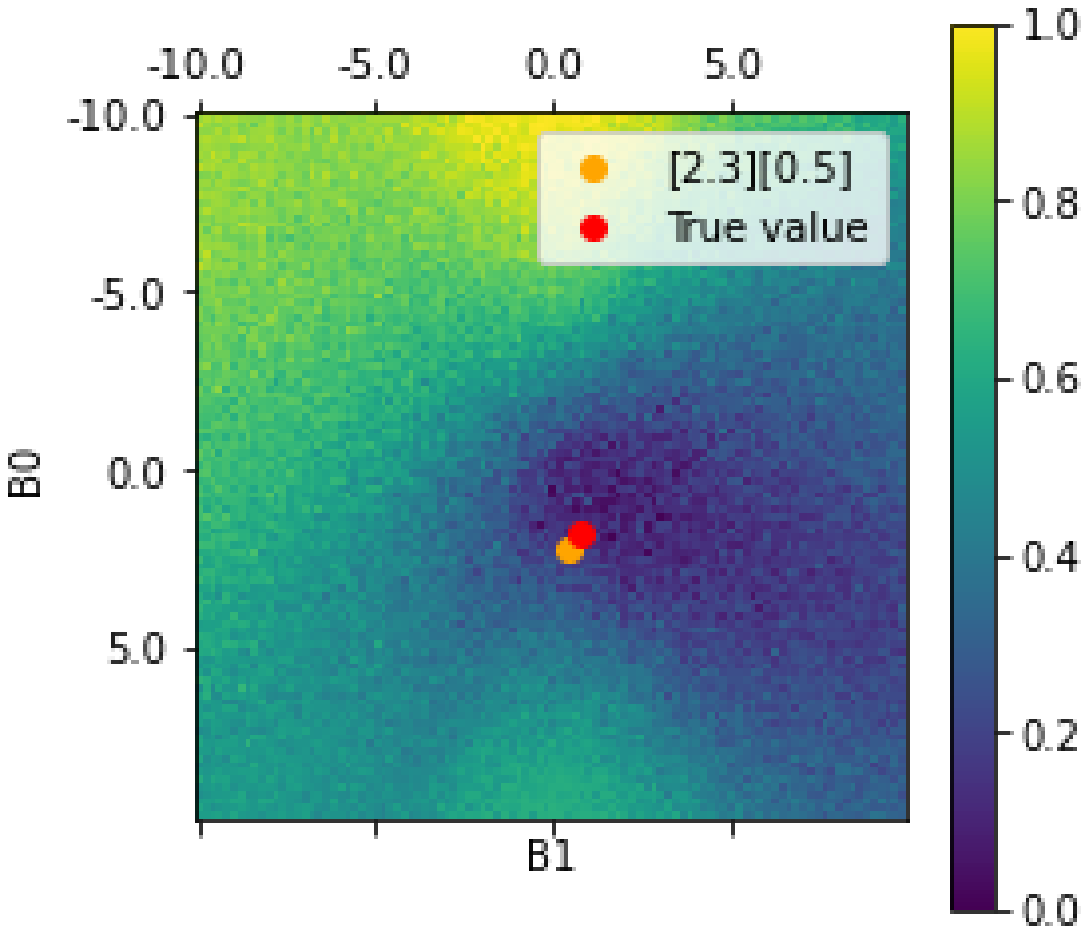

Figure: Experiment with $\theta = (1, 2, 1, 2, 3)$, showing β_0, β_1.

Simulation Study - Results

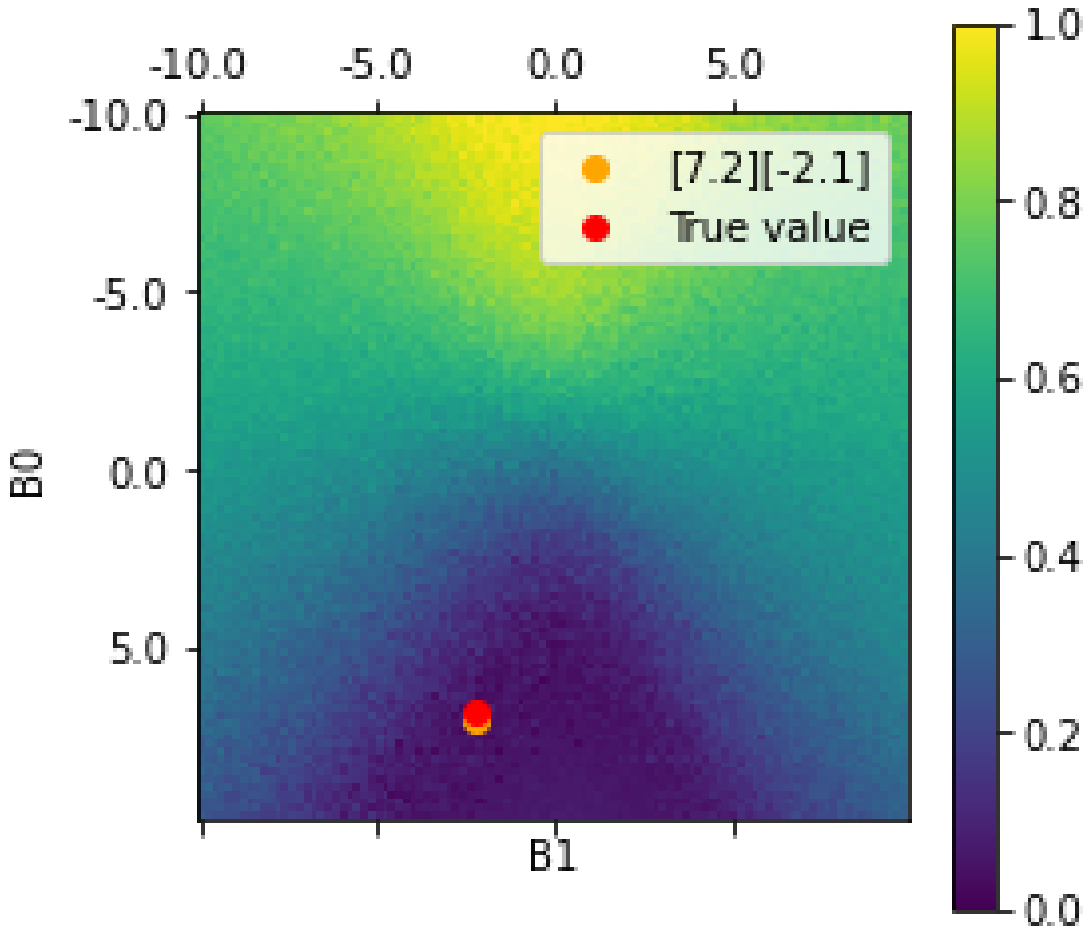

Figure: Experiment with $\theta = (7, -2, 1, 2, 3)$, showing β_0, β_1.

Simulation Study - Results

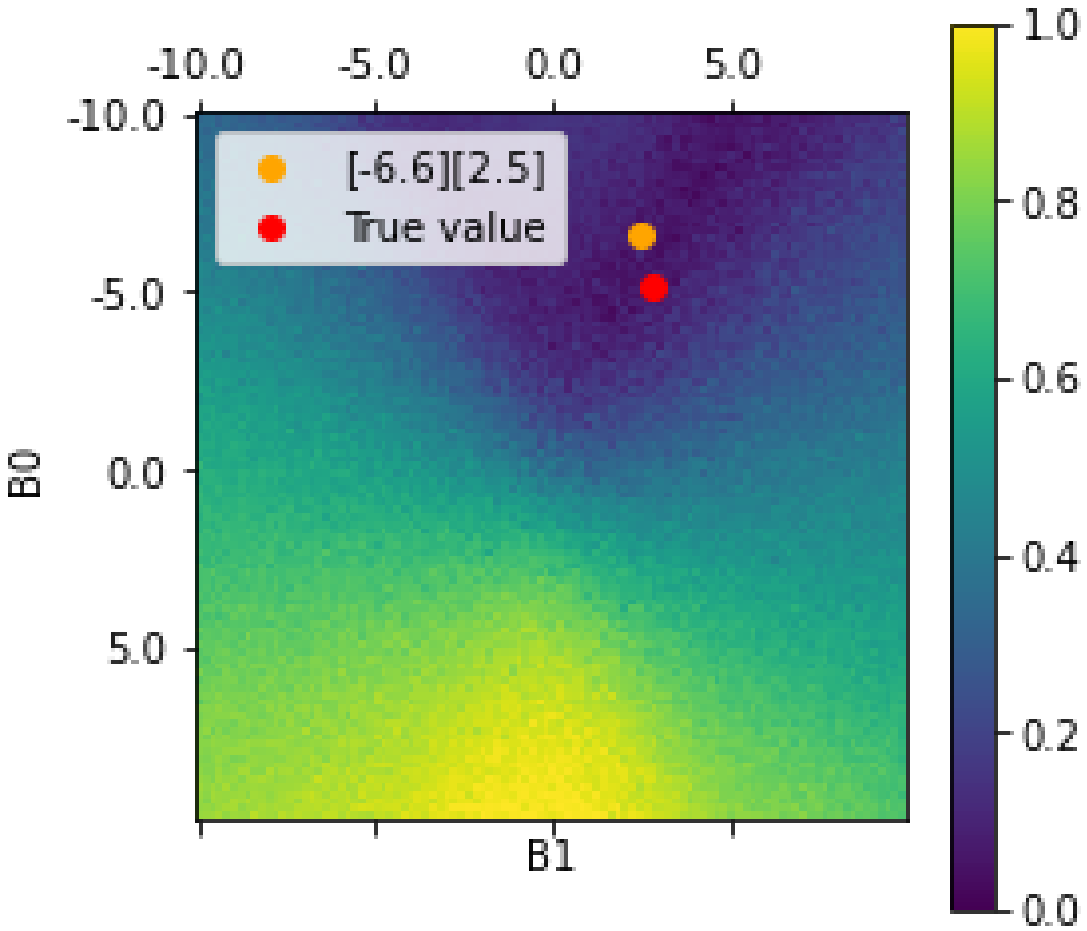

Figure: Experiment with $\theta = (-5, 3, 1, 2, 3)$, showing β_0, β_1.

Simulation Study - Results

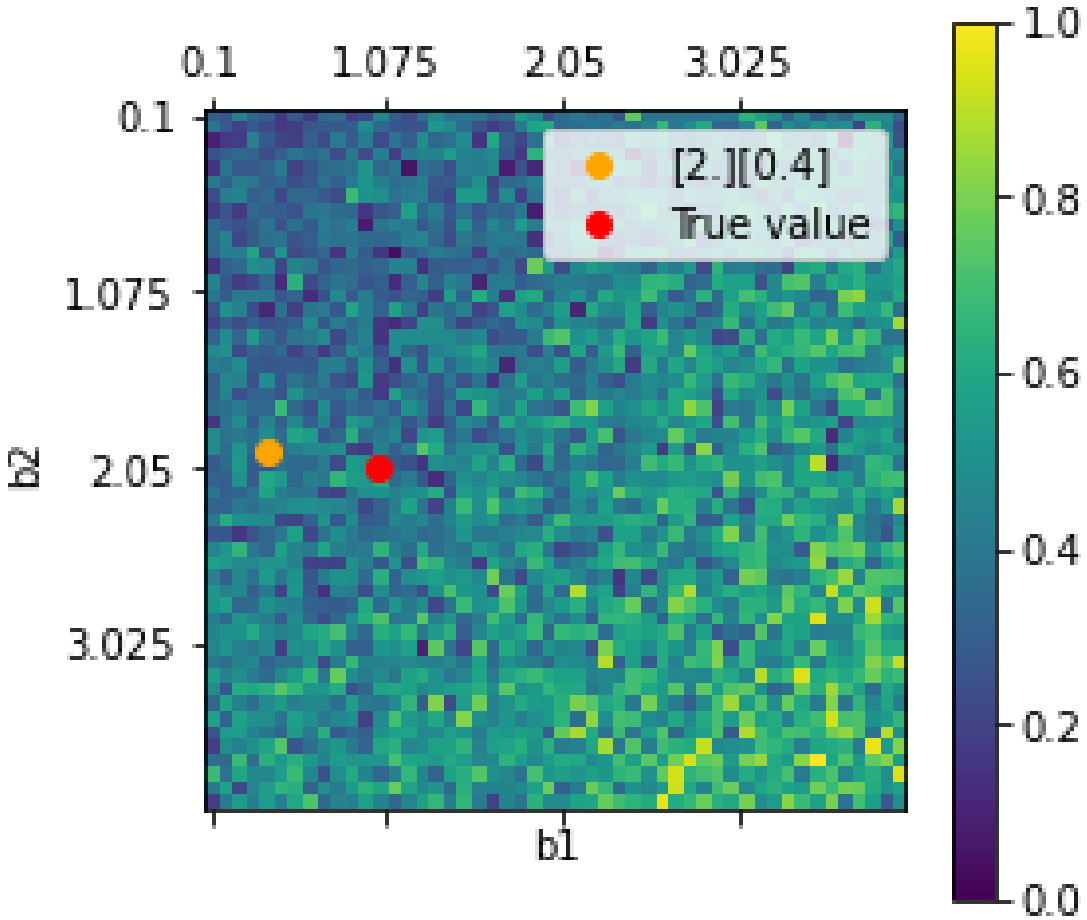

Figure: Experiment with $\theta = (1, 2, 1, 2, 3)$, showing b_1, b_2.

Simulation Study - Results

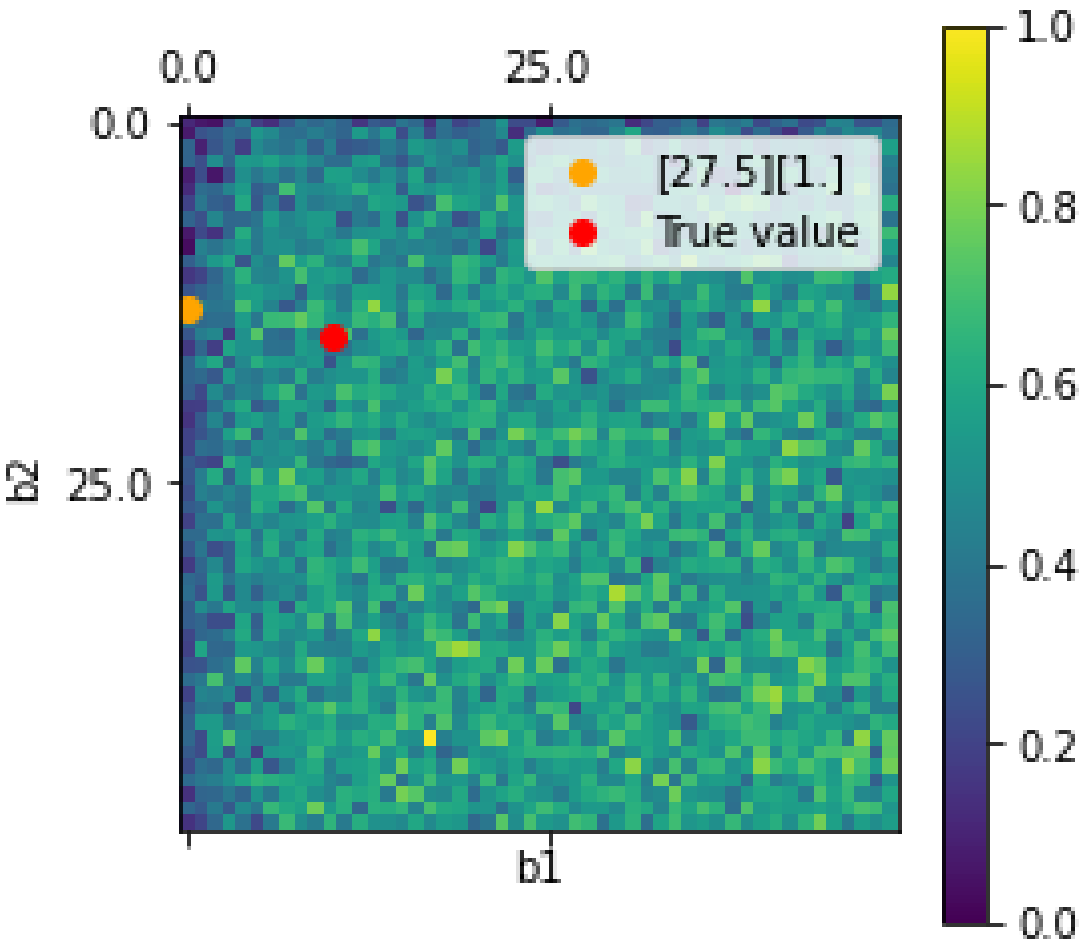

Figure: Experiment with $\theta = (1, 2, 20, 30, 3)$, showing b_1, b_2.

Future Work - Bayesian inference for GLMMs with ABC

Idea: Use an ABC framework in order to infer parameters in a GLMM. We can put priors on the unknown terms, which then allows us to sample from them. Then, for a given sample, we can compare it to observations by SR, yielding a score for that sample. We can do this repeatedly, and retain the best q% of such samples based on the score. This will then approximate the joint posterior of the parameters.

That is, doing ABC inference on GLMM parameters where the discrepancy between observations and samples is measured with the SR instead of a statistic or other.

Conclusion

- We reviewed LMM and their GLMM extension.
- We explained how inference is classically made for those models.
- We introduced Divergences and Scoring Rules.
- We explained our approach for LFI in GLMMs.
- We exemplified our approach on a simulated problem.
- We presented an idea to extend this inference to the Bayesian setting.

- Thank you for your attention!