Parameter estimation of Poisson generalized linear mixed models based on three different statistical principles: a Simulation Study

Martí Casals

Co-authors: Klaus Langohr, Josep Lluís Carrasco and Lars Rönnegård

Primeras Jornadas Científicas de Estudiantes de la SEB
Valencia

20th January, 2015

Introduction

- **Complex designs** where data is **hierarchically structured**
- Different meanings of the 'term hierarchical model':
  - To account for **clustering**
  - Different hierarchical levels (**multi-level analysis**)
  - **Bayesian framework** (multiple layers of data or prior information)
- Simple hierarchical models are relatively common (**random intercept models**)
- Applications in ecology, genetics, medicine, psychology, sports science...

Introduction to GLMM

GLMMs: a statistical modelling framework incorporating:

- **Linear combinations** of categorical and continuous predictors, and interactions
- Response distributions in the **exponential family** (binomial, Poisson, and extensions)
- Any smooth, monotonic **link function** (e.g. logistic, exponential models)
- Flexible combinations of **blocking factors** (clustering; random effects)
- The main difficulty of GLMMs is the **parameter estimation**; not viable an analytic solution that allows maximizing the marginal likelihood of data.

Definition of GLMM

- Basic idea: Just add random effects on the linear scale
- A function $g(\cdot)$ known as the link function and a linear predictor η as follows:

$$\eta = g(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i,\ i = 1, \ldots, n,$$

  - $\mathbf{X}_i$ and $\mathbf{Z}_i$: design matrices associated fixed and random effects
  - $\boldsymbol{\beta}$: fixed effects covariate vector,
  - $\mathbf{u}_i$: random effect vector,
  - Random effects: follow a multivariate Gaussian distribution with mean vector $\mathbf{0}$ and unknown positive definite covariance matrix $\boldsymbol{\Sigma}$.
- Tipically, a normal distribution is assumed for the random effects.

Problem: The likelihood

The conditional density of $\mathbf{Y}$ given $\mathbf{u}$ has the form:

$$f(\mathbf{Y}|\mathbf{u};\beta) = \prod_{i=1}^{n} f(\mathbf{Y}_i|\mathbf{u}_i;\beta).$$

We have to evaluate an integral of the form:

$$l(\beta,\mathbf{u}|Y) = f(Y;\mathbf{u},\beta) = \int f(Y|\mathbf{u};\beta) f(\mathbf{u};\Sigma)\, du$$

Different estimation methods based on approximation (Laplace,GHQ,PQL,...) or simulation have been developed in recent years.

Objective

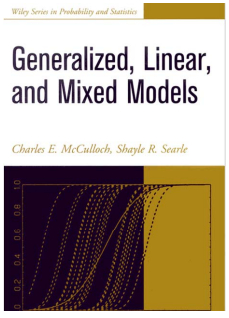
- The purpose of this study is to compare the performance of three different statistical principles –**Marginal likelihood, Extended likelihood, Bayesian approach**– via a simulation study with different scenarios of overdispersion.
- We want to highlight Gauss-Hermite quadrature **(GHQ)** estimation, hierarchical **(h-likelihood)**, and Bayesian methods (Integrated nested Laplace approximation **(INLA)**)

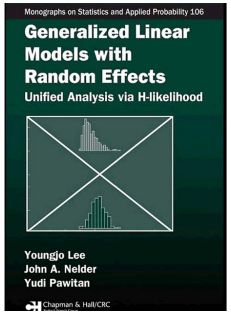
Principles, Methods and Algorithms

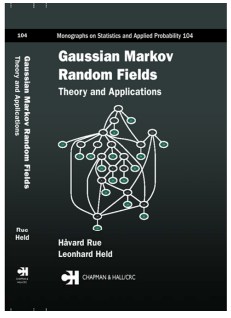
Table: Overview of statistical principles

Principle	**Method**	**Algorithms**
Marginal Likelihood	Maximum likelihood	Newton-Raphson, Fisher scoring, Penalized iteratively reweighted least squares, Adaptative Gauss Hermite Quadrature
Extended likelihood	h-likelihood	Newton-Raphson, Iterative weighted least squares
Bayesian	Posterior mean	MCMC, Integrated Nested Laplace Approximations

Books on likelihood principles







McCulloch, C. E., Searle, S. R., 2001. Generalized, Linear and Mixed Models. John Wiley & Sons, New York.

Lee, Y., Nelder, J., Pawitan, Y., 2006. Generalized Linear Models with Random Eects: Unied Analysis via H-likelihood. Chapman & Hall/CRC, Boca Raton.

Rue, Held. Gaussian Markov Random Fields. Theory and Applications. Boca Raton: Chapman & Hall 2005

Likelihood Principles

Classical Inference

- Uses marginal likelihood (random effects integrated out)
- Included fixed parameters and only observations are treated as random

Likelihood Principles

Classical Inference

- Uses marginal likelihood (random effects integrated out)
- Included fixed parameters and only observations are treated as random

Extended likelihood inference

- All information in the data about the random and fixed effects is included in a joint likelihood.
- Includes: fixed parameters, unobserved random effects, and observations as random.

Likelihood Principles

Classical Inference

- Uses marginal likelihood (random effects integrated out)
- Included fixed parameters and only observations are treated as random

Extended likelihood inference

- All information in the data about the random and fixed effects is included in a joint likelihood.
- Includes: fixed parameters, unobserved random effects, and observations as random.

Bayesian inference

- Probabilistic framework combines likelihood and prior information. All parameters and observations as random.

Statistical software packages in R

For Likelihood Principle

`lme4`: multiple/crossed; fast; Under active development. Laplace and GHQ.

Others: `glmmML`, `MASS (glmmPQL)`, `repeat`.

For Extended Likelihood Principle

`hglm`: GLMs with random effects (based on the h-likelihood). Hierarchical GLMs.

Others: `hglm and HGLMM`.

For Bayesian Principle

`R-INLA`: recent; Specialized; spatial and temporal correlation.

Others: `glmmBUGS`, `glmmAK`, `glmmADMB`,`MCMCglmm` and `R-INLA`

Integrated nested Laplace approximation (INLA)

Three main ingredients in INLA:

- Gaussian Markov random fields
- Latent Gaussian models (LGMs)
- Laplace approximations

which together (with a few other things) give a very nice tool for Bayesian inference:

- quick
- accurate

INLA is an alternative to MCMC

- much, much faster
- R-INLA makes coding very easy
- allows non-experts to fit complex models
- suitable for a specific class of models, LGMs (GLM, GAM, spline models, semi-parametric regression, spatial models, log Gaussian Cox processes, frailty models...)
- information webpage: `http://www.r-inla.org/`

Folk Wrestling data I

- Leonese Wrestling (LW) or Aluche is a traditional and popular sport of the province of Leon.

Folk Wrestling data I

- Leonese Wrestling (LW) or Aluche is a traditional and popular sport of the province of Leon.
- The main variable of interest: **Incidence of injury**

Figure: Historic photo of Leonese Wrestling (From: Vicente Martin)

Folk Wrestling data II

- There is not much information on the frequency of injuries, their incidence and their causes as a previous stage to carry out prevention and control programs in this sport.

Figure: A photo of Leonese Wrestling (From: Vicente Martin)

Real Data Example: Folk Wrestling data

Subjects

- The cohort consists of a sample of 213 wrestlers during the summer seasons 2005-2010.

Real Data Example: Folk Wrestling data

Subjects

- The cohort consists of a sample of 213 wrestlers during the summer seasons 2005-2010.

Design

- An unbalanced design with repeated measures.

Real Data Example: Folk Wrestling data

Subjects

- The cohort consists of a sample of 213 wrestlers during the summer seasons 2005-2010.

Design

- An unbalanced design with repeated measures.

Response Variable

- Frequency of total injuries per combat, follows a Poisson distribution.

Real Data Example: Folk Wrestling data

Subjects

- The cohort consists of a sample of 213 wrestlers during the summer seasons 2005-2010.

Design

- An unbalanced design with repeated measures.

Response Variable

- Frequency of total injuries per combat, follows a Poisson distribution.

Risk covariates considered

- **Winner**: More falls in favor had against (yes/no)
- **Weight Category**: Light, Medium (reference), Semi-heavy, Heavy

The model under study I

Poisson generalized linear mixed model:

$$\log(\boldsymbol{\mu}_i) = \log(\boldsymbol{\lambda}_i) + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i,$$

- $\boldsymbol{\lambda}_i$: model offset (number of combats)
- $\boldsymbol{\beta}$: fixed effects parameters,
- $\mathbf{u}_i$: random effect intercept,
- Random effects: independent and $\mathcal{N}(\mathbf{0}, \sigma^2)$.

The model under study II

Table: Results from the Poisson mixed model in the folk wrestling data

	GLM		lme4		hglm		INLA	
	$\hat{\beta}$	$s.e.(\hat{\beta})$	$\hat{\beta}$	$s.e.(\hat{\beta})$	$\hat{\beta}$	$s.e.(\hat{\beta})$	$\hat{\beta}$	$s.e.(\hat{\beta})$
Intercept	−4.34	0.18	−4.37	0.2	−4.33	0.2	−4.38	0.2
Category								
Light	0.25	0.23	0.24	0.25	0.25	0.26	0.25	0.24
Semiheavy	0.1	0.23	0.11	0.26	0.11	0.27	0.1	0.25
Heavy	0.39	0.24	0.4	0.27	0.41	0.28	0.4	0.26
Winner	−0.48	0.17	−0.45	0.18	−0.45	0.19	−0.46	0.18
σ_u^2	–		0.10		0.13		0.07	
Dispersion[1]	1.45		1.26		1.23		1.4[2]	

[1] Pearson residuals with function `glm` and for packages `lme4` and `hglm`

[2] Negative Binomial dispersion was calculated in the `INLA` package.

Simulation Study I

- Based on the Poisson model and the wrestling data.
- 1000 runs for all settings
- Two scenarios settings used; Two parameters of interest(intercept and Winner)
- We applied R function glm treating the data as if we dealt with a GLM.
- We assessed the performance of the three statistical principles in terms of **Bias, MSE, ratio and coverage**.
- Convergence problems and Computation times in all packages.

Simulation Study II: Scenario 1

- Real data set; generated values of number of injuries through Poisson distribution.
- Chose the values of fixed effects from lme4 and values of the variance of the random effect were set depending on the overdispersion generated.
- We examined **8 combinations**: 4 values of dispersion(1,1.5,3,10), 1 value of the offset (average number of combat per season: 60) and 2 injuries' marginal means (1(small) and 10 (moderate)).

Simulation Study III: Scenario 2

- A balanced design generated (random intercept model) with sample sizes N=30 and 100 subjects.
- A random number of repeated measures was generated using a discrete uniform distribution ranging from 1 through 6.
- Number of counts was generated for each subject and repeated measure.
- We examined **32 combinations**: 4 dispersion parameters (1, 1.5, 3, 10), 2 marginal means (1 and 10), 2 offsets (60 and 100), 2 sample sizes (30 and 100) denoting small and moderate samples.

Results of Slope parameter I

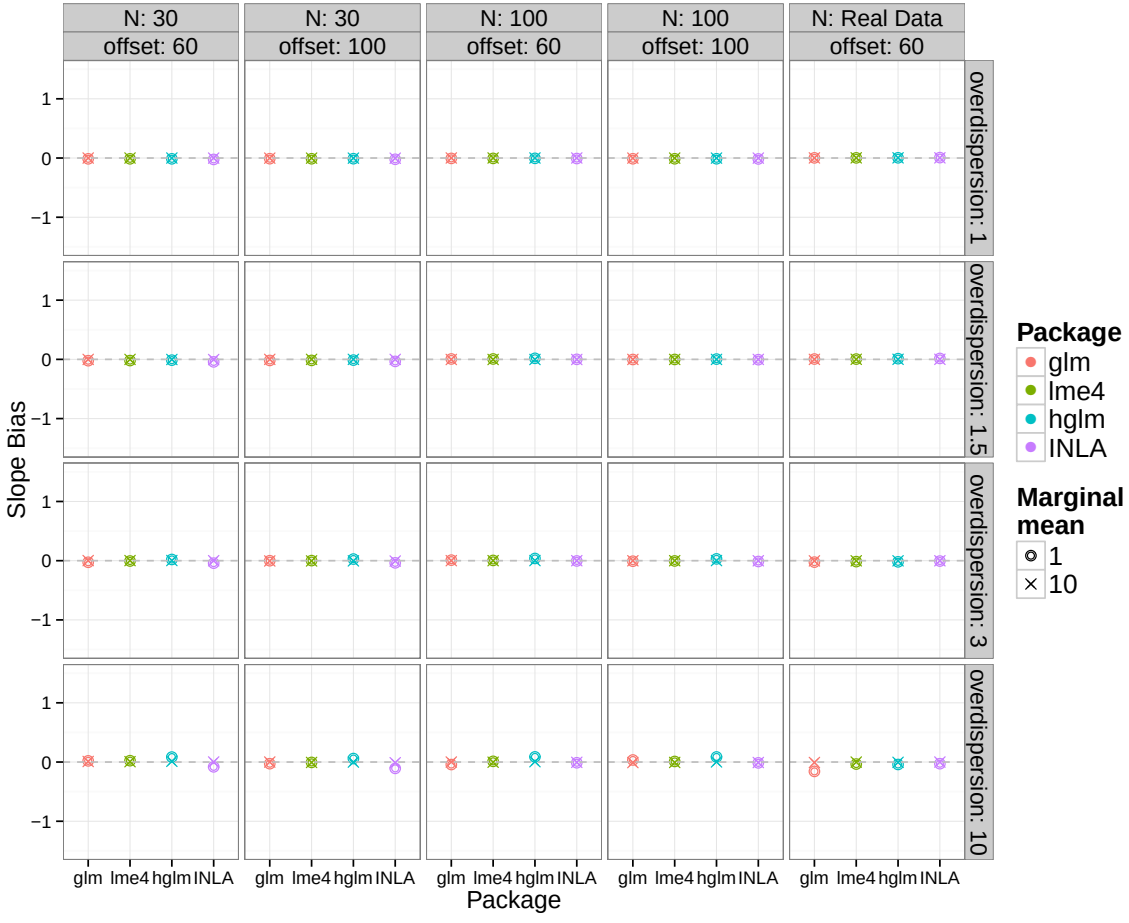


Figure: Bias of the slope estimate as a function of Φ, μ, offset, and N

Results of Slope parameter II

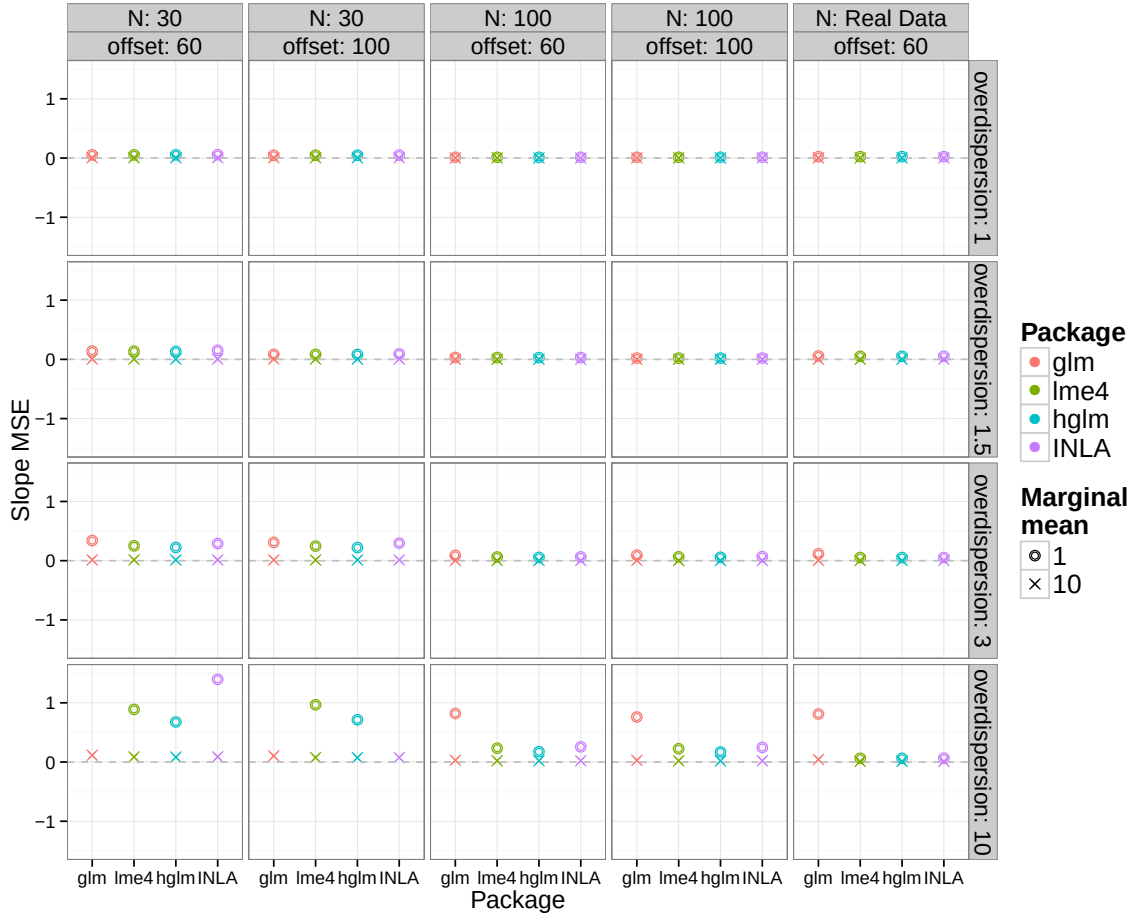


Figure: MSE of the slope estimate as a function of Φ, μ, offset, and N

Results of Slope parameter III

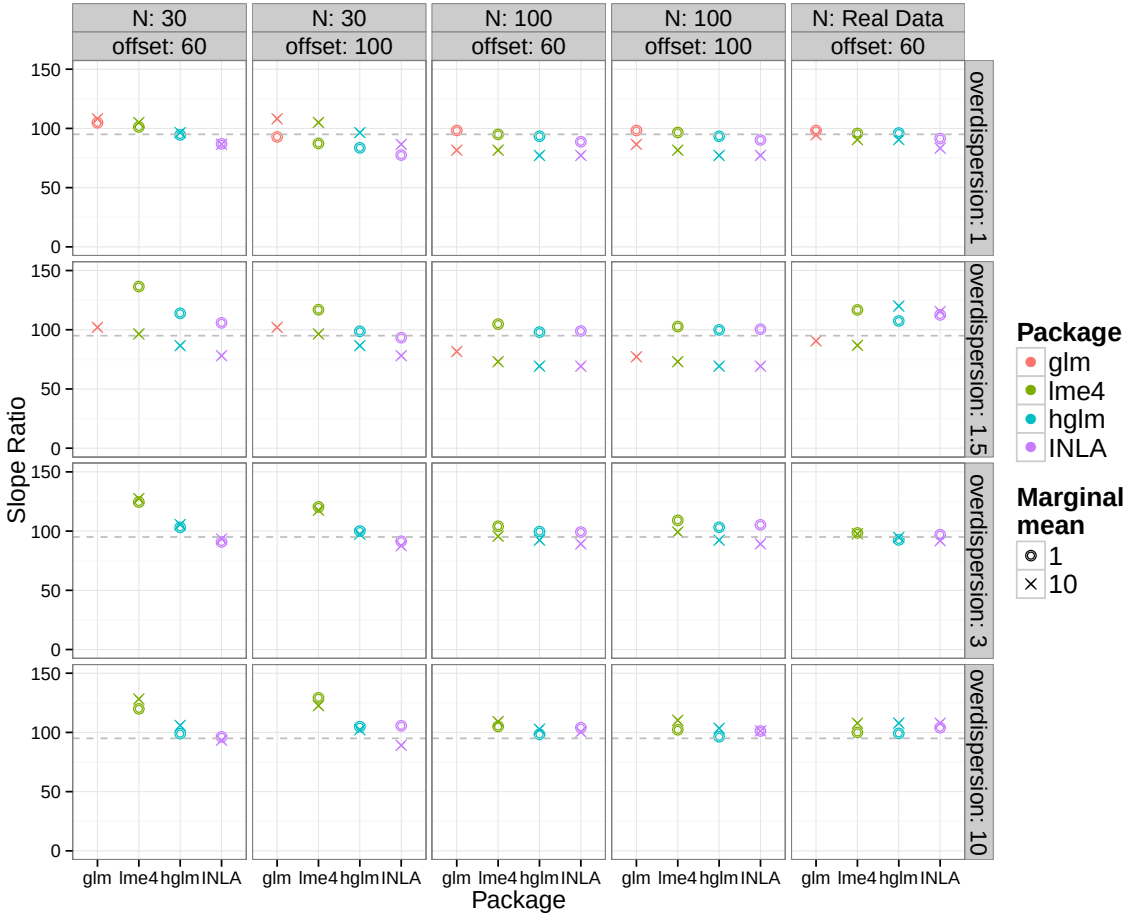


Figure: Ratio of the slope estimate as a function of Φ, μ, offset, and N

Results of Slope parameter IV

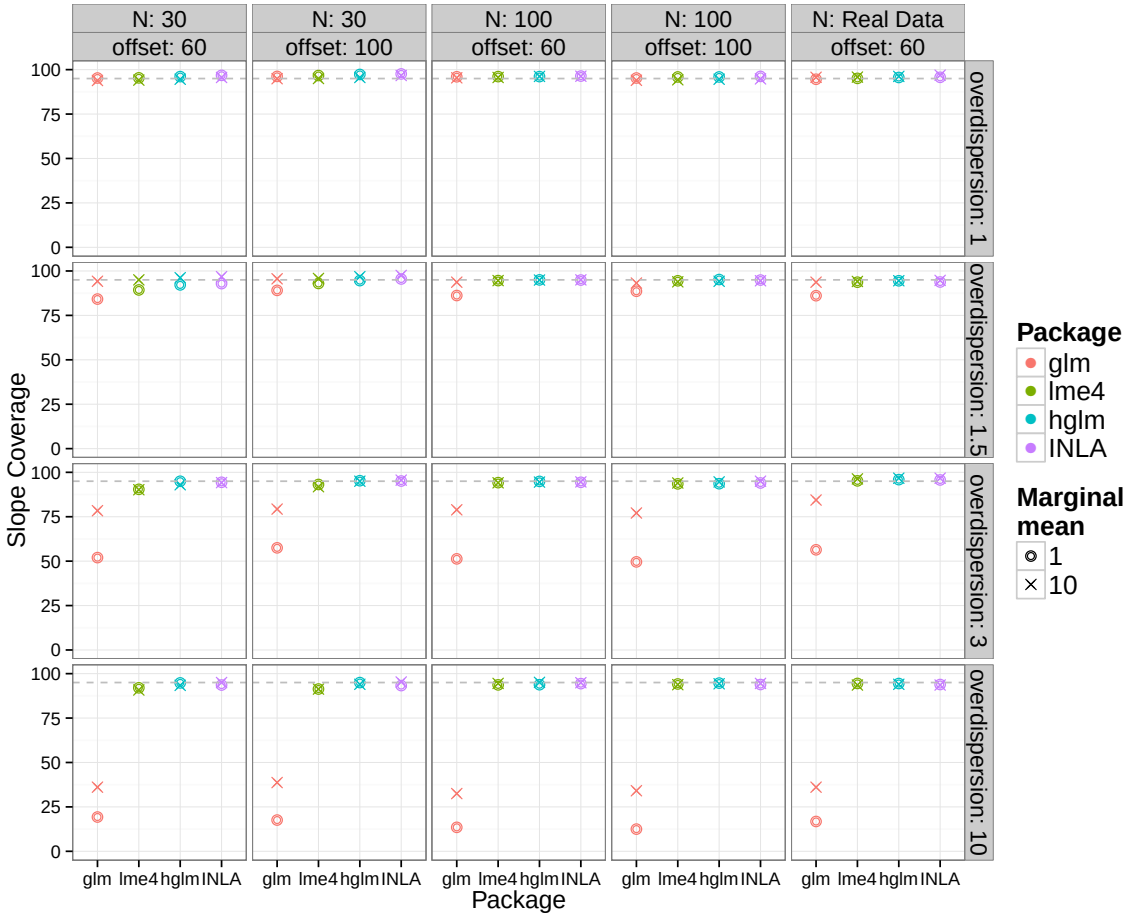


Figure: Coverage of the slope estimate as a function of Φ, μ, offset, and N

Results of Variance Component I

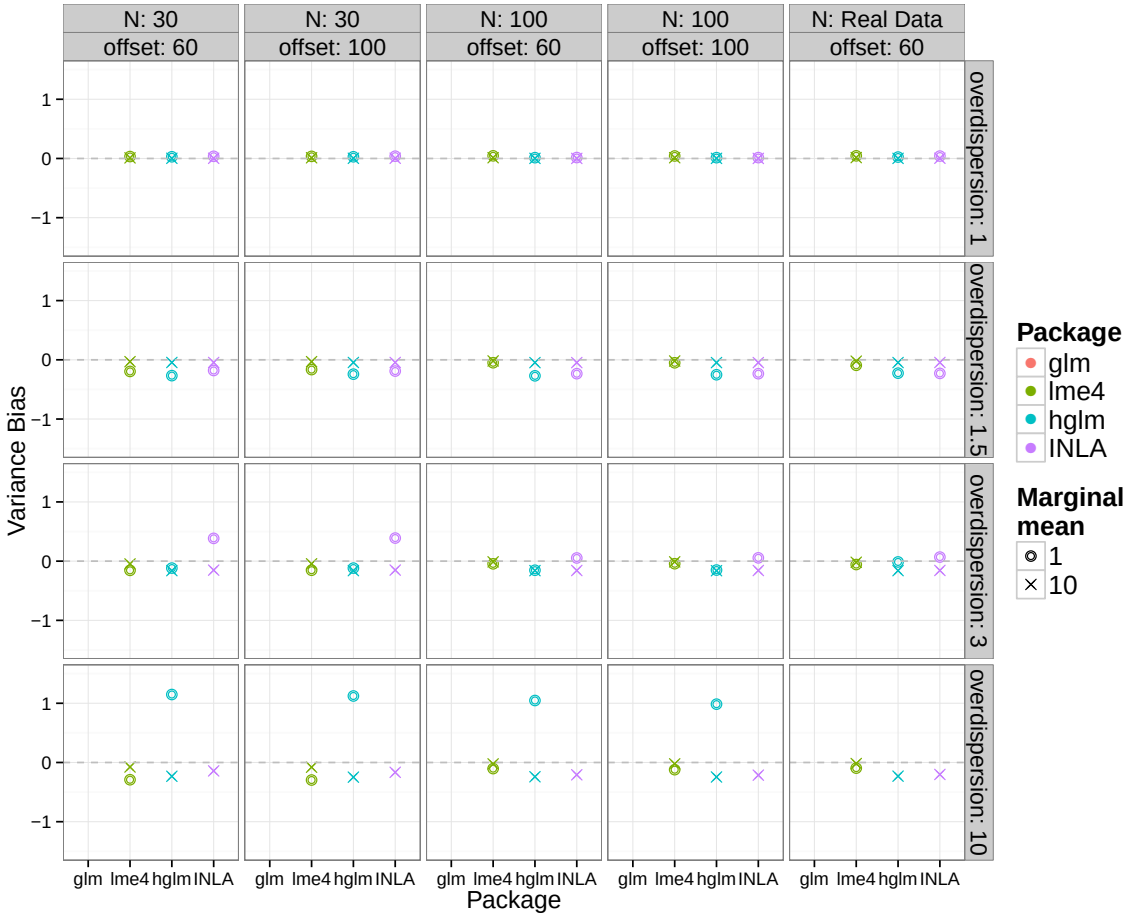


Figure: Bias of the variance component estimate as a function of Φ, μ, offset, and N

Results of Variance Component II

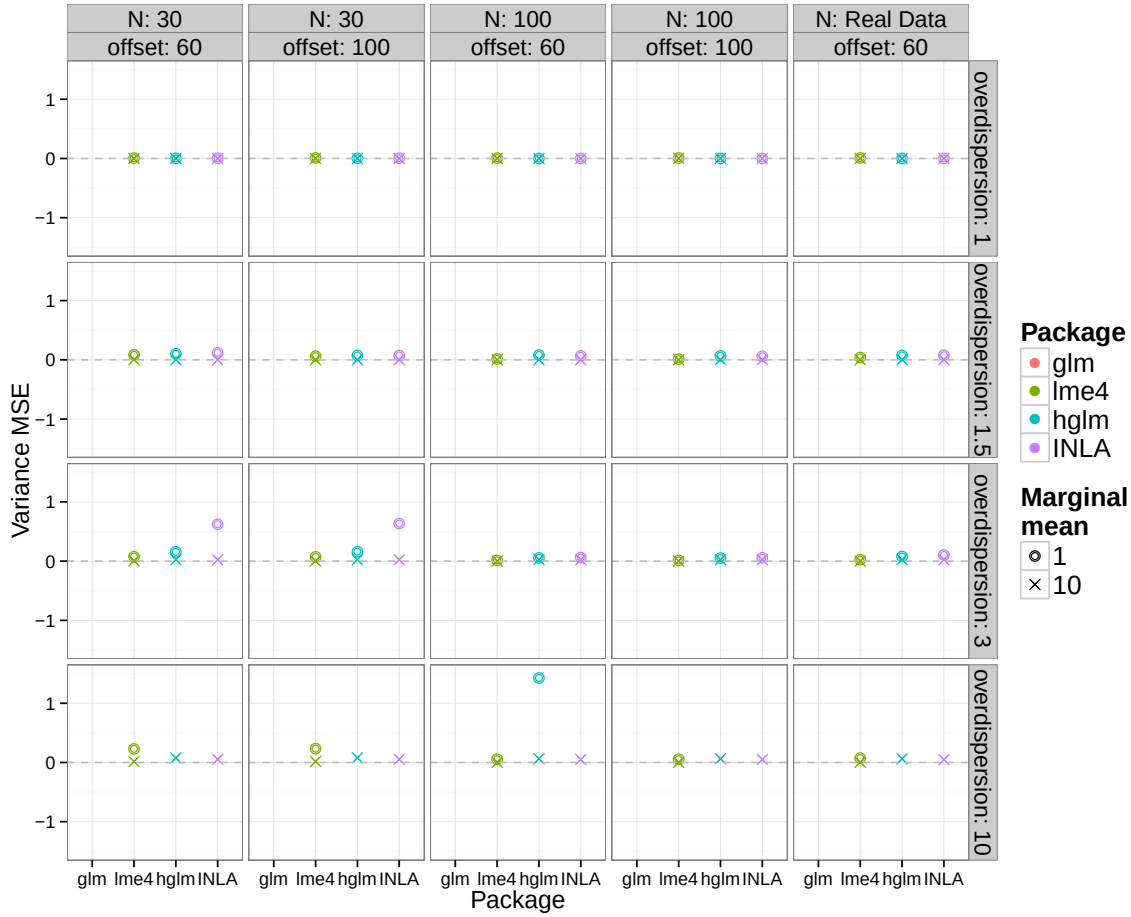


Figure: MSE of the variance component estimate as a function of Φ, μ, offset, and N

Results of Variance Component III

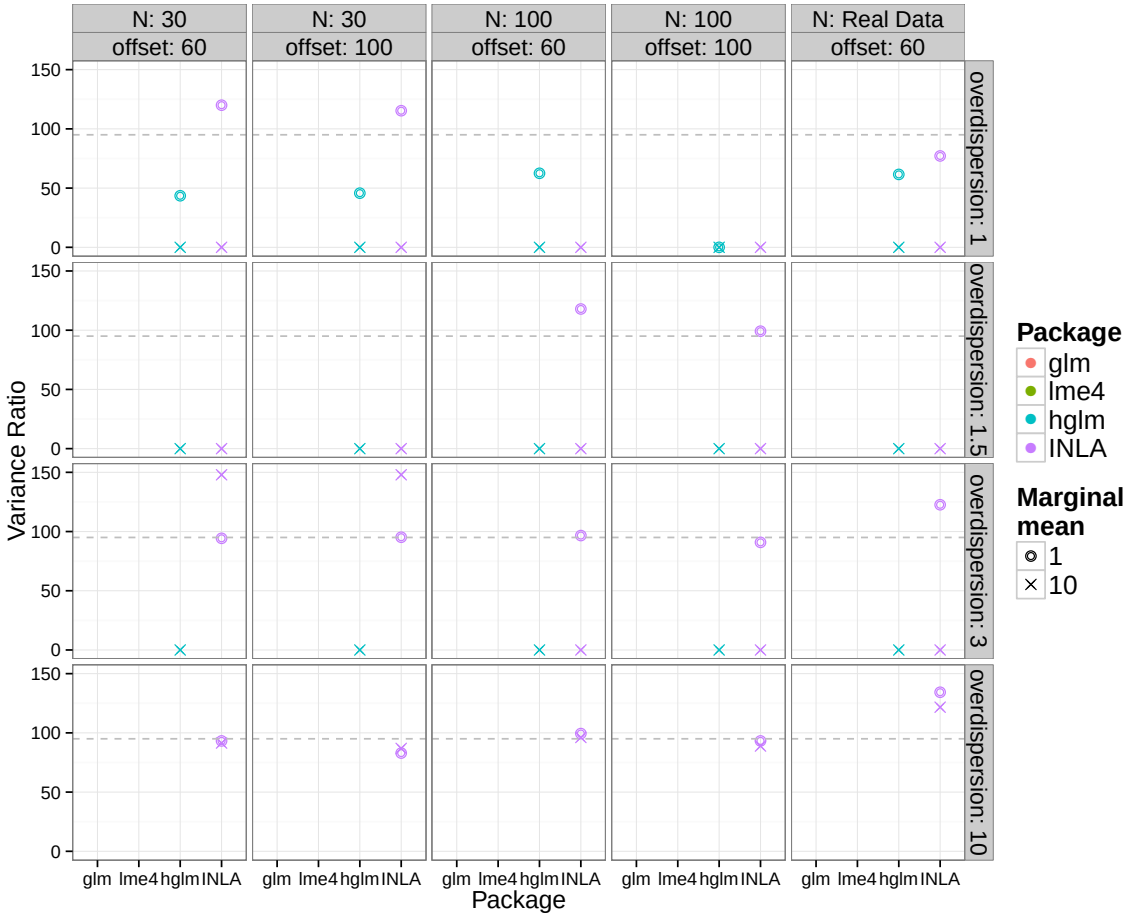


Figure: Ratio of the variance component estimate as a function of Φ, μ, offset, and N

Conclusions

- Approaches involving random effect `lme4`, `INLA` and `hglm` showed a **good performance except** in combinations with **huge overdispersion, small marginal mean and sample sizes**.
- In most extreme settings, we have found warnings of convergence in `hglm` and `lme4` packages. To solve a convergence problem we recommend specifying other starting values.
- In general, the `lme4` package seems more **adequate** than others in most combinations.
- We should take into account the estimation of data with small sample size and small marginal mean **without ignoring overdispersion**.