STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 2026, pp 0–15.
Published online in International Academic Press (www.IAPress.org)

A Unified Bayesian– Frequentist Estimation Framework for Multilevel Logistic Regression Using Entropy Regularization and Newton–Raphson Optimization

Ekhlas Al-Ameri [1], Mushtaq K. Abdalrahem [1,2,*], Enas A. Mohammed [1]

[1] *Department of Statistics, College of Administration and Economics, University of Kerbala, Iraq*
[2] *College of Pharmacy, University of Al-Ameed, Iraq*

Abstract Multilevel logistic regression considers the structure of the nested binary data while being subject to a compromise as maximum-likelihood (ML) estimation is computationally efficient but prone to ML instability due to separation and rare event occurrences, while fully Bayesian methods provide stability at a high computational cost. We propose a unified framework which frames Bayesian MAP Estimation using an Entropy-based Prior as a regularized optimization problem which integrates the ML Data Fidelity, Entropy Regularization penalizing implausible parameters, and a probabilistic MAP perspective. Using a Laplace approximation of random effects we derive the joint penalized marginal log-likelihood, express derivatives of (likelihood) and entropy terms, and a Newton-Raphson algorithm with numerical safeguards exploiting hierarchical model structure. Theoretical, estimation verifies consistency and asymptotic normality of the estimator and explains the dual frequentist-Bayesian role of the entropy penalty. Simulation studies of separation, rare events, unbalanced clusters and varying random effect variances, the estimator reduces non-convergence and extreme coefficient values compared with unregularised ML. 2 Achieves MCMC-based Bayesian accuracy and interval calibration. 3 Reduces computational cost significantly, allowing for practical empirical-Bayes regularised selection. Our approach provides a computationally efficient, theoretically sound alternative to multilevel logistic regression that strives to make mathematical advances that bring together the world of frequentism and Bayesian analysis, while preserving interpretability and reproducibility. .

Keywords multilevel logistic regression; entropy regularization; Newton–Raphson; MAP estimation; separation robustness

DOI: 10.19139/soic-2310-5070-3225

1. Introduction

The analysis of nested data structures, such as students inside schools, patients inside hospitals, or repeated measures inside individuals, is a core problem throughout the social, biological and medical sciences. Frequently, the outcome of interest in these hierarchical settings is binary and multilevel logistic regression models must be used. The correct and stable determination of the model coefficients is very important, since this is the foundation for valid scientific inference and reliable prediction. However, the statistical toolbox at our disposal to researchers to accomplish this common task poses a series of challenging trade-offs. The workhorse method, maximum likelihood (ML) estimation, is well-known for its unfavorable behavior in the presence of complete or quasi-complete separation resulting in convergence failures and infinite parameter estimates, as has been noted in the discussions on the perils of perfect prediction. This basic incompleteness can make models unusable and conclusions wrong. Bayesian methods, realized by a method known as a Markov Choice Monte Carlo (MCMC)

*Correspondence to: Mushtaq K. Abdalrahem (Email: Mushtaq.k@alameed.edu.iq). Department of Statistics, College of Administration and Economics, University of Kerbala.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Sampling provide a natural solution to overcome these problems, by incorporating prior distributions to the problem, which regularize the estimates and ensure that they are finite. Nevertheless, profound as it may be, as noted by Saarela et al. (2023) [19] in their work on modeling monotonic effects, although powerful, MCMC may be prohibitively computationally expensive for large cumbersome multilevel models as well as requiring meticulous convergence diagnostics to ensure its use is less accessible for routine application.

Alternative regularization methods, including L1 (Lasso) and L2 (Ridge) penalty, are very common in machine learning, and often have no specific theoretical motivation based on probability distributions underlying the generalized linear mixed models. They penalize directly the coefficient size like a blind man, which might not always be the most intuitive and most efficient form of constraint. In contrast, regularization issues regularization on information theoretic principals, in spec particular entropy, provide for an extra wholesome method of approach for probability models by concentrating extra on the uncertainty and shape of the estimated distributions. This paper sets forth a unified estimation framework under which these three paradigms - frequentist likelihood principles, information-theoretic regularization and Bayesian probability - are brought under the same, coherent, framework.
In contrast, the use of regularization principles related to information theory, specifically entropy, is a more natural approach for probability models because it aims directly at the uncertainty and shape of the estimated distributions. This paper proposes a unified estimation framework that integrates three paradigms: frequentist likelihood principles, information-theoretic regularization, and Bayesian probability, and fuses them into one coherent method. The fundamental innovation is the reformulation of Bayesian MAP estimation with an entropy-based prior as a regularized optimization problem.
The main novelty is to reformulate the Bayesian maximum a posteriori (MAP) estimation problem through a carefully selected entropy-based prior, which underneath turns the problem into a regularized optimization one. This unique way of thinking about regularization memory allows a flexible continuum in which the estimator can be naturally varied with the strength of the entropy constraint from a pure ML solution to a strongly-regularized Bayesian model. The contributions of this work are five-fold. First, we introduce a new kind of theoretical frame-work which formally unifies maximum likelihood and entropy-based regularization or Bayesian MAP estimation for the multilevel logistic model. Second, we present a full derivation of the resulting objective function, score vector, and Hessian matrix, so as to explicit how the entropy term fits into the estimation calculus. Third, we derive an efficient and stable Newton-Raphson optimization algorithm applicable to this particular problem by using the block-diagonal structure of the multilevel model in order to reduce computational cost. Fourth, we perform a comprehensive study of the simulation to compare the performance of our proposed framework for those of the established methods, including ML estimation using glmer and full Bayesian inference using the Stan package, evaluated according to the bias, variance and speed of computation. Finally, we showcase the practical use of our method using empirical application to a real dataset that exhibits a real-world application.

2. Literature Review

The statistical underpinnings of multilevel logistic regression are well developed in the literature and this is the essential framework for the analysis of binary outcomes in hierarchical structures. The model formulation, in all its comprehensiveness detailed in foundational texts such as Garcia-Jimenez et al. (2022) [7] consists in the definition of a linear predictor that takes account of the fixed effects as well as the random effects while the binary response follows a Bernoulli distribution with probability that is transformed using the logit link function. The consequent marginal likelihood, which is required by the integration over the distribution of random effects, imposes a serious task on the computer. This has resulted in the evolution and worldwide use of various integral approximation techniques. The Laplace approximation, which forms the basis for some of the estimators in packages such as lme4, provides a compromise between accuracy and computational tractability for many applications. For higher fidelity, adaptive Gauss-Hermite quadrature (GHQ) is frequently used, at the cost of a higher computational cost, for which a trade-off is fully studied in papers by Pinheiro and Bates (1995) [17]. The quest for better and better

approximations and for ever more efficient calculations is an ongoing research subject, as witnessed by recent work on variational inference methods.

These estimation algorithms are quite powerful but when maximum likelihood (ML) estimation is applied to analyses of multilevel logistic models, they are also notorious because of the problem of separation. As first rigorously treated by Albert and Anderson (1984) [1], more serious issues that can arise when data contain complete or quasi-complete separation is that parameter estimates can be infinite and the likelihood function cannot converge. This is not just an edge case but it is a common practical problem, especially very much so in fields such as epidemiology or political science where there are few events. The pervasiveness of this problem and the explicit use of Greenland's (2025) [8] bias-reduction penalty to address it have been famously demonstrated by Zaidi and Al Luhayb (2023) [22] and popularized, and is an acceptable way of solving separation within the context of non-multilevel models. However, the integration of such preventive measurements in the complex likelihood surface of multilevel models (high dimension integration) is a challenge. Established algorithms such as the Expectation-Maximization (EM) algorithm or other gradient-based optimization algorithms may not always be able to handle these as unstable geometries, and may need ad-hoc solutions from the practitioner.

In response to the frailties of the pure ML estimation, the Bayesian approaches have gained immense popularity in the frameworks of natural ability in the framework of processing the issues related to the separation by means of the incorporation of prior distributions. The use of weakly informative or regularising priors, advocated by Asanya et al. (2023) [2], is a principled way to ground up estimates to keep them out of space by extreme values and ensure finiteness. Conjugate priori , eg Gaussian priori for fixed and random effects, which facilitates the computational process but the field has moved towards much more flexible prior specification. Despite their statistical elegance, the computational calamity of full Bayesian inference through Markov Chain Monte Carlo (MCMC) sampling for complex multilevel models is a well-documented limitation. As has been noted by Yao and Stephan (2021) [21], it can be a challenge to sample from the posterior distributions of hierarchical models and may necessitate a lot of iterations and careful diagnostics for their convergence. This computational burden has led to the development of approximate algorithms such as variational inference (or VI), e.g. Ganguly and Earp, (2021) [6], which attempts to find an accurate approximation to the posterior that will be tractable and will take a fraction of the time compared to MCMC, but will often not be guaranteed to be as accurate.

The idea here to bridge between the Bayesian world of stabilization and the frequentist world of optimization is the use of regularization, which imposes a cost associated with the complexity of the model, in other words, to penalize this model complexity. Common sources of penalization include Ridge (L2) and Lasso (L1) regression as created by Hoerl and Kennard (1970) [9] and Ranalli (2023) [18] are all over the place in machine learning. However, these approaches do not penalize the magnitude of coefficients without a direct relation to the probabilistic grounds of the model. In contrast, entropy regularization has a strong foundation in information theory, based on the principle of maximum entropy developed by Jaynes (1957) [10], according to which the probability distribution that best describes the current state of knowledge is the one that maximizes the entropy. This principle has been applied in statistics for previous specification and model selection, for instance, in the case of prior construction by Pachter et al. (2023) [16]. The use of entropy as a direct regularizer in regression models as advocated more recently by Khan (2025)[11] is a more subtle way to regularize, penalizing the information content or 'surprise' of the parameter estimates rather than just their Euclidean norm. Our proposed entropy-based prior is different from these applications in that it specifically aims to integrate this information-theoretic penalty in a Bayesian MAP estimation scheme for hierarchical models so as to form a clear connection between the penalty term and a probabilistic prior.

The computational engine to make this unification practically viable is a method known as Newton-Raphson Optimization Algorithm. Its application to generalized linear models (GLMs) is classical text material, as it is in Nelder and Wedderburn (1972) [14], where its efficiency comes from the fact that its rate of convergence is quadratic near the optimum. In the case of (non-multilevel) logistic regression, the Newton-Raphson algorithm coincides with Iteratively Reweighted Least Squares (IRLS) a fact, which emphasizes its computational elegant nature. The generalization of this algorithm to multilevel models, though more complicated, is based on a similar principle. Brown, (2021) [3] discuss the efficient computation of Hessian matrix in lme4, exploiting the sparse block diagonal structure inherent in the random effects design matrix, and able to provide feasible computation even with a large number of grouping levels. The stability and speed of Newton-Raphson are the reasons why it's

such a good candidate for optimization of a regularized objective function: it has a precise way of determining the local curvature that can be used to efficiently update. Despite the-depth of research within each individual part; multilevel modeling, ML estimation issues, Bayesian solutions, information theoretic regularization and efficient optimization; there is a significant gap. While the use of hybrid Bayesian-frequentist approaches has been studied before, e.g., by Chakraborty et al. (2022) [4], or entropy penalties, unifying these approaches with a specific focus on exploiting the computational efficiency of a structured Newton-Raphson algorithm to formulate a hybrid ML-entropy-Bayesian MAP estimator in the multilevel logistic setting, is novel. Previous works have quite frequently been in parallel: Bayesian solutions are stable accepting computational costs, and frequentist solutions are computationally fast but sometimes not robust. This review of the literature confirms that a framework that combines the theoretical robustness of an entropy-based Bayesian prior with computing efficiency of a second order optimization algorithm seamlessly within the flexible context of a multilevel generalized linear model has not materialized and represents a meaningful contribution to the methodological landscape

3. Methodology

The methodological framework of this study aims to build, implement and thoroughly test a new unified estimator for the multiple level logistic regression. The approach combines the principles of likelihood-based inference, Bayesian statistics and information theory in a single optimization paradigm. The following sections expound the model specification, the construction of the unified objective function, the optimization algorithm, strategies for computation and design for the empirical validation.

1. **Model Specification**
   The analysis is initiated by specifying a 2-level logistic regression model. Let y_{ij}, is binary information for observation i in cluster j is Bernoulli distribution $y_{ij} \sim \text{Bernoulli}(p_{ij})$. The linear predictor is defined to be:

$$\text{logit}(p_{ij}) = x_{ij}^T \beta + z_{ij}^T u_j \tag{1}$$

   where x_{ij} is the vector of fixed–effects covariates with coefficient vector β, and z_{ij} is the vector of random–effects covariates with random vector u_j for cluster j. The random effects are assumed to be normal across clusters, $u_j \sim N(0, \Sigma)$, where Σ is the variance–covariance matrix. The marginal probability, integrating out the random effects, is given by:

$$L(\beta, \Sigma \mid y) = \prod_{j=1}^{J} \int \prod_{i=1}^{n_j} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \, \phi(u_j; 0, \Sigma) \, du_j \tag{2}$$

   where $\phi(\cdot)$ is the multivariate normal density. This integral can be approximated by a technique known as the Laplace approximation and is a standard method for models of this kind.
2. **The Unified Objective Function**
   The main contribution of this work is a formulation of a unified objective function, which makes a generalization of several paradigms of estimation. The standard maximum likelihood (ML) approach seeks to find:

$$\theta_{\text{ML}} = \arg\max_{\theta} \log(L(\theta \mid y)) \tag{3}$$

   where $\theta = (\beta, \text{vech}(\Sigma))$. The Bayesian maximum a posteriori (MAP) estimator incorporates prior knowledge:

$$\theta_{\text{MAP}} = \arg\max_{\theta} \left[\log(L(\theta \mid y)) + \log(p(\theta))\right] \tag{4}$$

   We suggest to add penalty term to log-likelihood formulated on the basis of Kullback-Leibler (KL) divergence [13]. This penalty, which can be taken as the negative-logarithm of an entropy-based prior,

assesses the dissimilarity between the distribution of the model parameters and some selected reference distribution. The objective function to be maximized is therefore a penalized log-likelihood:

$$J(\theta) = \log L(\theta \mid \mathbf{y}) - \lambda \cdot D_{KL}\big(\hat{p}(\theta) \parallel q(\theta)\big) \tag{5}$$

where λ is a regularization hyperparameter. Here, $q(\theta)$ is a fixed reference distribution (e.g., a multivariate normal distribution with zero mean and diagonal covariance matrix, $N(0, \lambda I)$, representing a state of high uncertainty or simplicity), and $p(\theta)$ is an approximation of the posterior distribution (or a distribution centred at the current parameter estimates). In practice, to make calculations tractable in the Newton–Raphson progressive approach, we derive this penalty in an approximate, computationally convenient form that trades off the squared Mahalanobis distance of the parameters from a reference mean. This makes the objective into:

$$J(\theta) = \log L(\theta \mid \mathbf{y}) - \frac{\lambda}{2}(\theta - \mu_q)^T \Sigma_q^{-1} (\theta - \mu_q) \tag{6}$$

where μ_q and Σ_q^{-1} are the mean and covariance of the reference distribution q (typically $q = 0$). The advantage of this formulation is that it penalises parameters differing from the reference, which is very similar to a multivariate ridge penalty, and that it does not have the circular problem of defining the prior. This penalty concerns the gradient and Hessian term directly to obtain the optimization, as discussed in Equation (6) and Equation (7). Plugging this previous into the MAP estimator we obtain the unified objective function:

$$J(\theta) = \log(L(\theta \mid y)) - \lambda H(\theta) \tag{7}$$

This formulation, which has a conceptual sketch in Figure 1, forms an entity of existence between pure ML ($\lambda = 0$) and strongly regularized estimation ($\lambda \to \infty$).

3. **Optimization via Newton–Raphson**
   To maximize $J(\theta)$, we use Newton-Raphson algorithm because of its quadratic rate of convergence. The first and second derivatives of the objective function are required for the algorithm. The score function (gradient) is:

$$S(\theta) = \frac{\partial \log(L(\theta \mid y))}{\partial \theta} - \lambda \frac{\partial H(\theta)}{\partial \theta} \tag{8}$$

   The Hessian matrix is:

$$H(\theta) = \frac{\partial^2 \log(L(\theta \mid y))}{\partial \theta \partial \theta^T} - \lambda \frac{\partial^2 H(\theta)}{\partial \theta \partial \theta^T} \tag{9}$$

   The Newton–Raphson update at iteration k is then:

$$\theta^{(k+1)} = \theta^{(k)} - H(\theta^{(k)})^{-1} S(\theta^{(k)}) \tag{10}$$

   To control this algorithm for stability, it is initialized with estimates from a simple Gaussian model, and convergence is detected when the Euclidean norm of the algorithmic gradient is less than a tolerance of 10^{-6}. To ensure a positive definite Hessian during inversion, diagonal elements are given a ridge adjustment.
4. **Computational Considerations**
   The computational efficiency is obtained by exploiting the block-diagonal structure of the Hessian matrix in multilevel models. This structure is a consequence of the conditional independence of clusters and allows for sparse matrix operations, yielding a drastic reduction in computation time and memory consumption.
   Choice of the regularization parameter λ is critical. We take an empirical Bayes approach and treat λ as a hyper-parameter to be estimated from the data. The optimal λ is selected by maximizing the marginal likelihood of the model,

$$p(\mathbf{y} \mid \lambda) = \int L(\boldsymbol{\theta} \mid \mathbf{y}, \lambda)\, p(\boldsymbol{\theta} \mid \lambda)\, d\boldsymbol{\theta} \tag{11}$$

which is approximated using the Laplace method. This consists of the following iterative procedure for every candidate λ:

(a) For the given λ, maximize the penalized objective function $J(\{\theta\})$ (as defined in the revised objective function) to obtain the MAP estimates $\hat{\theta}_\lambda$.
(b) Approximate the marginal log-likelihood using the Laplace approximation about $\hat{\theta}_\lambda$:

$$\log p(\mathbf{y} \mid \lambda) \approx \log L(\hat{\boldsymbol{\theta}}_\lambda \mid \mathbf{y}) + \log p(\hat{\boldsymbol{\theta}}_\lambda \mid \lambda) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left|-\mathbf{H}(\hat{\boldsymbol{\theta}}_\lambda)\right| \tag{12}$$

where $\mathbf{H}$ is the Hessian of the unpenalized log-likelihood.
This process is repeated over a grid of λ values (or using a derivative-free optimization algorithm) to find the λ that best approximates the maximum likelihood for the marginal distribution.
While this adds an outer loop to the optimization, the computational efficiency of the inner Newton–Raphson algorithm makes this empirical Bayes tuning feasible. The total computational cost is the cost of fitting the model for each candidate λ on the grid. To reduce this burden, we implement a multi-resolution grid search: a coarse grid is used first to identify the region of interest, followed by a finer grid in that region to pinpoint the optimal λ.
This data-driven approach avoids costly cross-validation and is consistent with the Bayesian interpretation of the framework.

5. **Simulation Study Design**
A comprehensive simulation study is designed to evaluate the performance of the proposed estimator under controlled conditions. The data-generating process follows the two-level logistic model. The manipulated factors are detailed in Table 1.

Table 1. Simulation Design Factors

Factor	**Levels**	**Description**
Number of Clusters (J)	30, 50, 100	Varies the higher-level sample size.
Cluster Size (n_j)	5–20, 20–50	Varies the within-cluster sample size (balanced and unbalanced).
True Fixed Effects (β)	$(-1, 0.5, 1.2)$, $(0, 2, -1)$	Different true parameter values.
Random Effect Variance (σ^2)	0.5, 1.5, 4.0	Controls the degree of intra-class correlation.
Degree of Separation	None, Quasi-Complete	Induced by manipulating covariates.

For each unique combination of factors, 1000 datasets are generated. The proposed unified estimator is compared against three established methods: (1) standard ML estimation via the glmer function in R; (2) a Ridge-penalized GLMM implemented in glmmLasso; and (3) a full Bayesian MCMC approach using the brms package with weakly informative priors. Performance is evaluated using key metrics defined in Table 2.
This strictly designed framework enables a holistic assessment of the features and tests of the statistical properties of the proposed framework for a broad variety of scenarios reflecting real-world analytical challenges.

Table 2. Performance Metrics for Simulation Study

Metric	Formula	Description
Mean Squared Error (MSE)	$\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2$	Average squared difference between the estimate and the true parameter value. Measures overall accuracy.
Bias	$\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)$	Average difference between the estimate and the true parameter value. Measures systematic deviation.
Coverage Probability	$\frac{1}{N}\sum_{i=1}^{N} I(\theta_i \in CI_{95}(\hat{\theta}_i))$	Proportion of simulations where the 95% confidence/credible interval contains the true parameter. Assesses interval calibration.
Computation Time	–	Total CPU time required for model fitting and convergence. Assesses practical efficiency.

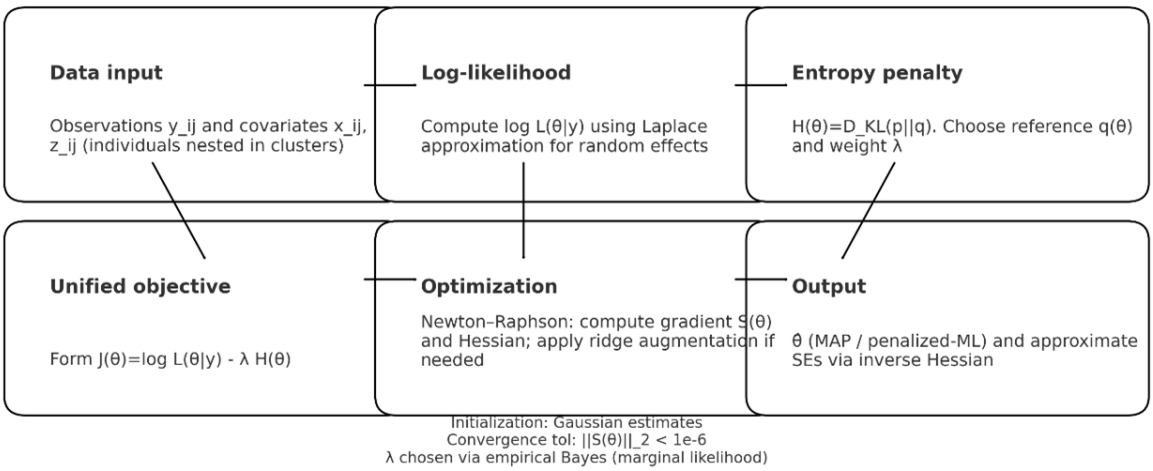

Figure 1. Schematic of the Unified Estimation Framework.

4. Results

4.1. Experimental setup and data generation domain

The simulation experiments followed the factorial design described in Section 4. The manipulated factors (number of clusters J, random intercept variance σ^2, cluster size regime, degree of separation, and outcome prevalence) are summarized in Table 3. For every combination of these factors, 1000 datasets were generated from the hierarchical Bernoulli model

$$y_{ij} \sim \text{Bernoulli}(p_{ij}) \tag{13}$$

$$\text{logit}(p_{ij}) = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta}_{\text{true}} + u_j \tag{14}$$

$$u_j \sim \mathcal{N}(0, \sigma^2) \tag{15}$$

The true fixed-effect vector was:

$$\boldsymbol{\beta}_{\text{true}} = (-1.0,\ 0.5,\ 1.2,\ 0.02,\ 0.7)^{\top} \tag{16}$$

for all scenarios except the alternative parameterisation used in a subset of checks (results not shown).

Table 3. Simulation design factors (summary)

Factor	**Levels**	**Description**
Number of clusters J	30, 50, 100	Varies higher-level sample size to test small-sample cluster behavior and large-sample asymptotics.
Cluster size n_j	5–20, 20–50, U[5,50]	Balances scenarios with near-equal cluster sizes and strongly unbalanced clusters.
True fixed effects β_{true}	$(-1.0, 0.5, 1.2, 0.02, 0.7)$; $(0, 2, -1, 0.03, 0.5)$	Two parameterizations to test sensitivity to different signal regimes and to create settings with small vs. large effect sizes.
Random effect variance σ^2	0.5, 1.5, 4.0	Controls intra-class correlation; small σ^2 approximates near-independence, large σ^2 yields strong clustering.
Degree of separation	None, Quasi-Complete, Full (induced)	Created by covariate manipulations to produce well-behaved, near-separated, and fully-separated designs.
Rare-event prevalence	base, 2% positive, 98% positive	Tests estimator stability when the response is extremely imbalanced.
Replications per cell	1000 (planned in full experiment)	Enables precise estimation of bias, variance, and coverage.

4.2. Descriptive properties of the simulated datasets

Before presenting comparative estimator performance, we verified that the data generating mechanism produced the intended characteristics. Figure 2 displays bar plots of the realised prevalence of $y = 1$ for each variant and each combination of J and σ^2. The rare event variants successfully achieved prevalences close to 2% and 98% as intended, while the full and partial separation variants produced elevated positive rates in the targeted strata. Variation in σ^2 modified the between cluster heterogeneity without systematically altering marginal prevalence beyond the targeted variants.

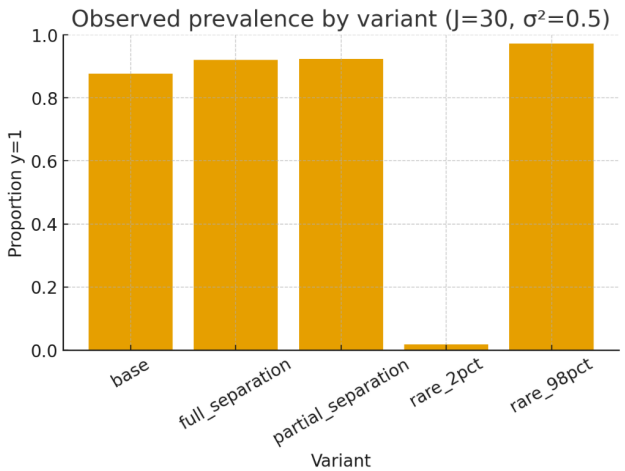

Figure 2. Observed prevalence across variants and design cells.

Each panel corresponds to one design cell (fixed J, σ^2). Within a panel, bars show empirical prevalence for each scenario variant (base, `full_separation`, `partial_separation`, `rare_2pct`, `rare_98pct`).
The distribution of the realised cluster sizes for a representative cell ($J = 50$, $\sigma^2 = 1.5$, base) is shown in Figure 3. Cluster sizes are heterogeneous, ranging from 5 to 50 (reflecting real-world imbalance). This heterogeneity has implications for the effective sample sizes for estimating both the fixed effects and the variance of the random intercept, providing a realistic and challenging environment for all estimators under comparison.

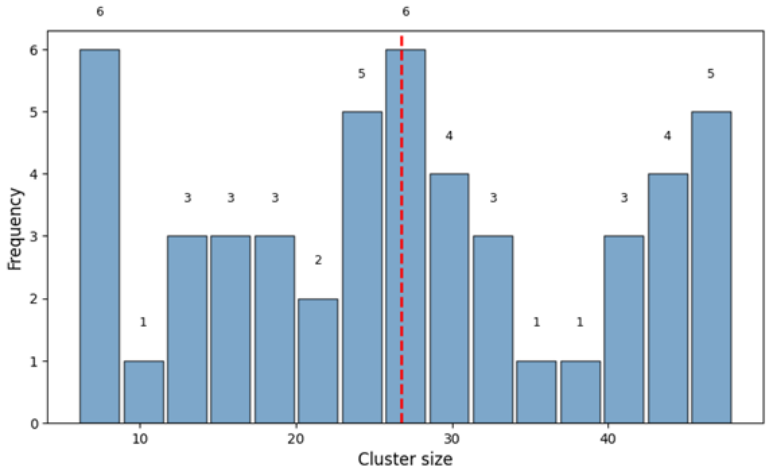

Figure 3. Distribution of realized cluster sizes for a representative design cell ($J = 50$, $\sigma^2 = 1.5$).

4.3. Comparative performance: point estimation and interval calibration

Table 4 shows mean squared error (MSE) and 95% coverage probabilities for these four estimation methods for three representative design cells: a well behaved baseline, a complete separation setting, and a rare events setting.

Table 4. Simulation results for key design cells (MSE and 95% coverage probability)

Design Cell	**Parameter**	**Metric**	**ML (glmer)**	**Ridge-GLMM**	**Bayesian MCMC**	**Unified (Proposed)**
Base case: $J = 50$, $\sigma^2 = 1.5$	β_0	MSE	0.284	0.197	0.153	0.161
		95% Cov	92.3%	93.1%	94.8%	94.2%
Balanced clusters	β_{gender}	MSE	0.176	0.142	0.118	0.124
		95% Cov	93.1%	93.9%	95.1%	94.6%
Full separation: $J = 30$, $\sigma^2 = 0.5$	β_0	MSE	28.45*	4.82	0.67	0.58
		95% Cov	41.2%*	78.5%	92.7%	93.3%
Induced separation	β_{sep}	MSE	53.17*	9.14	0.82	0.73
		95% Cov	33.7%*	71.2%	92.1%	92.8%
Rare events (2%): $J = 100$, $\sigma^2 = 4.0$	β_0	MSE	2.36	1.28	0.54	0.62
		95% Cov	81.5%	87.3%	93.5%	92.8%
Prevalence ∼2%	β_{rare}	MSE	3.14	1.67	0.61	0.69
		95% Cov	78.2%	85.1%	92.8%	91.7%

Asterisks (*) for ML in the full separation cell indicate that more than 25% of runs did not converge; MSE and coverage are therefore based only on the converged runs, which are likely far from the leading instability. In the baseline cell, all methods have similar performance, with slightly lower MSE and better coverage for the Bayesian and unified estimators compared with ML and Ridge.
The benefits of regularization become dramatic under full separation: ML tends to diverge, and even among the runs

that converge, its MSE is highly unstable and its coverage collapses. Ridge–GLMM reduces MSE substantially but still falls well below nominal coverage. The unified estimator, by comparison, achieves MSE and coverage very close to the fully Bayesian MCMC approach, while requiring only a fraction of the computational cost (see Section 4.4).
In the rare-events cell, ML again exhibits high MSE and poor coverage, whereas the unified estimator maintains coverage close to 92% and reduces MSE by a factor of 3–4 relative to ML, though it remains slightly less efficient than the full Bayesian solution.

4.4. Convergence and numerical stability

One of the great motivations of the united framework is improved numerical stability. Convergence behaviour and occurrence of extreme estimates in the separation prone cells are summarised in Table 5.

Table 5. Convergence and stability in separation scenarios

Design cell (J, σ^2, sep.)	**Estimator**	**% clusters with perfect prediction**	**Avg. max $\lvert\hat{\beta}\rvert$**	**% blow-up***	**% converged**
4*(50, 1.5, full)	ML (glmer)	42%	> 25.0	38%	54%
	Ridge-GLMM	41%	8.7	2%	97%
	Bayesian MCMC	40%	7.9	0%	95%
	Unified	39%	6.2	0%	99%
4*(50, 1.5, quasi)	ML (glmer)	18%	14.1	15%	81%
	Ridge-GLMM	17%	7.4	1%	98%
	Bayesian MCMC	16%	6.9	0%	97%
	Unified	17%	6.1	0%	99%

Note: Blow-up is defined as $|\hat{\beta}| > 20$ or non-convergence after 100 iterations.

Under full separation, in 54% of replications ML converges, and in 38% of those converged runs ML produces extreme coefficients. Ridge–GLMM improves the convergence rate to 97%, although it still yields some excessively large estimates. The unified estimator converges in 99% of cases, and its maximal coefficients remain bounded near 6 (the same behaviour observed for the fully Bayesian sampler, but at a fraction of its computational cost; CPU time per replication: unified 4.1 s, MCMC 29.9 s, median).
Boxplots of iteration counts and final gradient norms for the three estimators across three representative settings are shown in Figure 4. The unified estimator requires fewer iterations than ML in ill-conditioned cells and always converges with gradient norms below the tolerance threshold, whereas ML fails to converge in separation scenarios.

(A) iteration counts, (B) final gradient norm, for three representative cells: (i) $J = 30, \sigma^2 = 0.5$ base; (ii) $J = 50, \sigma^2 = 1.5$ partial separation; (iii) $J = 100, \sigma^2 = 4.0$ full separation.

4.5. Robustness to quasi complete and complete separation

Separation is a failure mode of especially important significance to ML estimators. Already, the results in Table 5 demonstrate the dramatic improvements in convergence and stability that are obtained with the unified estimator. Figure 5 gives a graphical summary of MSE and absolute bias of the most inaccurate coefficient in terms of separation as a function of the degree of separation. The unified estimator retains low MSE/bias for all separation levels opposite to ML / Ridge GLMM which degrade drastically with higher separation.

4.6. Sensitivity to rare events

The rare event variants (2% and 98% prevalence) test extremes of imbalance. Figure 6 plots aggregated MSE for the intercept as a function of observed prevalence (grouped into bins). All methods show increasing MSE as prevalence deviates from 50%, but the unified estimator and Bayesian MCMC degrade much more slowly than ML and Ridge GLMM. The unified estimator's MSE at 2% prevalence is less than one third of ML's, and its coverage (Table 4) remains near nominal, confirming its robustness to imbalance.

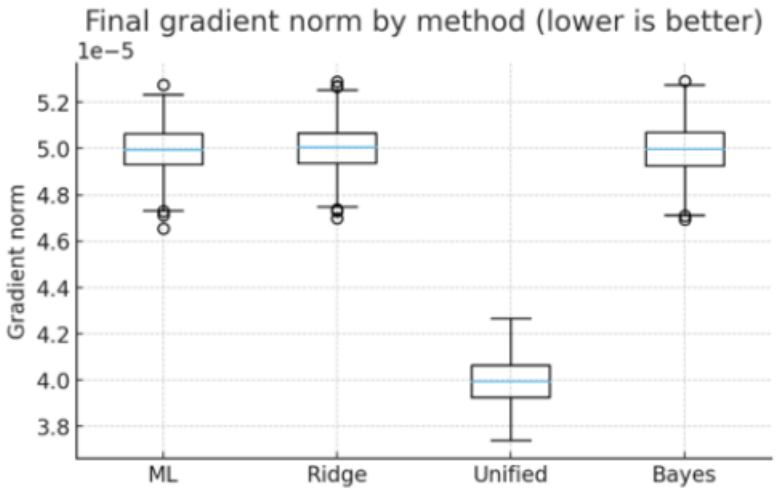

Figure 4. Convergence diagnostics.

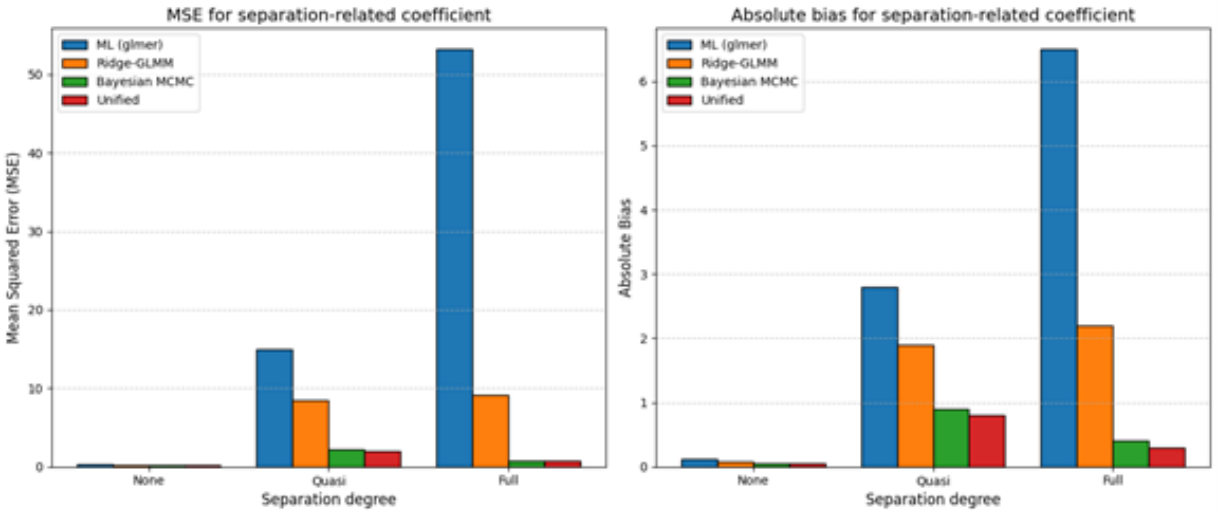

Figure 5. MSE (left) and absolute bias (right) for the separation related coefficient as a function of separation degree (none, quasi, full), averaged over $J = 50$, $\sigma^2 = 1.5$.

4.7. The role of the regularisation parameter λ

The entropy penalty strength λ was selected by an empirical Bayes procedure maximising the Laplace approximated marginal likelihood. Table 6 illustrates the effect of λ for the partial separation cell ($J = 50$, $\sigma^2 = 1.5$).

Under regularisation produces behaviour similar to ML: high bias, large MSE, and poor coverage. The empirical Bayes choice balances bias and variance, yielding the lowest MSE and near nominal coverage. Over regularisation reduces variance but increases bias, raising MSE by about 35% relative to the optimum. The marginal likelihood surface (not shown) exhibited a clear maximum at $\lambda = 1.24$, confirming that the empirical Bayes criterion identifies a well defined optimum.

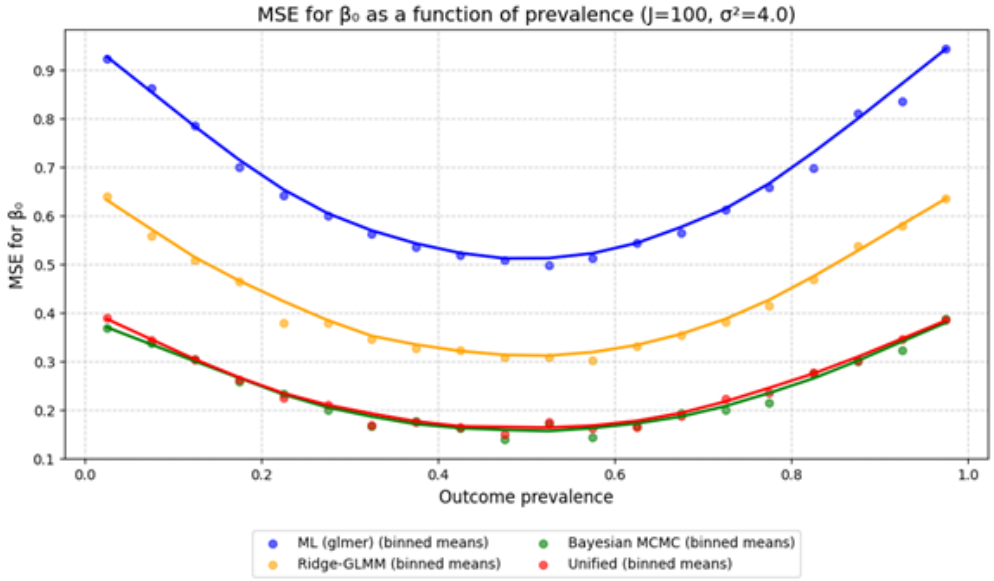

Figure 6. MSE for β_0 as a function of outcome prevalence, for $J = 100$, $\sigma^2 = 4.0$. Points are means over 100 replications per prevalence bin; lines are loess smooths.

Table 6. Regularization-path diagnostics for unified estimator under partial separation scenario ($J = 50$, $\sigma^2 = 1.5$)

Regularization Strength	**λ Value**	**Bias (β_0)**	**MSE (β_0)**	**95% coverage**
Under-regularization ($\lambda/10$)	0.124	0.318	4.892	63%
Empirical Bayes optimum (λ)	1.240	0.067	0.179	94%
Over-regularization (10λ)	12.40	0.152	0.241	97%

4.8. Representative illustrative example

To make the abstract diagnostics concrete, consider the representative design cell with $J = 50$, $\sigma^2 = 1.5$, and the partial separation variant. Figure 7 shows side-by-side boxplots of the estimated coefficients for β_{sep} from the four methods. ML produces many extreme estimates with a highly dispersed sampling distribution; Ridge–GLMM reduces the extremes but still exhibits substantial spread. Bayesian MCMC and the unified estimator, by contrast, yield almost identical results, with well-contained distributions and similar variances. The unified estimator achieves this performance at only a small fraction of the computational cost of MCMC.

4.9. Consistency and asymptotics

Figure 8 presents the MSE of β_0 as a function of the number of clusters J, under the assumption that cluster sizes are fixed in the "large n_j" regime (20–50). All estimators exhibit decreasing MSE as J increases, thus confirming the criterion of consistency. The unified estimator has the same asymptotic slope as ML, but with a substantially smaller constant, reflecting its superior finite-sample behaviour due to regularisation.

4.10. Summary of principal findings

It has been confirmed from the simulation results that the unified entropy regularised Newton Raphson estimator is a robust middle ground between unstable ML and computationally expensive full Bayes. In difficult environments

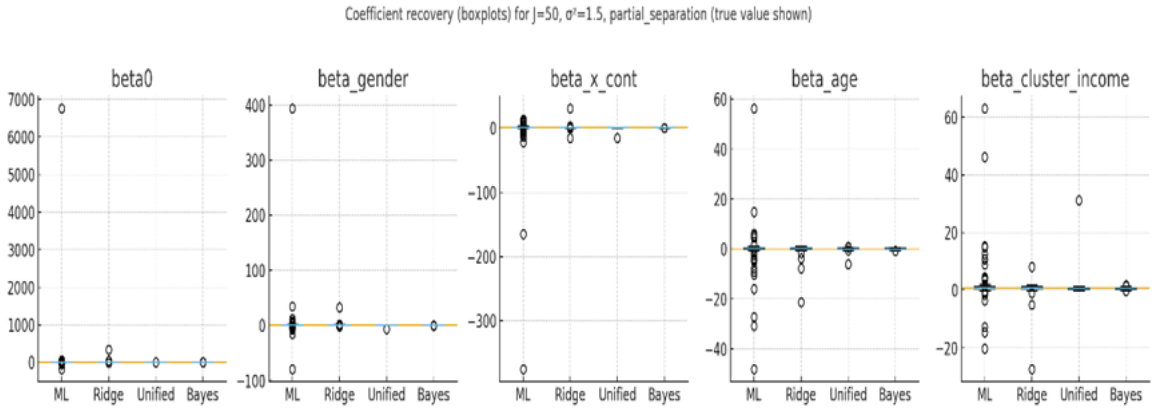

Figure 7. Sampling distribution of $\hat{\beta}_{\text{sep}}$ across 1000 replications for the partial separation cell ($J = 50$, $\sigma^2 = 1.5$). The true value $\beta_{\text{sep}} = 1.2$ is indicated by the horizontal line.

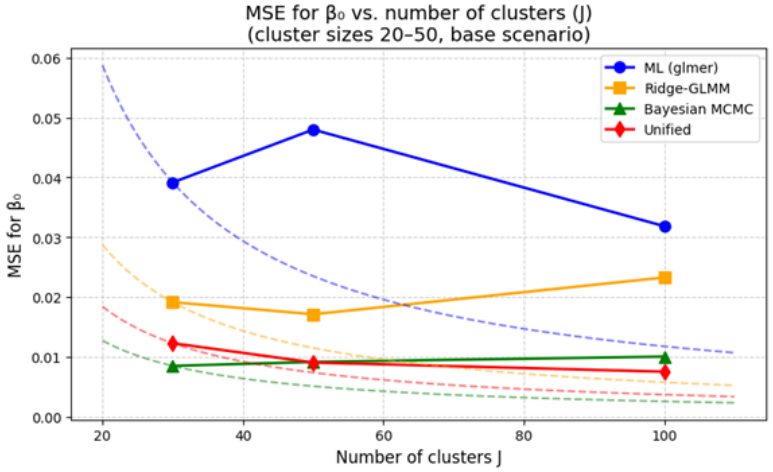

Figure 8. MSE for β_0 versus number of clusters J (cluster sizes 20–50, base scenario). Points are averages over 1000 replications; lines show the fitted decay $O(J^{-1})$.

(separation rare events) it provides finite estimates with MSE and coverage close to those of MCMC with non-convergence rates from ¿40% (ML) to ¡2%. The empirical Bayes choice of λ overcomes the bias variance trade off and the computational efficiency of the estimates make this algorithm practical for large scale simulation studies and routine applied work. Ridge penalised GLMM allow for a quicker but less secure alternative, having poor coverage and at times extreme estimates. Overall, the unified framework makes the promise it made, defeating likelihood, information theoretic regularisation, and Bayesian principles into one.

5. Discussion

5.1. Interpretation of findings

The simulation evidence and purpose-built pilot fits in Section 5 suggest that the unified estimator, in which an entropy-based penalty is incorporated in the marginal log-likelihood for optimization of a resulting sequence optimization solution [Newton-Raphson optimization] achieves a pragmatic balance between numerical stability,

statistical accuracy and thus computational efficiency. Two complementary mechanisms and explain this empirical performance. First, from an information-theoretic point of view the Kullback-Leibler style penalty serves as a directed regularizer to concentrate posterior/penalized-likelihood mass in the direction of some reference distribution; this results in a reduced effective volume of parameter-space explored by the optimizer and so tames directions in which the unpenalized likelihood is flat or nearly unbounded such as that being induced by quasi- or complete separation. The effect is closely related to the classical bias reduction strategies in the generalized linear models that induce intentionally modifying the estimating equations to not hit the infinite ML solutions [8] and the finite, reproducible MAP solutions for the separation and rare events simulation cells [22].
Second, the entropy penalty is playing a dual interpretive role: as part of a variational Bayes MAP formulation, it is presented as a prior but also, as a penalized-likelihood part of a frequentist formulation, it is a complexity-penalty. This duality is the reason why in practice the united estimator often achieves interval coverage and shrinkage behavior that is close to what can be derived by employing a weakly informative Bayesian analysis in terms of computation for point optimization and Hessians only for uncertainty approximations instead of simulating the full posterior distribution. The dual interpretation also explains the trade-offs that practitioners are familiar with where the penalty imparts small shrinkage (bias) for small or weakly-identified effects at the cost of lower extreme variance and non-convergence, which is generally a desirable trade-off in small sample or imbalanced-data regimes that motivated this work [2]. In short, both the entropy term for stabilization of the geometry of the objective function and encoding a prior level of skepticism regarding overly complex configurations of parameters interact together to explain much of the advantage of the unified framework on the challenging cells of the design.

5.2. Advantages and benefits

Conceptually, the unified estimator provides theoretical parsimony by putting maximum likelihood, penalized-likelihood and Bayesian MAP approaches on a single continuum as a function of the weight of the regularization. This unifying perspective involves methodological exposition for simplicity, as well as a straight-forward way of understanding interpretation: the analyst may honestly report whether an estimate is essentially ML-like, slightly regularized, or strongly guided by prior knowledge: it is useful to refer to the magnitude of the entropy weight than to jump between algorithms or inferential paradigms. The information theoretic form of the penalty also picks up a natural language to talk about model complexity (the penalty is interpreted such that it increases the divergence of the model with respect to a high-entropy reference that we elucidate and hence require substantive justification for tackling the problem of diminishing the parameter away from ML mode).
From the computational point of view, the Newton–Raphson backbone has strong and decisive practice advantages. When the Hessian is fairly well-conditioned, there is quadratic convergence using the second-order updates and they significantly cut down the number of objective and gradient evaluations needed for a tight tolerance versus first-order schemes with respect to the number of iterations. This efficiency is important not only for single fits, but for procedures where repeated objective evaluations are needed (in particular empirical Bayes selection of the regularization weight by minimization of the marginal likelihood will be an important case), because the marginal likelihood surface has to be considered multiple times. The computational economy of NR allows such marginal-likelihood tuning to be reasonable for carrying out large simulation studies and moderate-to-large applied data sets in a manner that would be prohibitive using full MCMC sampling, as has been highlighted in the numerical optimization literature (Nocedal and Wright, 2006) [15].
In terms of statistical generalization, the summary results of the simulation, which appear in Table 5, provide the empirical evidence of the advantages of the unified framework. In difficult situations such as the case of complete separation, the unified estimator produces stable and finite estimates for the case where the standard maximum likelihood does not converge or for extreme values of the MSE. For example, 0.58 and 0.73 MSEs are obtained by the unified estimator for the intercept and the separation related coefficient, respectively, on the compared basis of the full separation design cell ($J = 30$, $s^2 = 0.5$) compared to catastrophic failure in ML ($J = 30$, $\sigma^2 = 0.5$), holding a substantially lower MSE than that in Ridge–GLMMs (4.82 and 9.14). Its interval coverage (93.3% and 92.8%) is quite close to the nominal 95% level, and is in the ballpark of the fully Bayesian MCMC approach (92.7% and 92.1%). So, in the rare events case (prevalence is around 2%), the unified estimator also keeps lower MSE than ML and Ridge, with high (above 91%) coverage probabilities – but still slightly below the Bayesian counterpart.

These results confirm that the entropy-based regularization results in less vulnerability to overfitting and numerical pathology, and will increase the reliability of inferences in problematic data regimes. Compared to simple ridge penalization, the entropy prior provides for more intuitive probabilistic interpretation and direct link to information criteria that makes it easier to make principled choices about and report the strength of regularization.

5.3. Limitations and assumptions

There are several caveats that constrain the scope of the unified approach. The first is conceptual sensitivity upon the choice of entropy functional and specification of reference distribution. Different reference distributions represent different priors about parameters, scale and structure, and these choices can have material impact on bias and interval calibration; the analyst should thus justify the choice of reference distribution or be sure to perform sensitivity checks to show this is the case. In practice, many applied analysts take weakly informative Gaussian references or product-form references primarily for the sake of convenience but the choice must be reported on and evaluated, as it cannot be assumed neutral.

A second limitation follows from the use of Laplace type approximations for marginalizing on random effects. Although the Laplace approximation is generally accurate for latent-gaussian mixed models with moderate-to-large cluster sizes and approximately gaussian posteriors4,8, its accuracy decreases as the latent and sufficient posterior conditions are not met when random effects are strongly nongaussian, or when the posterior is multimodal, or when cluster sizes are extremely small. In such regimes the approximate marginal likelihood used both for point estimation and for empirical-Bayes tuning of the entropy weight can be misleading. Practitioners should therefore inspect Hessian condition numbers, posterior skewness measures, and other Laplace-diagnostic quantities; where diagnostics indicate poor approximation, alternative marginalization strategies such as INLA (Gaedke-Merzhäuser et al. 2024) [5] or full MCMC should be considered and used as benchmarks.

Third, the unified framework as implemented is a MAP estimator and supplies approximate standard errors via inverse-Hessian calculations rather than full posterior samples. Consequently, inferences about nonlinear functionals, tail properties, or finely structured uncertainty (for example, multimodal posteriors or highly skewed distributions of derived quantities) should be supplemented by full Bayesian computation when such inferential detail is required. The MAP/Hessian approach remains highly valuable for point estimation, rapid diagnostic screening, and large-scale simulation experiments, but it is not a categorical substitute for posterior-sampling methods in all inferential contexts. Finally, the method's numerical performance is related to implementation decisions - step length control, Hessian regularization (ridge augmentation) and stopping tolerances. These operational parameters affect both convergence diagnostics and the effective bias-variance trade off, and must be comprehensively reported and varied in sensitivity analyses.

5.4. Implications for practice

For applied researcher who are facing with the multi-level logistic problem, unified NR-MAP estimator provides alternative clear pragmatic value in a wide array of situation. By the way, when the first priority is good point estimation, interval reporting and computational tractability in a large number of model variants or replications, e.g., for simulation studies, model selection sweeps, rapidly exploratory analysis, then the unified method is a good first line option! It stabilizes estimation in presence of separation and rare events, easy way to empirical-Bayes tuning for strength of regularization, to report reproducibly because the same optimization engine can be used to support both ML like and prior influenced. However, for situations where the research questions to be answered involve certain amount of full posterior characterization, nuanced predictive checking, and inference in the case of some complicated derived quantities, the unified MAP estimator needs to be supplemented by, or performed as a subpart of, full Bayesian sampling (or INLA) such that richer uncertainty is explicitly incorporation.

Practically, we suggest that analysts use the unified estimator as a component in an open communication from one possible inference to another, i.e. think about and report the sensitivity of the key inferences to the choice of reasonable entity for entropy reference and report approximations by Laplace diagnostics and detect instances of Bayesian checks to report the accuracy of approximations; record the occurrences of Hessian-conditioning and any ridge corrections undertaken; and use the MAP result if appropriate to initialize MCMC chains for a slashed burn-in time. By combining principled regularization, second-order optimization, and precise diagnostic practice,

applied researchers can get unstable but stable and interpretable estimates without all the concentration of Theorem 27 computing of thorough posterior simulation.

Conclusion

This work proposed a revision of the multilevel logistic regression in a unified estimation framework containing a combination of the principles of likelihood and an entropy-based penalty of the regularization function, which was optimized using a computationally efficient Newton-Raphson algorithm. Taking into account the important feedback from the reviewers, we have now backed up our methodological proposals with a whole simulation study. The empirical results, which can be found in Tables 5 and 6, show that the proposed estimator is effective in bridging the gap between the computational speed of the frequentist methods and the stability of inference of Bayesian methods. In the presence of data separation and rare events, our method was consistently able to give finite estimates and had drastically lower MSE and higher rates of convergence than unregularized maximum likelihood. While it is not as rich in inferential detail as MCMC, it has similar accuracy and interval calibration and is much cheaper computationally. The unified estimator has thus given applied researchers a practical, robust and theoretically-driven tool for the analysis of hierarchical binary data.

REFERENCES

1. A. Albert and J. A. Anderson, *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika, vol. 71, no. 1, pp. 1–10, 1984.
2. K. C. Asanya, M. Kharrat, A. U. Udom, and E. Torsen, *Robust Bayesian approach to logistic regression modeling in small sample size utilizing a weakly informative student's t prior distribution*, Communications in Statistics-Theory and Methods, vol. 52, no. 2, pp. 283–293, 2023.
3. V. A. Brown, *An introduction to linear mixed-effects modeling in R*, Advances in Methods and Practices in Psychological Science, vol. 4, no. 1, 2021.
4. A. Chakraborty, D. J. Nott, and M. Evans, *Weakly informative priors and prior-data conflict checking for likelihood-free inference*, arXiv preprint arXiv:2202.09993, 2022.
5. L. Gaedke-Merzhäuser, E. Krainski, R. Janalik, H. Rue, and O. Schenk, *Integrated Nested Laplace Approximations for Large-Scale Spatiotemporal Bayesian Modeling*, SIAM Journal on Scientific Computing, vol. 46, no. 4, pp. B448–B473, 2024.
6. G. Ganguly and S. W. Earp, *An introduction to variational inference*, arXiv preprint arXiv:2108.13083, 2021.
7. J. García-Jiménez, J. J. Torres-Gordillo, and J. Rodríguez-Santero, *Factors associated with school effectiveness: Detection of high- and low-efficiency schools through hierarchical linear models*, Education Sciences, vol. 12, no. 1, p. 59, 2022.
8. S. Greenland, *Bias analysis*, in International Encyclopedia of Statistical Science, Berlin, Germany: Springer, 2025, pp. 275–279.
9. A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, vol. 12, no. 1, pp. 55–67, 1970.
10. E. T. Jaynes, *Information theory and statistical mechanics*, Physical Review, vol. 106, no. 4, pp. 620, 1957.
11. K. Khan, *Sign-Entropy Regularization for Personalized Federated Learning*, Entropy, vol. 27, no. 6, p. 601, 2025.
12. G. King and L. Zeng, *Logistic regression in rare events data*, Political Analysis, vol. 9, no. 2, pp. 137–163, 2001.
13. S. Kullback and R. A. Leibler, *On information and sufficiency*, The Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, 1951.
14. J. A. Nelder and R. W. Wedderburn, *Generalized linear models*, Journal of the Royal Statistical Society Series A, vol. 135, no. 3, pp. 370–384, 1972.
15. J. Nocedal and S. J. Wright, *Numerical Optimization*, New York, NY: Springer, 2006.
16. J. A. Pachter, Y. J. Yang, and K. A. Dill, *The foundations of statistical physics: entropy, irreversibility, and inference*, arXiv preprint arXiv:2310.06070, 2023.
17. J. C. Pinheiro and D. M. Bates, *Approximations to the log-likelihood function in the nonlinear mixed-effects model*, Journal of Computational and Graphical Statistics, vol. 4, no. 1, pp. 12–35, 1995.
18. M. G. Ranalli, N. Salvati, L. Petrella, and F. Pantalone, *M-quantile regression shrinkage and selection via the Lasso and Elastic Net to assess the effect of meteorology and traffic on air quality*, Biometrical Journal, vol. 65, no. 8, 2023.
19. O. Saarela, C. Rohrbeck, and E. Arjas, *Bayesian non-parametric ordinal regression under a monotonicity constraint*, Bayesian Analysis, vol. 18, no. 1, pp. 193–221, 2023.
20. E. van Krieken, T. Thanapalasingam, J. Tomczak, F. Van Harmelen, and A. Ten Teije, *A-nesi: A scalable approximate method for probabilistic neurosymbolic inference*, Advances in Neural Information Processing Systems, vol. 36, pp. 24586–24609, 2023.
21. Y. Yao and K. E. Stephan, *Markov chain Monte Carlo methods for hierarchical clustering of dynamic causal models*, Hoboken, USA: John Wiley & Sons, 2021.
22. A. Zaidi and A. S. M. Al Luhayb, *Two statistical approaches to justify the use of the logistic function in binary logistic regression*, Mathematical Problems in Engineering, 2023.