International Statistical Review

International Statistical Review (2026) doi: 10.1111/insr.70030

The Augmented Hat-Matrix of Hierarchical Generalised Linear Models and Its Use in Leverage Diagnostics

Gianfranco Lovison[1], Mariangela Sciandra[1], Alessandro Albano[1,2] and Chiara Di Maria[1]

[1]*Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, Palermo, Italy*
[2]*Sustainable Mobility Center (Centro Nazionale per la Mobilità Sostenibile–CNMS), Milan, Italy*
Corresponding to: Alessandro Albano, Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, Palermo, Italy. Email: alessandro.albano@unipa.it

Summary

Defining a 'hat-matrix' for a model is essential in many model diagnostic procedures, as it acts as an orthogonal projector from the observation space to the model space. In this paper, we introduce a unique Hat-matrix for the class of hierarchical generalised linear models (HGLMs), which includes, as a special case, the subclass of generalised linear mixed models (GLMMs). We provide a practical discussion on interpreting the hat matrix values in HGLMs across various settings, aimed at assisting practitioners in model diagnostics. Additionally, we propose two new empirical thresholds to identify high-leverage observations and clusters. We demonstrate the advantages of using these empirical thresholds over the traditional approach with a simulation study. Lastly, we present an application to real data to illustrate the effectiveness of our proposed methodology in real-world scenarios.

Key words: hat-matrix; diagnostic tools; HGLM; GLMM; LM.

1 Introduction

The possibility of defining a 'Hat-matrix' for a model, that is, an orthogonal projector from the observation space (or an appropriate transformation of it) to the model space (or an appropriate transformation of it), is of great importance in many diagnostic procedures. The hat-matrix is easily defined and widely used for linear models and generalised linear models since its diagonal values contribute to identifying leverages (Weisberg & Cook, 1982) and potential outliers through the inspection of jacknife residuals (Faraway, 2016).

The extension to the case of mixed models, that is, when both fixed and random parameters are present, is not straightforward. The literature presents some generalisations; for example, one of the first attempts to extend the hat matrix to a more general class of models can be found in Wei *et al.* (1998). The authors proposed the concept of generalised leverage, defined, in terms of derivatives, as the instantaneous rate of change of the i-th predicted value with respect to the j-th response variable. Although Wei *et al.* (1998) showed an application of their approach to

GLMs and did not extend it to mixed-effect model, their idea was subsequently exploited by Demidenko & Stukel (2005) and Nobre & Singer (2011), who proposed the definition of a generalised leverage for linear mixed models. In their setting, the hat matrix can be decomposed into two matrices, one corresponding to the fixed effects, the other to the random effects. Zewotir & Galpin (2007) defined a hat matrix for the conditional residuals instead, but their approach is developed in a setting of known variance components. The more complex the models become, the harder it is to find a formalisation of the hat matrix, and this holds for the case of HGLMs. To the best of our knowledge, the only works explicitly addressing the definition of a hat matrix in the context of hierarchical models are that by Loy & Hofmann (2013), who focused only on linear models, that by Rönnegård *et al.* (2010), which describes their R package `hglm` for fitting and making diagnostics on HGLMs, and the book by Lee *et al.* (2018), who provided theoretical foundations for h-likelihood-based diagnostics within their extended likelihood framework for HGLMs.

The approaches described so far suffer from different drawbacks. Those by Demidenko & Stukel (2005), Nobre & Singer (2011) and Loy & Hofmann (2013) are based on two different hat matrices, separating the fixed and the random effects, while the proposal of Zewotir & Galpin (2007) does not allow for heteroschedasticity. Another limitation, concerning hat matrices defined in Rönnegård *et al.* (2010) and Lee *et al.* (2018), is that their proposal is an oblique projector, and they do not offer decision criteria to perform leverage diagnostics.

The aim of this paper is to contribute to the current literature on leverages in random-effect models, including HGLMs, in a two-fold way: on the one hand, we want to provide a systematic and detailed formalisation of the hat matrix and its properties, currently sparse in several papers and textbooks. On the other hand, we propose a definition of hat matrix for HGLMs that is unique and includes both fixed and random effects, based on the introduction of an 'augmented' response vector (Hodges, 1998; Hodges & Sargent, 2001; Lee *et al.*, 2018). Such a definition, introduced by Lovison & Sciandra (2017), enjoys all the properties of orthogonal projection matrices. The proposed approach fundamentally differs from Demidenko & Stukel (2005), Nobre & Singer (2011) and Loy & Hofmann (2013) as it does not consider two separate matrices for fixed and random effects respectively, but instead provides a unified framework that simultaneously captures leverage information from both model components in a single hat matrix. Building upon the framework established by Rönnegård *et al.* (2010), and later systematised within the broader theoretical framework of HGLM theory and estimation by Lee *et al.* (2018), our method shares their unified approach to handling both fixed and random effects simultaneously. However, their hat matrix formulation is not symmetric, which means it fails to be an orthogonal projector; our approach overcomes this limitation, ensuring that the resulting hat matrix is symmetric and fully satisfies the orthogonal projection properties.

Moreover, we provide a practical discussion on leverage diagnostics, proposing two empirical thresholds to identify high-leverage observations and clusters, along with concrete examples illustrating their implications. In this respect, the proposed thresholds are simple, interpretable and useful for both observations and clusters leverage diagnostics. This represents an advancement over Zewotir & Galpin (2007), which focuses solely on observation-level leverage and relies on complex thresholds involving unknown variance ratios, resulting in 'informal and ad hoc' cutoff points. Furthermore, while Rönnegård *et al.* (2010) introduce leverage measures, they do not offer clear guidelines on how to determine whether an observation or cluster has high leverage, making their practical application less straightforward. Generally, decisions about what constitutes a high-leverage observation or cluster are often based on subjective judgment in the absence of objective criteria (this issue has also been discussed by Singer et al., 2017). Our proposed thresholds address this gap by providing a coherent, mathematically justified, and easily implementable rule.

The paper is organised as follows. Section 2 briefly introduces some notation for the case of data with clustered structure, which is the type of data for which mixed models are used. Section 3 provides a (short) recap on hierarchical generalised linear models, focusing on the joint maximum hierarchical likelihood (MHL) estimation of fixed and random parameters, carried out through the 'augmented GLM' approach, which finally provides the simultaneous Hat-matrix for this wide class of models. Section 4 presents numerical examples aiming at clarifying the practical use of the augmented hat matrix in terms of model diagnostics. We also propose new thresholds based on estimated hat values for identifying high-leverage points and we assessed their performance through a simulation study. An application to real data shows why the single threshold traditionally used in the literature can yield misleading or erroneous results. The conclusions are drawn in Section 5.

2 Data With Clustered Structure

Let the data have the following structure:

$$y_{ij},\, \boldsymbol{x}_{ij},\, \boldsymbol{z}_{ij} \quad i = 1,\, \ldots,\, m; \quad j = 1,\, \ldots,\, n_i;\, n = \sum_{i=1}^{m} n_i \tag{1}$$

where

i is the cluster index

j is the individual (within cluster) unit index

y_{ij} is the response variable

$\boldsymbol{x}_{ij}$ is a vector of p explanatory variables (with fixed parameters)

$\boldsymbol{z}_{ij}$ is a vector of q explanatory variables (with random parameters)

For the ease of presentation, we assume in this paper $n_i = k \;\; \forall i$, so that $n = km$.

We can arrange the data according to the clustered structure as

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i,\,1} \\ \vdots \\ y_{i,\,j} \\ \vdots \\ y_{i,\,k} \end{bmatrix}, \boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_{i,\,1}^T \\ \vdots \\ \boldsymbol{x}_{i,\,j}^T \\ \vdots \\ \boldsymbol{x}_{i,\,k}^T \end{bmatrix}, \boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{z}_{i,\,1}^T \\ \vdots \\ \boldsymbol{z}_{i,\,j}^T \\ \vdots \\ \boldsymbol{z}_{i,\,k}^T \end{bmatrix} \qquad i = 1,\, \ldots,\, m \tag{2}$$

For the subsequent developments, it is convenient to collect all the cluster-level components in vectorised form:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_i \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_i \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} \quad \boldsymbol{Z} = \oplus_{i=1}^{m} \boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{Z}_1 & \ldots & \boldsymbol{O} & \ldots & \boldsymbol{O} \\ \vdots & & \vdots & & \vdots \\ \boldsymbol{O} & \ldots & \boldsymbol{Z}_i & \ldots & \boldsymbol{O} \\ \vdots & & \vdots & & \vdots \\ \boldsymbol{O} & \ldots & \boldsymbol{O} & \ldots & \boldsymbol{Z}_m \end{bmatrix}$$

This representation is called *the stacked form* of the data; here $\boldsymbol{y}$ is an n-dimensional vector, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $n \times p$ and $n \times r$ matrices respectively, where $r = m \times q$.

In this paper, data with clustered structure will be analysed through mixed models, where two types of parameters appear: a p-dimensional vector of fixed parameters $\boldsymbol{\beta}$ and, in each cluster, a q -dimensional vector of random parameters $\boldsymbol{b}_i$, $i = 1, \ldots, m$, which can also be arranged, in stacked form, into an r-dimensional vector:

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_i \\ \vdots \\ \boldsymbol{b}_m \end{bmatrix} \tag{3}$$

3 A Brief Recap of Hierarchical Generalised Linear Models (including Generalised Linear Mixed Models)

Suppose that, conditionally on the random parameters $\boldsymbol{b}_i$, the distribution of y_{ij}, the j-th response in the i-cluster, belongs to the natural exponential class:

$$f(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i) = \exp\left\{\frac{y_{ij}\theta_{ij} - c(\theta_{ij})}{\phi} + d(y_{ij}, \phi)\right\} \forall i, j \tag{4}$$

where θ_{ij} is called the *natural parameter*. Denote $E[Y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i] = c'(\theta_{ij}) = \mu_{ij}$; then $g_\mu(\mu_{ij}) = \eta_{ij}$, for an invertible link function g_μ, and the linear predictor for the j-th observation in the i-th cluster η_{ij} is specified at cluster level as follows:

$$\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{v}_i \qquad i = 1, \ldots, m. \tag{5}$$

Table 1 reports the p.d.f.'s of the four distributions most commonly used in GLM's (Normal, Poisson, Binomial and Gamma) in the Exponential Class representation (4).

In Equation 5, $\boldsymbol{v}_i = g_b(\boldsymbol{b}_i)$, where $\boldsymbol{b}_i$ is a q-vector of random parameters and g_b is a strictly monotone link function which provides the transformation of $\boldsymbol{b}_i$ which combines linearly with the fixed effects $\boldsymbol{\beta}$ in the linear predictor. The rationale behind this choice will become clear later. We assume in the first part of this work that there is only one random parameter (typically,

Table 1. *Distributions belonging to the natural exponential class.*

Distribution	*Natural Parameter* θ	*Dispersion Parameter* ϕ	$c(\theta)$	$E(Y)= c'(\theta)$	*variance Function* $c''(\theta)$	$var(Y)= \phi\, c''(\theta)$
Normal $\mathcal{N}(\mu, \sigma)$	μ	σ^2	$\frac{\theta^2}{2}$	$\mu = \theta$	1	σ^2
Poisson $\mathcal{P}(\lambda)$	$\log(\lambda)$	1	$\exp(\theta)$	$\lambda = \exp(\theta)$	λ	λ
Binomial $\mathcal{B}(\pi; m)$	$\log\left(\frac{\pi}{1-\pi}\right)$	1	$m\log(1+\exp(\theta))$	$m\pi = m\frac{\exp(\theta)}{1+\exp(\theta)}$	$m\pi(1-\pi)$	$m\pi(1-\pi)$
Gamma $\mathcal{G}(\mu, \nu)$	$-\frac{1}{\mu}$	$\frac{1}{\nu}$	$-\log(-\theta)$	$\mu = -\frac{1}{\theta}$	μ^2	$\frac{\mu^2}{\nu}$

an intercept), denoted simply by b on the original scale and by v on the transformed scale. We shall later consider the more general case in which there are q random effects, and such random effects are independent. Generalisation to the case of correlated random effects is relatively easy (see Lee & Nelder 2001), but is beyond the scope of this paper.

In the HGLMs framework the distributional assumption about the random parameter b is more flexible than in LMMs. In particular, the distribution of b is not assumed to be necessarily $\mathcal{N}(0, \sigma_B^2)$, but rather to be any of the *conjugate distributions of the Exponential Class*; this implies (see Cox & Hinkley (1979), p.370) that the p.d.f. of b has the form:

$$f(b; \boldsymbol{\tau}) = \exp[a_1(\boldsymbol{\tau})g_b(b) - a_2(\boldsymbol{\tau})c(g_b(b)) + a_3(\boldsymbol{\tau})] \tag{6}$$

for some vector of parameters $\boldsymbol{\tau}$. For the Normal, Poisson, Binomial and Gamma distributions, the conjugate distributions are Normal, Gamma, Beta and Inverse Gamma respectively. Table 2 summarises such distributions, in the Cox–Hinkley representation (6).

For reasons of identifiability of the fixed effects $\boldsymbol{\beta}$ in the linear predictor (see Lee et al., 2018, p. 168), a further restriction is imposed upon the distributions of the random parameter b: they are parameterised in such a form that their expectation is fixed. In particular, the expectations are $E[b] = 0, 1, \frac{1}{2}, 1$ for the Normal, Gamma, Beta and Inverse Gamma, respectively. Since their expected value is fixed, these distributions depend now only on a scale parameter, denoted by λ, and (6) can be written:

$$f(v; \lambda) = \exp[a_1(\lambda)v - a_2(\lambda)c(v)) + a_3(\lambda)] \tag{7}$$

Table 3 reports the p.d.f.'s and the functions $a_1(\lambda)$, $a_2(\lambda)$, $a_3(\lambda)$ for the four distributions of interest in this paper. Under the aforementioned constraint for $E[b]$, the identities $\lambda a_2(\lambda) = 1$ and $\frac{a_1(\lambda)}{a_2(\lambda)} = E[b] = \psi$ hold, and therefore, by dividing the argument of the exponential by $\lambda a_2(\lambda)$, (7) can be simplified as

$$f(v; \lambda) = \exp\left[\frac{\psi v - c(v)}{\lambda} + d(\psi, \lambda)\right] \tag{8}$$

Here comes what Lee *et al.* (2018) (p. 182) called a 'sleight of hand': the logarithm of (8)

$$\log[f(v; \lambda)] = \ell(\lambda; v) = \frac{\psi v - c(v)}{\lambda} + d(\psi, \lambda) \tag{9}$$

is not an orthodox log-likelihood, due to the presence of the unobserved random parameter v. It can be interpreted, exchanging the roles of is arguments, as the log-likelihood kernel of a GLM for the 'pseudo-response' ψ with 'pseudo-fixed' natural parameter v:

$$\ell(v, \lambda; \psi) = \frac{\psi v - c(v)}{\lambda}. \tag{10}$$

Thanks to this mathematical artifice, we can assume that the 'pseudo-response' ψ is generated by a GLM with $E[\psi] = b$, link function $g_b(b)$, 'pseudo-fixed' natural parameter v, variance function $V_b(b) = \frac{\partial^2 c(v)}{\partial v^2}$ and dispersion parameter λ, so that $Var[\psi] = \lambda V_b(b)$. Table 4 reports these 'pseudo' log-likelihood kernels for the Normal, Gamma, Beta and Inverse Gamma distributions.

Let us now go back to the more general situation where there are q random parameters, arranged at the cluster level in the q-vector $\boldsymbol{b}_i$, or $\boldsymbol{v}_i$ on the $g_b(\boldsymbol{b}_i)$ scale, for $i = 1, \ldots, m$, and such

Table 2. *Conjugate distributions for the natural exponential class.*

Distribution	*p.d.f.*	$a_1(\tau_1, \tau_2)$	$g_b(b) = v$	$a_2(\tau_1, \tau_2)$	$c(g_b(b)) = c(v)$	$a_3(\tau_1, \tau_2)$
Normal $\mathcal{N}(\tau_1, \tau_2)$	$\frac{1}{\sqrt{2\pi\tau_2^2}}\exp\left[\frac{(b - \tau_1)^2}{2\tau_2^2}\right]$	$\frac{\tau_1}{\tau_2^2}$	b	$\frac{1}{\tau_2^2}$	$\frac{b^2}{2} = \frac{v^2}{2}$	$-\frac{\tau_1}{2\tau_2^2} - \log\sqrt{2\pi\tau_2^2}$
Gamma $\mathcal{G}(\tau_1, \tau_2)$	$\frac{\tau_2^{\tau_1} b^{\tau_1 - 1}\exp(-\tau_2 b)}{\Gamma(\tau_1)}$	τ_1	$\log(b)$	τ_2	$b = \exp(v)$	$\tau_1\log(\tau_2) - \log[\Gamma(\tau_1)] - \log(b)$
Beta $\mathcal{BE}(\tau_1, \tau_2)$	$\frac{b^{\tau_1 - 1}(1 - b)^{\tau_2 - 1}}{B(\tau_1, \tau_2)}$	τ_1	$\log\left(\frac{b}{1 - b}\right)$	$\tau_1 + \tau_2$	$-\log(1 - b) = 1 + \exp(v)$	$-\log(B(\tau_1, \tau_2)) - \log[b(1 - b)]$
Inverse Gamma $\mathcal{IG}(\tau_1, \tau_2)$	$\frac{\tau_2^{\tau_1}\exp(-\frac{\tau_2}{b})}{\Gamma(\tau_1) b^{\tau_1 + 1}}$	τ_2	$-\frac{1}{b}$	$\tau_1 - 1$	$\log(b) = -\log(-v)$	$\tau_1\log(\tau_2) - \log[\Gamma(\tau_1)] - 2\log(b)$

International Statistical Review (2026)

Table 3. *Constrained conjugate distributions for the natural exponential class.*

Distribution	*p.d.f.*	$a_1(\lambda)$	$g_b(b) = \nu$	$a_2(\lambda)$	$c(\nu)$	$a_3(\lambda, b)$
Normal	$\frac{1}{\sqrt{2\pi\lambda}}\exp\frac{b^2}{2\lambda}$	0	b	$\frac{1}{\lambda}$	$\frac{b^2}{2}$	$-\log\sqrt{2\pi\lambda}$
$\mathcal{N}(0, \lambda)$						
Gamma	$\frac{\left(\frac{1}{\lambda}\right)^{\frac{1}{\lambda}} b^{\frac{1}{\lambda}} - 1\exp\left(-\frac{b}{\lambda}\right)}{\Gamma\left(\frac{1}{\lambda}\right)}$	$\frac{1}{\lambda}$	$\log(b)$	$\frac{1}{\lambda}$	b	$\frac{1}{\lambda}\log\left(\frac{1}{\lambda}\right) - \log\left[\Gamma\left(\frac{1}{\lambda}\right)\right] - \log(b)$
$\mathcal{G}\left(\frac{1}{\lambda}, \frac{1}{\lambda}\right)$						
Beta	$\frac{b^{\frac{1}{2\lambda}} - 1(1-b)^{\frac{1}{2\lambda}} - 1}{B\left(\frac{1}{2\lambda}, \frac{1}{2\lambda}\right)}$	$\frac{1}{2\lambda}$	$\log\left(\frac{b}{1-b}\right)$	$\frac{1}{\lambda}$	$\log(1-b)$	$-\log\left[B\left(\frac{1}{2\lambda}, \frac{1}{2\lambda}\right)\right] - \log[b(1-b)]$
$\mathcal{BE}\left(\frac{1}{2\lambda}, \frac{1}{2\lambda}\right)$ Inverse Gamma	$\frac{\frac{1}{\lambda}^{1+\frac{1}{\lambda}}\exp\left(-\frac{1}{\lambda b}\right)}{\Gamma\left(1+\frac{1}{\lambda}\right)b^{2+\frac{1}{\lambda}}}$	$\frac{1}{\lambda}$	$-\frac{1}{b}$	$\frac{1}{\lambda}$	$\log(b)$	$\left(1+\frac{1}{\lambda}\right)\log\left(\frac{1}{\lambda}\right) - \log\left[\Gamma\left(1+\frac{1}{\lambda}\right)\right] - 2\log(b)$
$\mathcal{IG}\left(1+\frac{1}{\lambda}, \frac{1}{\lambda}\right)$						

Table 4. *Log-likelihood kernels for the pseudo-response ψ.*

Distribution	*Log-likelihood kernel*	*Pseudo-response ψ*	*Pseudo-natural parameter ν*	$c(\nu)$	$E[\psi] = c'(\nu)$	*Variance function* $V_b(b) = c''(\nu)$
Normal $\mathcal{N}(0, \lambda)$	$\frac{\nu^2}{2\lambda}$	0	b	$\frac{\nu^2}{2} = \frac{b^2}{2}$	ν	1
Gamma $\mathcal{G}\left(\frac{1}{\lambda}, \frac{1}{\lambda}\right)$	$\frac{\nu - \exp(\nu)}{\lambda}$	1	$\log(b)$	$\exp(\nu) = b$	$\exp(\nu) = b$	$\exp(\nu)$
Beta $\mathcal{BE}\left(\frac{1}{\lambda}, \frac{1}{\lambda}\right)$	$\frac{\frac{1}{2}\nu - \log[1+\exp(\nu)]}{\lambda}$	$\frac{1}{2}$	$\log\left(\frac{b}{1-b}\right)$	$\log[1+\exp(\nu)] = -\log(1-b)$	$\frac{\exp(\nu)}{1+\exp(\nu)}$	$\frac{\exp(\nu)}{[1+\exp(\nu)]^2}$
Inverse Gamma $\mathcal{IG}\left(1+\frac{1}{\lambda}, \frac{1}{\lambda}\right)$	$\frac{\nu - [-\log(-\nu)]}{\lambda}$	1	$-\frac{1}{b}$	$-\log(-\nu) = \log(b)$	$-\frac{1}{\nu}$	$\frac{1}{\nu^2} = b^2$

International Statistical Review (2026)

parameters are independent. In what follows, we focus on the $\boldsymbol{v}$ parameters, but analogous results can be obtained on the original $\boldsymbol{b}$ scale, since $f(\boldsymbol{y}_i|\boldsymbol{v}_i; \boldsymbol{\beta}, \phi) = f(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\beta}, \phi)$, thanks to the monotonicity of $g_b()$, and $f(\boldsymbol{v}_i; \lambda) = f(\boldsymbol{b}_i; \lambda)J(\boldsymbol{b})$, where $J(\boldsymbol{b})$ is the Jacobian of the one-to-one transformation $g_b(\boldsymbol{b}) = \boldsymbol{v}$.

The joint density of the responses $\boldsymbol{y}_i$ and the random parameters $\boldsymbol{v}_i$ can be written, for each cluster $i = 1, \ldots, m$:

$$f(\boldsymbol{y}_i, \boldsymbol{v}_i; \boldsymbol{\beta}, \phi, \lambda) = f(\boldsymbol{y}_i|\boldsymbol{v}_i; \boldsymbol{\beta}, \phi)f(\boldsymbol{v}_i; \lambda) \tag{11}$$

Owing to the assumption of conditional independence of the individual observations in each cluster given the random parameters $\boldsymbol{v}_i$, we can write

$$f(\boldsymbol{y}_i|\boldsymbol{v}_i; \boldsymbol{\beta}, \phi) = \prod_{j=1}^{k} f(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{v}_i)$$

where $f(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{v}_i)$ is a density from the natural exponential class (in particular, in this paper, one of those listed in Table 1) and, owing to the assumption of independently distributed random parameters

$$f(\boldsymbol{v}_i; \lambda) = \prod_{u=1}^{q} f(v_{iu}; \lambda) \qquad \forall i$$

where $f(v_{iu}; \lambda)$ is the density function of the random parameter v_{iu}, $u = 1, \ldots, q$, having constrained expected value and hence depending only on the scale parameter λ (in particular, in this paper, one of those listed in Table 3).

Furthermore, using the assumption of between-clusters independence, we can easily obtain from (11) the global joint density of the 'stacked' observations and random parameters $\boldsymbol{y}$ and $\boldsymbol{v}$:

$$f(\boldsymbol{y}, \boldsymbol{v}; \boldsymbol{\beta}, \phi, \lambda, \boldsymbol{v}) = \prod_{i=1}^{m} f(\boldsymbol{y}_i|\boldsymbol{v}_i; \boldsymbol{\beta}, \phi)f(\boldsymbol{v}_i; \lambda) \tag{12}$$

The way in which the distribution for $\boldsymbol{y}_i|\boldsymbol{v}_i$ is combined with the distribution for $\boldsymbol{b}_i$ and with that of $\boldsymbol{v}_i$ through the choice of link $g_b()$, gives rise to different sub-classes of HGLM's

- if each distribution for $\boldsymbol{y}_i|\boldsymbol{v}_i$ is combined with its conjugate for $\boldsymbol{b}_i$, and $g_b = g_\mu$, then the resulting models (Normal-Normal, Poisson-Gamma, Binomial-Beta and Gamma-Inverse Gamma) are called *conjugate hierarchical generalised linear models* (CHGLM's);
- if, whatever the distribution chosen for $\boldsymbol{y}_i|\boldsymbol{v}_i$, $\boldsymbol{b}_i$ is assumed to be $\mathcal{N}(0, \lambda)$, and $g_b(\boldsymbol{b})$ to be the identity link, so that $\boldsymbol{v}$ is itself $\mathcal{N}(0, \lambda)$, then the resulting models (Normal-Normal, Poisson-Normal, Binomial-Normal, Gamma-Normal) are called *generalised linear mixed models* (GLMM's). Notice the special role played by the Normal-Normal case: this is the only GLMM which is also a CHGLM, and actually corresponds to the widely used linear mixed model with Gaussian random components;
- all other combinations of distributions for $\boldsymbol{y}_i|\boldsymbol{b}_i$ from the Natural Exponential Class, distributions for $\boldsymbol{b}_i$ conjugate to the natural exponential class and link functions g_b can be employed: for example, a Binomial distribution can be selected for $\boldsymbol{y}_i|\boldsymbol{b}_i$, a Gamma distribution for $\boldsymbol{b}$ and $g_b(\boldsymbol{b}) = \log(\boldsymbol{b})$: this would still be a HGLM, but neither a CHGLM nor a GLMM. These non-conjugate, and non-GLMM, HGLM's are not widely used, but can be useful in some specific applications: for some examples, see Table 6.2 of Lee *et al.* (2018) book.

The logarithm of (12) is proportional to

$$h(\boldsymbol{y},\ \boldsymbol{v};\ \boldsymbol{\beta},\ \phi,\ \lambda,\ \boldsymbol{v}) = \sum_{i}^{m} \ell(\boldsymbol{\beta},\ \phi;\ \boldsymbol{y}|\boldsymbol{v}_i) + \sum_{i}^{m} \ell(\lambda;\ \boldsymbol{v}_i) \tag{13}$$

which is called the *hierarchical likelihood* of $(\boldsymbol{y},\ \boldsymbol{v})$ by Lee & Nelder (2001). As pointed out earlier, (13) is not a standard log-likelihood, due to the presence of the unobserved random parameters $\boldsymbol{v}_i,\ i = 1,\ \ldots,\ m$; rather, it is a special form of *extended likelihood*, due to the specific scale $\boldsymbol{v} = g_b(\boldsymbol{b})$ with which the random parameters $\boldsymbol{b}$ enter the linear predictor. This scale is particularly important, because it guarantees that the *(weak) canonical property* (see Lee *et al.* (2018), p. 169) holds: according to this property, the random parameters do not interfere with the estimation of the fixed parameters, so that the estimators of $\boldsymbol{\beta}$ obtained maximising the marginal likelihood, that is, the likelihood obtained by integrating out the random parameters $\boldsymbol{v}$ in (13), are the same as those obtained by maximising (13) jointly for $\boldsymbol{\beta}$ and $\boldsymbol{v}$. This is crucial in our effort to obtain a unique, simultaneous, 'hat-matrix' for $\boldsymbol{\beta}$ and $\boldsymbol{v}$ (or $\boldsymbol{b}$). Summarising, joint modelling of $\boldsymbol{y}_i|\boldsymbol{v}_i$ and $\boldsymbol{v}_i$ involves two intertwined GLM's:

1. conditionally on the random parameters $\boldsymbol{v}_i = g_b(\boldsymbol{b}_i)$, the response vector $\boldsymbol{y}_i$ in the i-cluster, is generated by a generalised linear model with $E(\boldsymbol{y}_i|\boldsymbol{v}_i) = \boldsymbol{\mu}_i$, link function $g_\mu(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i$, variance function $\boldsymbol{V}_\mu(\boldsymbol{\mu}_i) = \left\{\text{diag}(V_\mu(\mu_{ij})\right\}$ and dispersion parameter ϕ, so that the dispersion matrix of the observations, conditional on the cluster-specific random parameters $\boldsymbol{v}_i$, is $\mathcal{D}[\boldsymbol{y}_i|\boldsymbol{v}_i] = \phi \boldsymbol{V}_\mu(\boldsymbol{\mu}_i)$.
2. the pseudo-data $\boldsymbol{\psi}_i$, with $E[\boldsymbol{\psi}] = \boldsymbol{b}_i,\ \ i = 1,\ \ldots,\ m$, are also assumed to follow a generalised linear model with link function $g_b(\boldsymbol{b}_i) = \boldsymbol{v}_i$, variance function $\boldsymbol{V}_b(\boldsymbol{b}_i) = \{\text{diag}(V_b(b_i))\}$ and dispersion parameter λ, so that the dispersion matrix of the pseudo-response is $\mathcal{D}[\boldsymbol{\psi}] = \lambda \boldsymbol{V}_b(\boldsymbol{b}_i)$.

3.1 Joint MHL Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{v}$ Though the 'Augmented GLM' Representation

3.1.1 Case 1: ϕ and λ known

Estimation methods for the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{b}$ (or $\boldsymbol{v}$) have been the object of several subsequent refinements and improvements by Lee *et al.* (2018), also in response to various criticisms raised against the use of the h-likelihood (see Kuk and Cheng 1999 and Waddington and Thompson 2004). We refer here to the original proposal contained in Lee & Nelder (1996), which focused on estimating $\boldsymbol{\beta}$ and $\boldsymbol{v}$ simultaneously by jointly maximising the h-likelihood, because this approach, although in some cases is not the optimal one in terms of statistical properties of the derived estimators, has the fundamental advantage of providing a unique hat-matrix for both the fixed and random effects. Let us begin by assuming ϕ and λ known. Then, MHL joint estimation of $\boldsymbol{\beta}$ and $\boldsymbol{v}$ is accomplished as usual by computing and setting to $\boldsymbol{0}$ the score functions for $\boldsymbol{\beta}$ and $\boldsymbol{v}$:

$$\begin{aligned} \frac{\partial h(\boldsymbol{y},\ \boldsymbol{v};\ \boldsymbol{\beta},\ \phi,\ \lambda)}{\partial \boldsymbol{\beta}} &= \boldsymbol{0} \\ \frac{\partial h(\boldsymbol{y},\ \boldsymbol{v};\ \boldsymbol{\beta},\ \phi,\ \lambda)}{\partial \boldsymbol{v}} &= \boldsymbol{0}. \end{aligned} \tag{14}$$

For the purpose of this paper, the key contribution of Lee & Nelder (2001) was to show that, similarly to what happens for linear mixed models, the solutions to (14), that is, the joint

MLE of $\boldsymbol{\beta}$ and the BLUP of $\boldsymbol{v}$ (and hence of $\boldsymbol{b}$), can be obtained as the IWLS solution of a unique 'augmented' GLM. This approach starts from constructing the 'augmented' response vector:

$$\boldsymbol{y}_a = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\psi} \end{bmatrix} \qquad \text{with} : E[\boldsymbol{y}_a] = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{b} \end{bmatrix} = \boldsymbol{\mu}_a$$

and the 'augmented' linear predictor:

$$\boldsymbol{\eta}_a = \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} g_\mu(\boldsymbol{\mu}) \\ g_b(\boldsymbol{b}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{O} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{v} \end{bmatrix} \tag{15}$$

or, more compactly:

$$\boldsymbol{\eta}_a = \boldsymbol{T}\boldsymbol{\gamma}$$

Let us denote

$$\boldsymbol{D}_\mu = \left\{ \text{diag}\left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right\} \qquad \boldsymbol{W}_\mu = \left\{ \text{diag}\left(\frac{\left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2}{V_\mu(\mu_i)} \right) \right\} = \boldsymbol{D}_\mu \boldsymbol{V}_\mu(\boldsymbol{\mu})^{-1} \boldsymbol{D}_\mu \tag{16}$$

$$\boldsymbol{D}_b = \left\{ \text{diag}\left(\frac{\partial b_i}{\partial v_i} \right) \right\} \qquad \boldsymbol{W}_b = \left\{ \text{diag}\left(\frac{\left(\frac{\partial b_i}{\partial v_i} \right)^2}{V_b(b_i)} \right) \right\} = \boldsymbol{D}_b \boldsymbol{V}_b(\boldsymbol{b})^{-1} \boldsymbol{D}_b \tag{17}$$

$$\boldsymbol{s}_\mu = \boldsymbol{\eta} + \boldsymbol{D}_\mu(\boldsymbol{y} - \boldsymbol{\mu}) \qquad \boldsymbol{s}_b = \boldsymbol{v} + \boldsymbol{D}_b(\boldsymbol{\psi} - \boldsymbol{b}) \tag{18}$$

or, in compact form:

$$\boldsymbol{s}_a = \begin{bmatrix} \boldsymbol{s}_\mu \\ \boldsymbol{s}_b \end{bmatrix} \qquad \boldsymbol{W}_a = \begin{bmatrix} \boldsymbol{W}_\mu & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{W}_b \end{bmatrix}$$

Clearly $\boldsymbol{s}_a$ and $\boldsymbol{W}_a$ are theoretical, unobservable quantities, but, similarly to what Firth (1991) did to justify the IWLS algorithm in the context of standard GLM's (see also Lovison 2014), it is useful to do a 'mental experiment' and assume for a while they can be, and have been, observed. Then, it is easily shown that

$$E[\boldsymbol{s}_a] = \boldsymbol{\eta}_a \text{ and } \mathcal{D}[\boldsymbol{s}_a] = \boldsymbol{\Sigma}_a$$

with

$$\boldsymbol{\Sigma}_a = \begin{bmatrix} \phi \boldsymbol{W}_\mu^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \lambda \boldsymbol{W}_b^{-1} \end{bmatrix}$$

so that the following (heteroscedastic) linear model holds for $\boldsymbol{s}_a$:

$$\boldsymbol{s}_a = \boldsymbol{T}\boldsymbol{\gamma} + \boldsymbol{\delta} \qquad \mathcal{D}[\boldsymbol{\delta}] = \boldsymbol{\Sigma}_a$$

Since Σ_a is a full variance/covariance matrix, if $\boldsymbol{s}_a$, $\boldsymbol{W}_a$, and hence $\boldsymbol{\Sigma}_a$ were known, the estimator of $\boldsymbol{\gamma}$ would be the solution of the weighted least squares system of equations:

$$\boldsymbol{T}^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{T}\boldsymbol{\gamma} = \boldsymbol{T}\boldsymbol{\Sigma}_a^{-1}\boldsymbol{s}_a \tag{19}$$

Of course, $\boldsymbol{s}_a$, $\boldsymbol{W}_a$, and hence $\boldsymbol{\Sigma}_a$ are not known, and hence the iterative weighted least squares algorithm must be employed, with the updating formula, at step $t+1$:

$$\boldsymbol{T}^T\boldsymbol{\Sigma}_{a,t}^{-1}\boldsymbol{T}\hat{\boldsymbol{\gamma}}_{t+1} = \boldsymbol{T}\boldsymbol{\Sigma}_{a,t}^{-1}\boldsymbol{s}_{a,t} \tag{20}$$

where

$$\boldsymbol{\Sigma}_{a,t} = \begin{bmatrix} \phi\boldsymbol{W}_{\mu,t}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \lambda\boldsymbol{W}_{b,t}^{-1} \end{bmatrix} \tag{21}$$

and the working vector $\boldsymbol{s}_{a,t}$ and weight matrices $\boldsymbol{W}_{\mu,t}$, $\boldsymbol{W}_{b,t}$ are evaluated at step t.

3.1.2 Case 2: ϕ and λ unknown

In real applications the dispersion parameters ϕ and λ are unknown, and must be estimated. Various estimators have been proposed in the literature, based on the idea of extending ML or REML methods, introduced for linear mixed models, to the hierarchical likelihood of HGLM's. We do not delve into the problems involved in the choice between alternative estimators; for the scope of this paper, it suffices to assume that appropriate estimates $\hat{\phi}$ and $\hat{\lambda}$ are available, and from them an estimate $\hat{\boldsymbol{\Sigma}}_a$ of $\boldsymbol{\Sigma}_a$ can be obtained.

3.2 The Hat-Matrix in the HGLM 'Augmented-Data' Representation

At convergence, the MHLE of $\boldsymbol{\gamma}$ is

$$\hat{\boldsymbol{\gamma}} = [\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\boldsymbol{T}]^{-1}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_a^{-1}\hat{\boldsymbol{s}}_a \tag{22}$$

or, using the extended expressions for $\boldsymbol{T}$, $\boldsymbol{s}_a$, $\boldsymbol{\Sigma}_a$:

$$\hat{\boldsymbol{\gamma}} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{v}} \end{bmatrix} = \begin{bmatrix} \hat{\phi}\boldsymbol{X}^T\hat{\boldsymbol{W}}_\mu^{-1}\boldsymbol{X} & \hat{\phi}\boldsymbol{X}^T\hat{\boldsymbol{W}}_\mu^{-1}\boldsymbol{Z} \\ \hat{\phi}\boldsymbol{Z}^T\hat{\boldsymbol{W}}_\mu^{-1}\boldsymbol{X} & \hat{\phi}\boldsymbol{Z}^T\hat{\boldsymbol{W}}_\mu^{-1}\boldsymbol{Z} + \hat{\lambda}\hat{\boldsymbol{W}}_b^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\phi}\boldsymbol{X}^T\hat{\boldsymbol{W}}_\mu^{-1}\hat{\boldsymbol{s}}_\mu \\ \hat{\lambda}\boldsymbol{Z}^T\hat{\boldsymbol{W}}_b^{-1}\hat{\boldsymbol{s}}_b \end{bmatrix} \tag{23}$$

The 'augmented-data' representation provides an orthogonal projection hat-matrix. From:

$$\hat{\boldsymbol{\eta}}_a = \boldsymbol{T}\hat{\boldsymbol{\gamma}} = \boldsymbol{T}[\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\boldsymbol{T}]^{-1}\boldsymbol{T}\hat{\boldsymbol{\Sigma}}_a^{-1}\hat{\boldsymbol{s}}_a$$

we see that the (unscaled) 'augmented-data' hat-matrix is (Rönnegård *et al.*, 2010; Lee *et al.*, 2018):

$$\hat{\boldsymbol{H}}_a = \boldsymbol{T}[\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\boldsymbol{T}]^{-1}\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1} \tag{24}$$

$\hat{\boldsymbol{H}}_a$ is not symmetric due to the presence of the covariance structure $\hat{\boldsymbol{\Sigma}}_a$, which prevents it from being a proper orthogonal projector. If we consider the scaled version of $\hat{\boldsymbol{s}}_a$ and $\hat{\boldsymbol{\eta}}_a$, $\hat{\boldsymbol{s}}_a^* = \hat{\boldsymbol{\Sigma}}_a^{-1/2}\hat{\boldsymbol{s}}_a$ and $\hat{\boldsymbol{\eta}}_a^* = \hat{\boldsymbol{\Sigma}}_a^{-1/2}\hat{\boldsymbol{\eta}}_a$, we obtain:

$$\hat{\boldsymbol{\eta}}_a^* = \hat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{T}(\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2}\hat{\boldsymbol{s}}_a^*$$

thus, the scaled 'augmented-data' hat-matrix is seen to be

$$\hat{\boldsymbol{H}}_a^* = \hat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{T}(\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2}. \tag{25}$$

The scaled hat-matrix $\hat{\boldsymbol{H}}_a^*$ possesses all the essential properties of an orthogonal projector:

1. Symmetry: $(\hat{\boldsymbol{H}}_a^*)^T = \hat{\boldsymbol{H}}_a^*$
   Proof

$$\begin{aligned}(\hat{\boldsymbol{H}}_a^*)^T &= \left(\hat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{T}(\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2}\right)^T \\ &= (\hat{\boldsymbol{\Sigma}}_a^{-1/2})^T\boldsymbol{T}((\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\boldsymbol{T})^{-1})^T(\boldsymbol{T}^T)(\hat{\boldsymbol{\Sigma}}_a^{-1/2})^T \\ &= \hat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{T}(\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2} = \hat{\boldsymbol{H}}_a^*\end{aligned}$$

□

2. Idempotency: $(\hat{\mathbf{H}}_a^*)^2 = \hat{\mathbf{H}}_a^*$
   Proof

$$\begin{aligned}(\hat{\mathbf{H}}_a^*)^2 &= \left[\hat{\boldsymbol{\Sigma}}_a^{-1/2}\mathbf{T}(\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T})^{-1}\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2}\right] \times \left[\hat{\boldsymbol{\Sigma}}_a^{-1/2}\mathbf{T}(\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T})^{-1}\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2}\right] \\ &= \hat{\boldsymbol{\Sigma}}_a^{-1/2}\mathbf{T}(\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T})^{-1}\underbrace{\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2}\hat{\boldsymbol{\Sigma}}_a^{-1/2}\mathbf{T}}_{=\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T}}(\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T})^{-1}\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2} \\ &= \hat{\boldsymbol{\Sigma}}_a^{-1/2}\mathbf{T}(\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T})^{-1}\left[\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T}\right](\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T})^{-1}\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2} \\ &= \hat{\boldsymbol{\Sigma}}_a^{-1/2}\mathbf{T}(\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T})^{-1}\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1/2} = \hat{\mathbf{H}}_a^*.\end{aligned}$$

□

All eigenvalues of $\hat{\boldsymbol{H}}_a^*$ lie in $[0, 1]$, which follows directly from the symmetry and idempotency properties.

The symmetry and idempotency of $\hat{\boldsymbol{H}}_a^*$ establish it as the orthogonal projection matrix that maps the scaled 'augmented' adjusted response vector $\hat{\boldsymbol{s}}_a^*$ onto the space of the scaled 'augmented' estimated linear predictor $\hat{\boldsymbol{\eta}}_a^*$. In this respect, it can be considered as the closest generalisation of the usual hat-matrix of linear models, generalised linear models and linear mixed models to hierarchical generalised linear models (including generalised linear mixed models, which are a special case of HGLM).

Recall that both the augmented covariance matrix (see Equation 21) and the design matrix decompose conformably into fixed- and random-effect blocks:

$$\hat{\boldsymbol{\Sigma}}_a = \begin{pmatrix}\hat{\boldsymbol{\Sigma}}_{\text{obs}} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_{\text{clust}}\end{pmatrix}, \qquad \mathbf{T} = \begin{pmatrix}\mathbf{X} \\ \mathbf{Z}\end{pmatrix}.$$

Hence,

$$\hat{\boldsymbol{\Sigma}}_a^{-1/2} = \begin{pmatrix}\hat{\boldsymbol{\Sigma}}_{\text{obs}}^{-1/2} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_{\text{clust}}^{-1/2}\end{pmatrix},$$

and all products and inverses in the definition of $\hat{\mathbf{H}}_a^*$ respect this same block structure. In particular, when we combine $\hat{\boldsymbol{\Sigma}}_a^{-1/2}\hat{\boldsymbol{\Sigma}}_a^{-1/2} = \hat{\boldsymbol{\Sigma}}_a^{-1}$, this block-diagonal multiplication simply leads to another block-diagonal matrix, and the middle factor $\mathbf{T}^T\hat{\boldsymbol{\Sigma}}_a^{-1}\mathbf{T}$ expands to $\mathbf{X}^T\hat{\boldsymbol{\Sigma}}_{\text{obs}}^{-1}\mathbf{X} + \mathbf{Z}^T\hat{\boldsymbol{\Sigma}}_{\text{clust}}^{-1}\mathbf{Z}$. Therefore, the scaled 'augmented-data' hat-matrix assumes the form:

$$\hat{\mathbf{H}}_a^* = \begin{pmatrix} \hat{\mathbf{H}}_{\text{obs}}^* & 0 \\ 0 & \hat{\mathbf{H}}_{\text{clust}}^* \end{pmatrix}, \tag{26}$$

where $\hat{\mathbf{H}}_{\text{obs}}^*$ is $n \times n$ (subject-level) and $\hat{\mathbf{H}}_{\text{clust}}^*$ is $r \times r$ (cluster-level).

4 Numerical Experiments

By construction, $\hat{\boldsymbol{H}}_a^*$ contains both *subject-specific* (or individual) hat values denoted as $\hat{h}_{a,jj}$, where $j = 1, \ldots, n$ and *cluster-specific* hat values denoted as $\hat{h}_{a,ii}$, where $i = (n+1), \ldots, (n+r)$.

In order to gain a better understanding of how individual- and cluster-specific hat values change in different settings and interpret them accordingly, we present a set of toy examples with artificial data. Although the data are generated from LMMs for ease of presentation, the results can be extended to HGLMs; these illustrations aid in understanding the outcomes and behavior of the augmented hat matrix in this class of models.

This issue is linked to the task of selecting an appropriate threshold for hat values for identifying high-leverage points since the classical one employed for fixed-effect models may not be appropriate. Thus, we introduce novel thresholds, one for subjects, t_s, and one for clusters, t_c, and examine their performance through a simulation study. The proposed thresholds are based on empirical hat values, considering random and fixed components in HGLMs separately.

4.1 Toy Examples

To investigate the behavior of individual and cluster hat values, we generate data from different scenarios. In each scenario, a linear mixed model with 10 clusters and 10 observations per cluster is fitted, with a single explanatory variable X, and a random intercept. The model can be written as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \varepsilon_{ij}, \;\; i = 1, \ldots, 10, \; j = 1, \ldots, 10 \tag{27}$$

where Y_{ij} is the response for the i-th observation in the j-th cluster, $X_{ij} \sim \mathcal{N}(0,1)$ is the explanatory variable for each observation, $\beta_0 = 0$ and $\beta_1 = 2$ are the fixed regression coefficients, $u_j \sim \mathcal{N}(0,1)$ is the random intercept for the j-th cluster, and $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ is the first-level error.

This model specification entails the presence of no high-leverage observations nor high-leverage clusters.

In the following toy examples we investigate how the inclusion of extreme values in the predictor variable X impacts the hat values. Specifically in each scenario, we manually modify the X_{ij} values of some observations to study how their leverage, as well as that of their respective cluster, varies. To ensure that the modified observations do not become influential, we adjust the corresponding Y_{ij} values according to Equation 27.

Example 1. *In this scenario, we focus on Cluster 1, where one observation, $x_{1,1}$, is assigned a large X-value of 100. This modification strongly affects the leverage of $x_{1,1}$, as its hat value is markedly high, approaching 1. In contrast, the hat values of the remaining observations in the cluster remain much lower, ranging from 0.056 to 0.062, with little variability among them. This difference is*

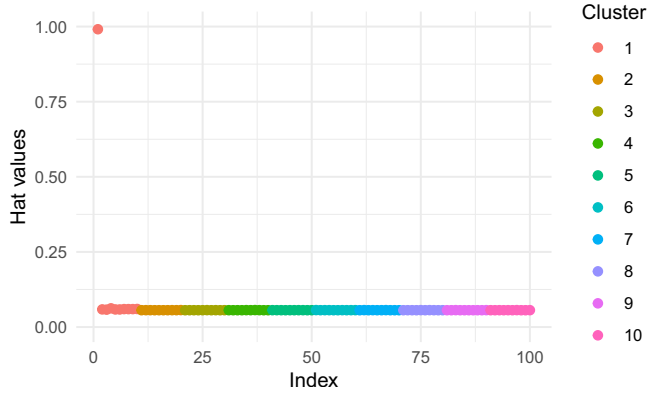


FIGURE 1. *Scenario 1: Subject-specific hat values.*

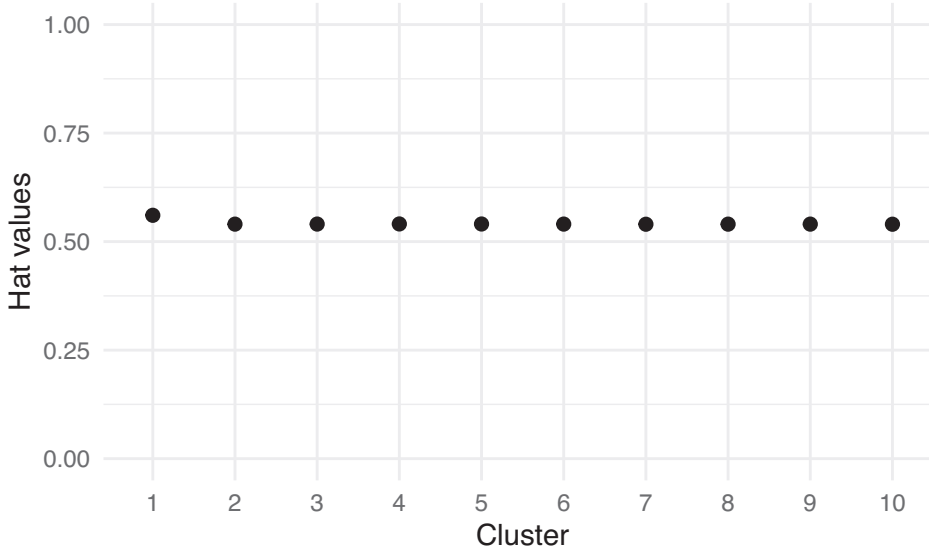


FIGURE 2. *Scenario 1: Cluster-specific hat values.*

clearly visible in Figure 1, where the modified observation stands out sharply compared to the others. Interestingly, when we look at the cluster-specific hat values in Figure 2, we see that the values across clusters are very similar to one another, indicating that the modification of a single observation in one cluster does not have a noticeable effect on the overall cluster-level leverage.

Example 2. *In this scenario, all observations in Cluster 1,* $\{x_{1,1}, \ldots, x_{1,10}\}$*, are assigned an X-value of 100. The results show that although all ten observations in Cluster 1 share this high X-value, none of their individual hat values is particularly large. As a matter of fact, their hat values are all the same, approximately equal to 0.09, while the other subjects' individual hat values range from 0.085 to 0.086 (Figure 3). In contrast, the hat value of Cluster 1 is notably high, very close to 1, indicating that Cluster 1, which includes all the observations with a high X value, can be considered to have high leverage (Figure 4). Compared to Scenario 1, we can see that a subject may be deemed to have high leverage*

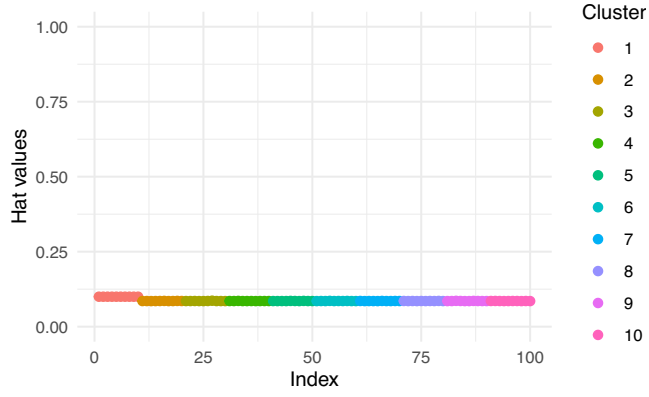


FIGURE 3. *Scenario 2: Subject-specific hat values.*

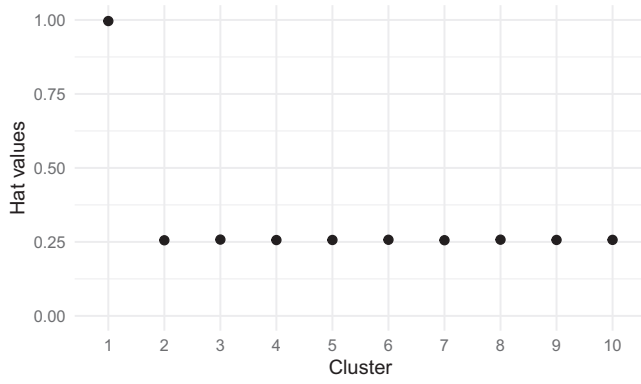


FIGURE 4. *Scenario 2: Cluster-specific hat values.*

when it exhibits a high X value, and it is part of a cluster where observations have standard X values. However, in the particular case where all the other observations within the same cluster share a similar X value, even if it is an extreme value, none of them will be considered individually as having high leverage. Instead, their collective influence 'transfers' the leverage to the entire cluster.

Example 3. *All the observations in Cluster 1, $\{x_{1,1}, \ldots, x_{1,10}\}$, and a single observation in Cluster 2, $x_{2,1}$, are assigned an X value of 100. As a consequence, as already noted in scenario 2, the ten observations in Cluster 1 have standard individual hat values, see Figure 5. Instead, $x_{2,1}$ has a high hat value equal to 0.49. Cluster 1 still has a high hat value equal to 0.638 (Fig 6), but its magnitude is reduced compared to scenario 2. Then, it is important to remark that although $x_{2,1}$ and $x_{1,1} \ldots x_{1,10}$ have exactly the same X value, the hat value of $x_{2,1}$ differs from those of $x_{1,1} \ldots x_{1,10}$. This is due to the fact that the observations in Cluster 1 lose their leverage in favour of their cluster, while $x_{2,1}$ retains it. The behaviour seen in scenario 2 holds, that is, having an extreme X value in a group of observations with the same X yields a different leverage effect than having an extreme X value in a group of standard values.*

Example 4. *All the observations but one in Cluster 1, $\{x_{1,2}, \ldots, x_{1,10}\}$, and a single observation in Cluster 2, $x_{2,1}$, are assigned an X value of 100.*

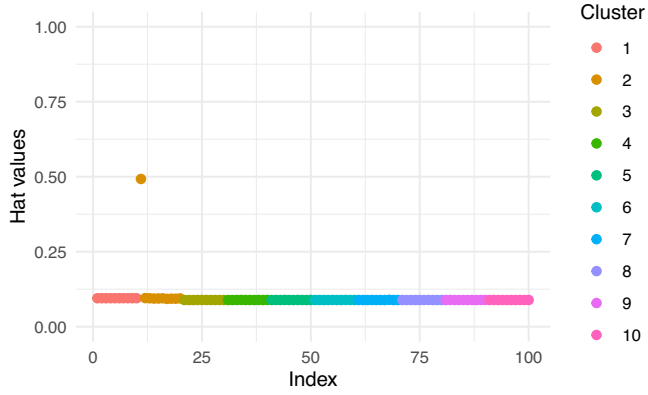


FIGURE 5. *Scenario 3: Subject-specific hat values.*

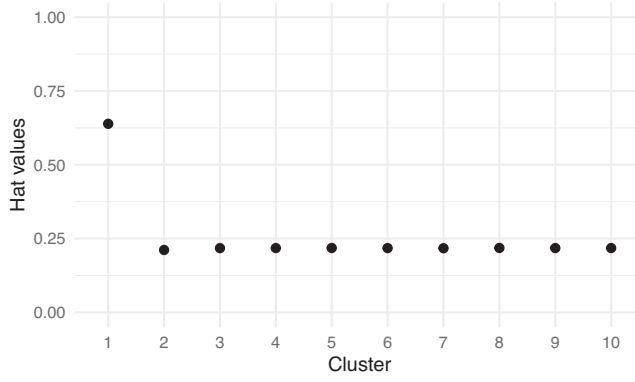


FIGURE 6. *Scenario 3: Cluster-specific hat values.*

In this particular setting, only two individuals, $x_{1,1}$ and $x_{2,1}$, have high individual hat values, equal to 0.339 and 0.392, respectively (Figure 7). Indeed, $x_{1,1}$ is the only subject in Cluster 1 having a standard X value, while the X values of all the other subjects are equal to 100. While, $x_{2,1}$ is the only individual in Cluster 2 having an extreme X value, and all the other subjects have standard X values. In this sense, the individual leverage of a subject in an HGLM can also be seen as a function of its discrepancy with respect to the other individuals in the same cluster. As regards cluster leverage, Cluster 1 has a hat value equal to 0.454 while other clusters range from 0.206 to 0.21, indicating that Cluster 1 leverage effect has been turned down but it is still present (Figure 8).

4.2 Two Empirical Thresholds for Hat Values: A Simulation Study

Identifying high-leverage points through the hat matrix is a widespread approach in model diagnostics. A commonly used rule of thumb is to consider twice the mean of the hat values as a threshold, since, in fixed-effects models, because of the idempotence property of the hat matrix, the mean of the values on the diagonal is always equal to $\frac{p}{n}$.

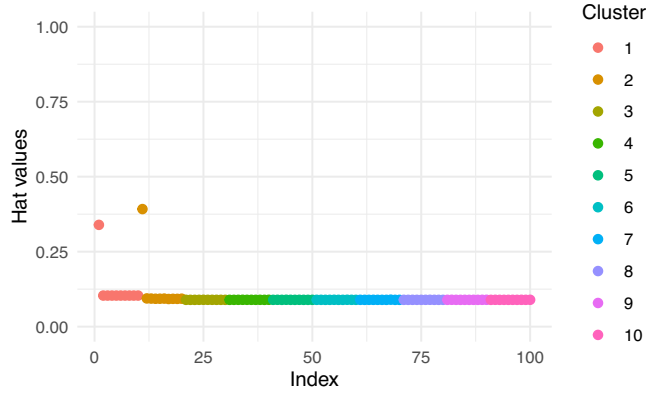


FIGURE 7. *Scenario 4: Subject-specific hat values.*

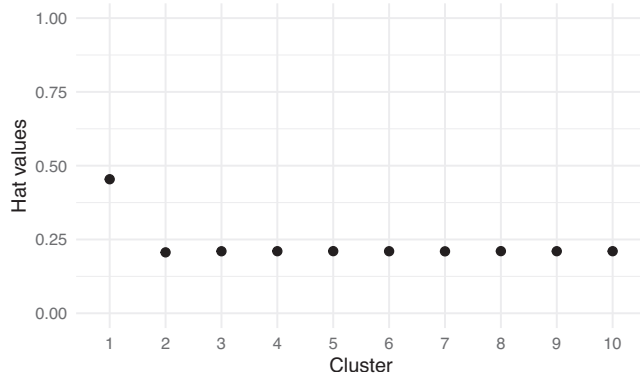


FIGURE 8. *Scenario 4: Cluster-specific hat values.*

However, as demonstrated in previous sections, the augmented hat matrix, $\hat{\boldsymbol{H}}_a^*$, includes subject-specific and cluster-specific hat values. Specifically, considering that the augmented hat matrix $\hat{\mathbf{H}}_a^*$ is a block matrix, then

$$Tr(\hat{\mathbf{H}}_a^*) = Tr(\hat{\mathbf{H}}_{\text{obs}}) + Tr(\hat{\mathbf{H}}_{\text{clust}}) = p + r.$$

Consequently, the classical threshold formula to identify high-leverage points in random-effects models becomes $2(p+r)/(n+r)$. However, we claim that this standard threshold may not be entirely appropriate for such models, as the hat values of the fixed component and the random component might not be directly comparable. While it may be tempting to assume that the traces of the sub-matrices are equal to p and r respectively, this is not true in HGLMs. Therefore, in the absence of a theoretical trace for each sub-matrix, we propose employing the average of the empirical hat values:

$$\overline{h}^{s} = \frac{1}{n}\sum_{j=1}^{n} \hat{h}_{a,jj}, \qquad \overline{h}^{c} = \frac{1}{r}\sum_{i=n+1}^{n+r} \hat{h}_{a,ii}.$$

The trace constraint can be rewritten as

$$Tr(\hat{\mathbf{H}}_a^*) = p + r = n\overline{h}^s + r\overline{h}^c,$$

then

$$\overline{h}^s = p + r - r\frac{\overline{h}^c}{n}.$$

Thus the two averages are linearly linked, and neither can be chosen independently of the other. $\hat{\mathbf{H}}_a^*$ has a fixed total trace and bounded diagonal entries, any increase in leverage for a subset (e.g., certain clusters) forces a decrease elsewhere (e.g., subject-level). Therefore, the proposed thresholds for subjects and clusters, respectively, are

$$t_s = 2\overline{h}^s, \quad t_c = 2\overline{h}^c. \tag{28}$$

Then it can be easily shown that

$$\frac{nt_s + rt_c}{n+r} = 2(n\overline{h}^s + r\frac{\overline{h}^c)}{n+r} = \frac{2(p+r)}{n+r},$$

so that the weighted average of our level-specific thresholds recovers exactly the classical 'twice the average' threshold for the full augmented matrix and ensures that high-leverage diagnostics remain robust without introducing additional assumptions or complexities. This global coherence shows that the proposed thresholds can be considered as a refinement of established leverage diagnostics. As a matter of fact, whenever the random intercepts collapse to fixed constants ($\sigma_u^2 \to 0$), the model reduces to an ANOVA with r fixed effects, so that

$$\lim_{\sigma_u^2 \to 0} \overline{h}^c = 1, \qquad \lim_{\sigma_u^2 \to 0} \overline{h}^s = \frac{p}{n},$$

and hence,

$$\lim_{\sigma_u^2 \to 0} t_c = 2, \quad \lim_{\sigma_u^2 \to 0} t_s = \frac{2p}{n}.$$

Thus each cluster uses up one full degree of freedom, while the observation threshold recovers the classical fixed-effects rule.

To demonstrate the effectiveness of the proposed thresholds, we conduct a simulation study encompassing the three different types of HGLM mentioned in section 3: GLMM, Non-conjugate HGLM, and Conjugate HGLM. The three models chosen are LMM, Normal-Gamma, and Poisson-Gamma.

We generate data from 9 scenarios for each HGLM, of which six are designed to assess the performance of the proposed threshold in the presence of high-leverage observations. At the same time, the remaining three scenarios are utilised to investigate the performance of the proposed threshold in scenarios where high-leverage clusters are present. We make the number of clusters vary over 5, 10 and 15, while the number of observations in each cluster is always fixed at 15. Moreover, for each scenario involving high-leverage observations, we select a different proportion of clusters (0.2, 0.4) in which a high-leverage point is present.

In scenarios involving clusters, just one cluster is high-leverage. In each scenario, one explanatory variable generated from a standard Normal is included, as well as a random intercept. High-leverage observations are included by multiplying the maximum observed X value by a set of multipliers $\alpha = \{1.5, 2, 2.5, 3.5\}$, according to the number of high-leverage points to be generated. The multiplier α can be considered as a factor which amplifies the leverage of an observation.

In the six scenarios involving individuals, increasing values of α are selected when multiple leverage points are present. For instance, in the second scenario with $m = 5$, $n = 75$, and a proportion equal to 0.4, there are two high-leverage points, for which we select the values of $\alpha = 1.5$ and $\alpha = 2.5$, respectively. In the sixth scenario, with $m = 15$, $n = 225$, and a proportion equal to 0.4, there are six high-leverage points, and for each pair, we choose values of $\alpha = 1.5$, $\alpha = 2.5$, and $\alpha = 3.5$. In contrast, $\alpha = 10$ in the three scenarios regarding clusters. Regression coefficients are chosen arbitrarily, and, for each scenario, we generate 500 replicates.

To assess the performance of the proposed thresholds against the traditional threshold $\frac{p+r}{n+r}$, we compute the F1 scores, which is a measure that takes into account the precision and recall of a test (Van Rijsbergen, 1979).

The results, as presented in Table 5, demonstrate that the proposed thresholds, denoted as t_s and t_c, outperform the traditional threshold across almost all scenarios. This indicates that they more accurately identify high-leverage points and clusters compared to the competitor, regardless of the scenario or model.

The only exception is observed in Scenario 1 (m = 5, prop = 0.2), where only a single high-leverage observation exists within one cluster, and in this specific case, the traditional threshold exhibits better performance. As a matter of fact, the traditional threshold is slightly larger than t_s, but notably smaller than t_c (see Table 6 displaying the average thresholds in the scenarios). Consequently, the underperformance of the traditional threshold in identifying high-leverage clusters can be attributed to an increased number of false positives, as the it tends to be overly liberal in this context. Conversely, its slightly lower performance in detecting high-leverage observations is due to false negatives, as the threshold is relatively more conservative in identifying these points.

In sparse scenarios, such as Scenario 1 ($m = 5$, prop = 0.2), the traditional threshold performs well, as there is minimal risk of false negatives. However, as the proportion of high-leverage points increases the proposed thresholds t_s and t_c offer a clear advantage by better balancing precision and recall.

Table 5. *Results of the simulation study comparing the traditional threshold used in the case of fixed models and the two thresholds proposed in this work, in terms of F1 scores. The first six rows concern the scenarios aiming at evaluating the impact of high-leverage points on the hat values, while the last three rows refer to the scenarios with high-leverage clusters. Numbers in bold denote better performance.*

				Norm-Norm		*Norm-Gamma*		*Pois-Gamma*	
Type	*m*	*n*	*Proportion*	$2\frac{p+r}{n+r}$	t_s	$2\frac{p+r}{n+r}$	t_s	$2\frac{p+r}{n+r}$	t_s
Observations	5	75	0.2	**0.978**	0.678	1	1	**0.715**	0.605
Observations	5	75	0.4	0.707	**0.884**	0.732	**0.878**	0.759	**0.798**
Observations	10	150	0.2	0.685	**0.864**	0.701	**0.85**	0.745	0.746
Observations	10	150	0.4	0.656	**0.832**	0.658	**0.766**	0.807	**0.929**
Observations	15	225	0.2	0.76	**0.852**	0.776	**0.82**	0.821	**0.834**
Observations	15	225	0.4	0.542	**0.812**	0.618	**0.784**	0.837	**0.852**

				Norm-Norm		*Norm-Gamma*		*Pois-Gamma*	
Type	*m*	*n*	-	$2\frac{p+r}{n+r}$	t_c	$2\frac{p+r}{n+r}$	t_c	$2\frac{p+r}{n+r}$	t_c
Clusters	5	75	-	0.333	**0.904**	0.398	**0.61**	0.374	**0.491**
Clusters	10	150	-	0.184	**0.946**	0.309	**0.757**	0.309	**0.687**
Clusters	15	225	-	0.135	**0.982**	0.268	**0.668**	0.281	**0.601**

Table 6. *Average thresholds in the nine scenarios.*

				$2\frac{p+r}{n+r}$	t_s		
					Norm-Norm	*Norm-Gamma*	*Pois-Gamma*
Observations	5	75	0.2	0.088	0.055	0.061	0.034
Observations	5	75	0.4	0.088	0.055	0.061	0.072
Observations	10	150	0.2	0.075	0.048	0.058	0.057
Observations	10	150	0.4	0.075	0.047	0.057	0.069
Observations	15	225	0.2	0.071	0.046	0.057	0.068
Observations	15	225	0.4	0.071	0.046	0.057	0.068
				$2\frac{p+r}{n+r}$	t_c		
					Norm-Norm	*Norm-Gamma*	*Pois-Gamma*
Clusters	5	75	-	0.088	0.405	0.496	0.470
Clusters	10	150	-	0.075	0.350	0.258	0.220
Clusters	15	225	-	0.071	0.277	0.181	0.144

4.3 Real Data Application

In this section, we show how the traditional threshold and our proposal can lead to different conclusions by using a real-world data set. The `home_prices` dataset used in this example, available in the `R` package `KingCountyHouses` (Kuhn, 2021), contains several pieces of information on housing prices in King County, Washington State. The dataset comprises 21 613 observations, and three crucial variables that are central to our analysis: *nn_sqft_lot* (numeric), *zip_code* (factor), and *price* (numeric).

The *nn_sqft_lot* variable represents the size of the lot square feet of the 15 houses that are closest in proximity (nearest neighbors) to the house in question, and it acts as explanatory variable. Larger lots might indicate more spacious and potentially desirable neighbourhoods, while smaller lots might indicate denser urban areas. The *zip_code* variable serves as a clustering identifier. It is a categorical variable denoting the zip code of the location where each house is situated, grouping houses together based on their geographical proximity. There are 70 different zip codes. Finally, the *price* variable acts as response. It represents the sale price of each house, and it is reported on the log10 scale.

Figure 9 serves as a tool for identifying potential high-leverage values within the data set, both among individual observations and within different clusters.

Figure 9 highlights the presence of many anomalous observations in the conditional distributions of the *nn_sqft_lot* variable, suggesting that these data points might have high hat values.

In terms of clusters, the figure shows that the cluster associated with the zip code '98039' exhibits, on average, higher values of *nn_sqft_lot* compared to the other clusters. Consequently, we can reasonably expect that this cluster may have a higher hat value compared to the other ones. However, it is essential to further investigate this issue.

Given the nature of the data and the results obtained from the scenarios in the simulation study, we decided to fit a Normal-Gamma model as an example of non-conjugated HGLM. Specifically, the model assumes a linear relationship between the expected price and the predictor *nn_sqft_lot*, and includes a random intercept for each *zip_code*, modeled using a Gamma distribution with a logarithmic link function. The response is assumed to be normally distributed, with expectation:

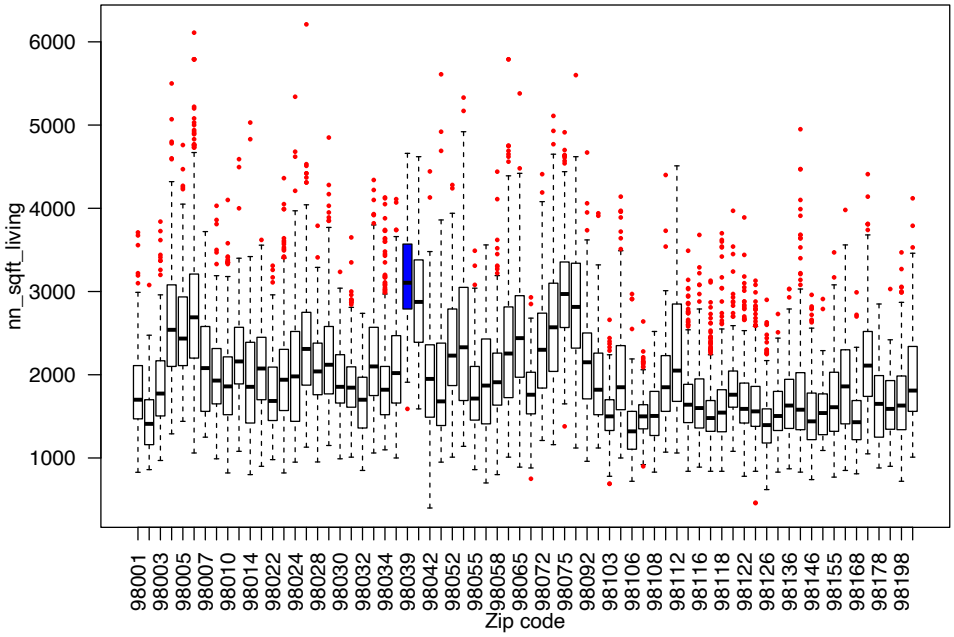
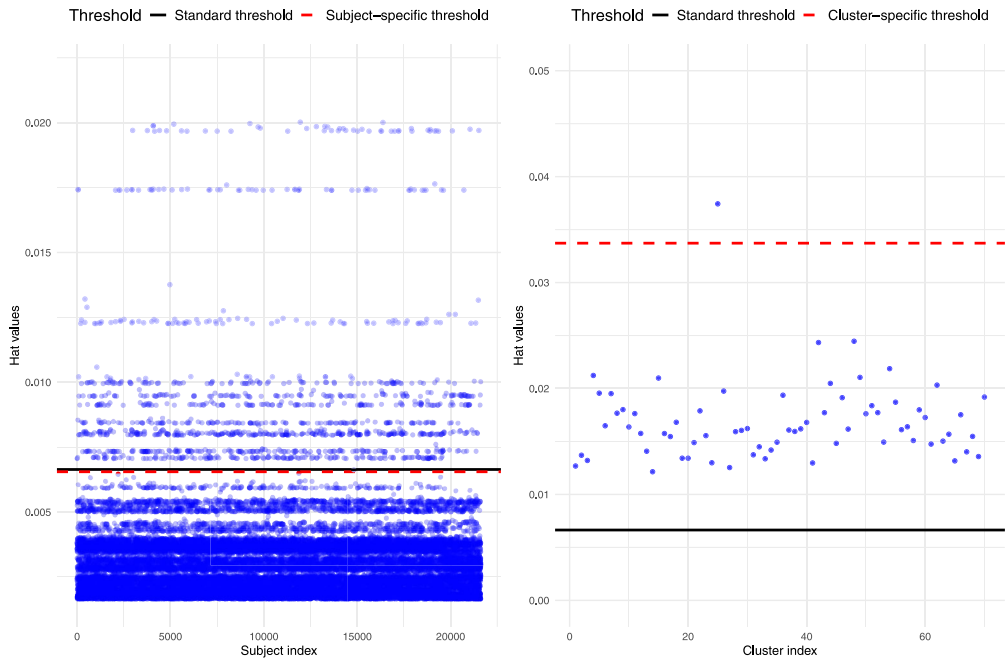


FIGURE 9. *Conditional distributions of nn_sqft_lot with respect to zip_code. Potential outliers within each cluster are highlighted in red, while the potential high-leverage cluster is marked in blue.*

$$E[\text{price}_{ij}] = \beta_0 + \beta_1 \cdot \text{nn}_\text{s}\text{qft}_\text{l}\text{iving}_{ij} + u_j,$$
$$u_j \sim \mathcal{G}\left(\frac{1}{\lambda}, \frac{1}{\lambda}\right).$$

Figure 10 shows subject-specific and cluster-specific hat values, as well as the two different thresholds applied. In Figure 10a, which displays subject-specific hat values, it is evident that the two thresholds are nearly identical. As a result, the same number of high-leverage points are identified using both thresholds. The number of observations classified as high-leverage can be attributed to the substantial variability in the explanatory variable within the dataset. Such variability contributes to the presence of many high-leverage points, particularly for data points that differ significantly from other observations within the same cluster. However, we highlight that high leverage, as identified in our analysis, does not necessarily imply that an observation or cluster is influential on the regression coefficients. Leverage merely indicates a greater potential to influence the results, which warrants further diagnostic investigation.

On the other hand, Figure 10b exhibits cluster-specific hat values, where notable differences between the two thresholds are observed. According to the traditional threshold, all 70 observed clusters are considered high-leverage, which clearly does not make sense and indicates a problem with this approach. This issue is similar to the one encountered in simulations, where the traditional threshold also yielded a higher number of false positives for individual high-leverage clusters. In contrast, using the proposed threshold leads to a much more reliable result, with only cluster '98039' identified as having high leverage. The traditional threshold may not be well-suited for this context, as it can lead to erroneous conclusions, such as incorrectly identifying all clusters as high-leverage. In contrast, the proposed threshold offers a more reliable and accurate assessment of high-leverage clusters, leading to better-informed analyses.



(a) Scenario 1: Subject-specific hat values

(b) Scenario 1: Cluster-specific hat values.

FIGURE 10. *Individual and cluster hat values in the home_prices dataset.*

5 Conclusions

In this work, our primary focus was exploring the use of the hat matrix in the context of HGLMs. The hat matrix plays a crucial role in diagnostics within classical regression settings that involve only fixed effects, and its utility has been well-documented. However, when it comes to models with random effects, the practical use of the hat matrix for diagnostics has been relatively limited, resulting in a lack of practical insights.

The primary challenges revolve around identifying and interpreting the cluster hat values, primarily because of two key reasons. Firstly, there is a lack of well-established thresholds for high-leverage clusters within the existing literature. Secondly, the leverage effect of an entire cluster in a Mixed Effect Model remains a scarcely explored area in previous research. Thus, in this paper, we reviewed the derivation of the augmented hat matrix in HGLMs, which proves to help address the issue of identifying high-leverage points. Subsequently, we showed a scaled version of the hat matrix, which enjoys desirable properties such as symmetry and idempotency.

Additionally, we offered a practical discussion of how to interpret the hat matrix values in various settings of HGLMs, with the goal of facilitating model diagnostics. By doing so, we aimed to provide researchers and practitioners with valuable insights into assessing model fit and identifying high-leverage observations and clusters.

To demonstrate the effectiveness and limitations of existing approaches, we conducted a simulation study. This study revealed the drawbacks of employing a single threshold, that is, the one typically used for fixed-effect models, when applied to HGLMs. In light of this, we proposed novel thresholds based on estimated hat values, which promise to offer suitable and

robust diagnostic tools for HGLMs. The empirical application showed that the use of a single threshold for hat values referring to fixed and random effects can lead to wrong conclusions, suggesting that our proposal is more appropriate in the context of HGLMs.

blackThis work paves the way for several future developments. In particular, while our approach focuses on standard HGLMs with random effects in the mean structure, the theoretical framework of the scaled augmented hat-matrix naturally extends to more complex HGLM formulations. For structured dispersion models (Lee *et al.*, 2018), the augmented representation can be expanded to include dispersion parameters and corresponding design matrix. Similarly, double HGLMs incorporating random effects in both mean and dispersion structures can be accommodated through further expansion of the augmented parameter vector and design matrix. However, these extensions present substantial computational challenges due to increased matrix dimensionality and complex covariance structures, along with theoretical challenges in interpreting multi-component leverage measures. The development of efficient computational algorithms and comprehensive diagnostic frameworks for these advanced structures represents an important direction for future research, building upon the theoretical foundation established in our current work.

In summary, our work aimed at improving the understanding and practical application of the hat matrix in the domain of HGLMs, filling a critical gap in the literature, and paving the way for more effective model diagnostics in complex hierarchical modeling scenarios.

ACKNOWLEDGEMENTS

The research work of Alessandro Albano has been supported by the European Union - NextGenerationEU - National Sustainable Mobility Center (CN00000023) and Italian Ministry of University and Research Decree n. 1033– 17/06/2022, Spoke 2 (CUP B73C2200076000). Open access publishing facilitated by Universita degli Studi di Palermo, as part of the Wiley - CRUI-CARE agreement.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

Cox, D.R. & Hinkley, D.V. (1979) *Theoretical statistics*. Taylor & Francis. ISBN 9780412161605. Available from: https://books.google.it/books?id%3Dppoujo-BInsC

Demidenko, E. & Stukel, T.A. (2005) Influence analysis for linear mixed-effects models. *Statistics in Medicine*, **24**(6), 893–909.

Faraway, J.J. (2016) *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. New York: Chapman and Hall/CRC.

Firth, D. (1991) Generalized linear models. In Hinkley, D.V., Reid, N. & Snell, E.J. (Eds) *Statistical Theory and Modelling*. London: Chapman & Hall.

Hodges, J.S. (1998) Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**(3), 497–536.

Hodges, J.S. & Sargent, D.J. (2001) Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**(2), 367–379.

Kuhn, M. (2021) Kingcountyhouses: data on house sales in King County WA. Available from: https://CRAN.R-project.org/package%3DKingCountyHouses.Rpackageversion0.1.0

Kuk, A.Y.C. & Cheng, Y.W. (1999) Pointwise and functional approximations in Monte Carlo maximum likelihood estimation. *Statistics and Computing*, **9**(2), 91–99.

Lee, Y. & Nelder, J.A. (1996) Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(4), 619–656.

Lee, Y. & Nelder, J.A. (2001) Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **4**(88), 987–1006.
Lee, Y., Nelder, J.A. & Pawitan, Y. (2018) *Generalized linear models with random effects: unified analysis via h-likelihood*. Chapman and Hall/CRC.
Lovison, G. (2014) A note on adjusted responses, fitted values and residuals in generalized linear models. *Statistical Modelling*, **14**(4), 337–359.
Lovison, G. & Sciandra, M. (2017) A recap on linear mixed models and their hat-matrices. *d/SEAS Working Paper*, **18**, 1–24.
Loy, A. & Hofmann, H. (2013) Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics*, **5**(1), 48–61.
Nobre, J.S. & Singer, J.M. (2011) Leverage analysis for linear mixed models. *Journal of Applied Statistics*, **38**(5), 1063–1072.
Rönnegård, L., Shen, X. & Alam, M. (2010) HGLM: a package for fitting hierarchical generalized linear models. *The R Journal*, **2**(2), 20–28.
Singer, J.M., Rocha, F. & Nobre, J. (2017) Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review*, **85**(2), 290–324.
Van Rijsbergen, C.J. (1979) *Information retrieval*, 2nd edition. USA: Butterworth-Heinemann.
Waddington, D. & Thompson, R. (2004) Using a correlated probit model approximation to estimate the variance for binary matched pairs. *Statistics and Computing*, **14**(2), 83–90.
Wei, B.-C., Hu, Y.-Q. & Fung, W.-K. (1998) Generalized leverage and its applications. *Scandinavian Journal of statistics*, **25**(1), 25–37.
Weisberg, S. & Cook, R.D. (1982) *Residuals and influence in regression*. NewYork: Chapman & Hall/CRC.
Zewotir, T. & Galpin, J.S. (2007) A unified approach on residuals, leverages and outliers in the linear mixed model. *Test*, **16**, 58–75.

APPENDIX A: Additional toy examples

Example 5. *All the observations in Cluster 1,*$\{x_{1,1}, \ldots, x_{1,10}\}$ *are assigned an X value of 100. An observation in Cluster 2,*$x_{2,1}$*, is assigned an X value of 50.*
As a result, the ten observations in Cluster 1 have standard individual hat values (Figure A1). Instead, $x_{2,1}$ *has a hat value equal to 0.234, still higher than the rest of the observations, but smaller than in Scenario 3. Inturn, Cluster 1 increases its hat value be ingequal to 0.863 (Figure A2). The insights are similar to Scenario 3, but it is note worthy that reducing the magnitude of* $x_{2,1}$ *attenuates the results obtained in Scenario 3.*

Example 6. *All the observations in Cluster 1 and Cluster 2,*$\{x_{1,1}, \ldots, x_{1,10}\}$ *and* $\{x_{2,1}, \ldots, x_{2,10}\}$*, are assigned an X value of 100. The consequence is that none of the subjects has high leverage*

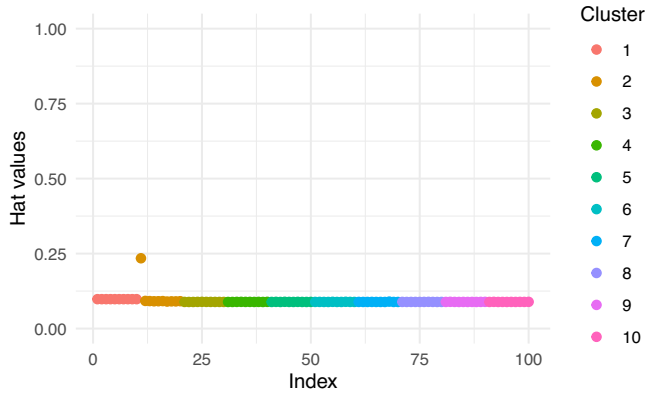


FIGURE A1. *Scenario 5: Subject-specific hat values.*

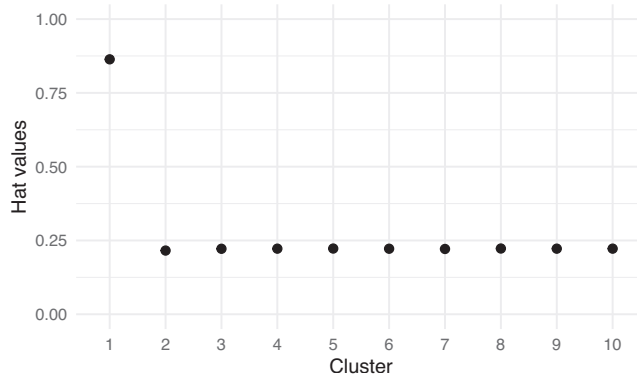


FIGURE A2. *Scenario 5: Cluster-specific hat values.*

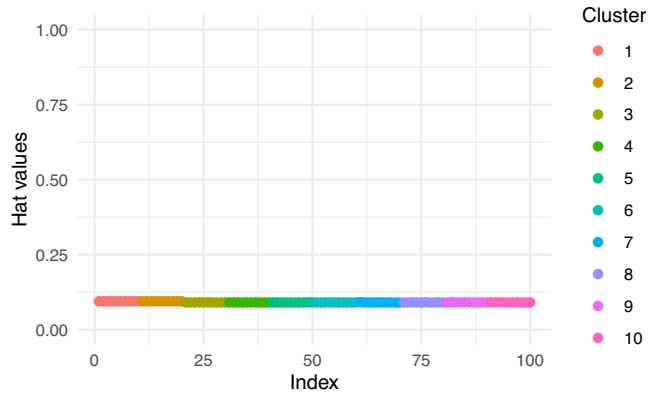


FIGURE A3. *Scenario 6: Subject-specific hat values.*

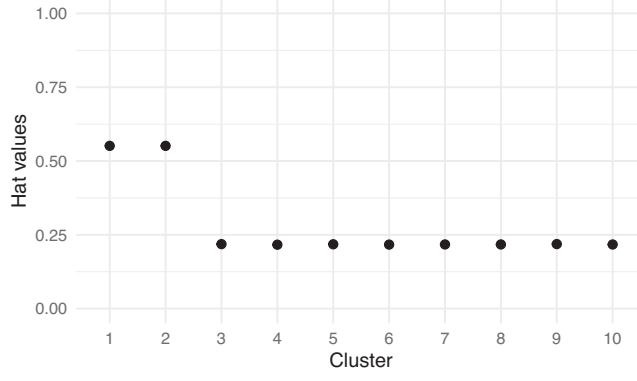


FIGURE A4. *Scenario 6: Cluster-specific hat values.*

(Figure A3). Instead, Cluster 1 and Cluster 2 have high hat values equal to 0.551 (Figure A4). This indicates that the presence of two high-leverage clusters reduced the magnitude of their hat values, as they compete to split the total leverage.

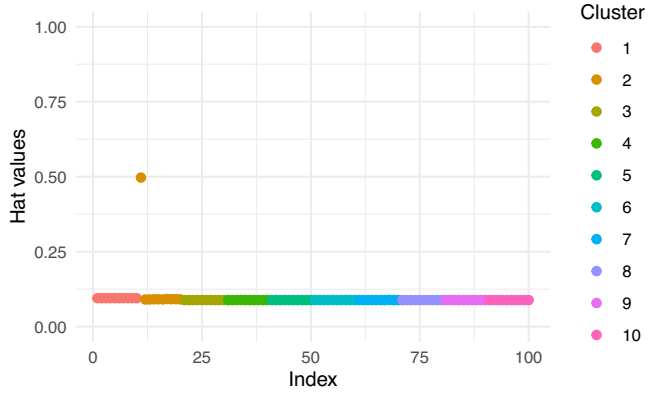


FIGURE A5. *Scenario 7: Subject-specific hat values.*

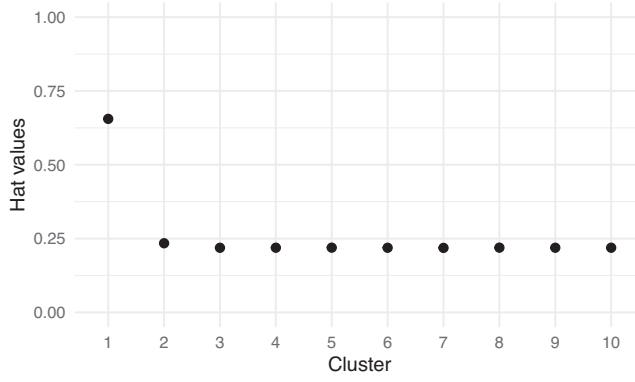


FIGURE A6. *Scenario 7: Cluster-specific hat values.*

Example 7. *All the observations in Cluster 1,* $\{x_{1,1}, x_{1,2}, \ldots, x_{1,10}\}$*, are assigned an X value of 100. The single observation in Cluster 2,*$x_{2,1}$*, has an X value of-100.*
In this setting, only the individual $x_{2,1}$ *has a high individual hat value, equal to 0.497 (Figure A5). Indeed,* $x_{2,1}$ *is the only individual in Cluster 2 with an extreme X value, and all the other subject sin Cluster 2 have standard X values. As for cluster leverage, Cluster 1 has a high hat value of 0.655, while the other clusters'hat values range from 0.218 to 0.234, showing a lower leverage effect of those clusters (Figure A6).*

[*Received May 2024; accepted February 2026*]