Post-Selection Inference for Multiverse Analysis in Mixed-Effects Models (PIMAX)

Angela Andreella[1(✉)] and Anna Vesely[2]

[1] Department of Economics, Ca' Foscari University of Venice, Venice, Italy
angela.andreella@unive.it
[2] Department of Statistics, University of Bologna, Bologna, Italy
https://angeella.github.io/

Abstract. Sign-flipping score tests provide robust inference in generalized linear models under variance misspecification and form the basis of two recently proposed inferential frameworks: Post-selection Inference in Multiverse Analysis (PIMA) and the flip2sss two-stage summary statistics approach. PIMA enables valid inference across a multiverse of model specifications with asymptotic control of the family-wise error rate, whereas flip2sss extends the resampling-based score test to longitudinal and hierarchical data settings. In this paper, we unify these two frameworks by extending multiverse inference to contexts involving dependent observations. Heteroscedasticity, unbalanced designs, and within-cluster dependence are accommodated nonparametrically, without requiring specification of a random-effects correlation structure, as is typically assumed in conventional generalized linear mixed models.

Keywords: Generalized linear mixed model · multiverse analysis · score test · two-stage summary statistic

1 Introduction

High-dimensional data settings, in which many covariates may plausibly explain the response, have intensified the need for flexible and transparent modeling strategies. Multiverse analysis [7] addresses this need by exploring statistical results across a collection of plausible model specifications. However, post-selection inference becomes considerably more intricate when observations exhibit dependence structures because valid inference then hinges on additional assumptions. As model complexity increases, so does the risk of misspecification. For instance, although generalized linear mixed models (GLMMs) provide a flexible framework for dependent data, inference depends critically on the specification of the random-effects structure, which is typically unknown.

Post-selection Inference in Multiverse Analysis (PIMA) [3] establishes asymptotically valid inference across a multiverse of generalized linear models (GLMs)

F. Martella et al. (Eds.): SIS-FENStatS 2026, ISSSAS, pp. 76–82, 2026.
https://doi.org/10.1007/978-3-032-30877-1_13

via sign-flipping score tests [2,4], but assumes independent observations. The subsequently proposed flipscores two-stage summary statistics framework (flip2sss) [1] extends resampling-based score testing to grouped data without requiring explicit specification of the random-effects structure, while retaining robustness to variance misspecification. In this work, we unify these contributions by developing a multiverse inference framework for dependent data. By embedding flip2sss within PIMA, we obtain post-selection inference that accommodates heteroscedasticity, unbalanced designs, and within-cluster dependence nonparametrically, under weaker structural assumptions than conventional approaches such as GLMMs or generalized estimating equations.

The paper is divided as follows. Section 2 introduces the proposed PIMA approach for mixed models, called PIMAX. Section 3 shows the validity and power of the approach with competitive methods through simulations.

2 Methodology

We introduce PIMAX, a unified framework that combines PIMA with flip2sss to enable post-selection inference under dependent observations. PIMAX extends post-selection inference to GLMM-like settings. In particular, the proposed approach retains the robustness to variance misspecification of resampling-based score tests and avoids explicit specification of the random-effects correlation structure, as flip2sss. Being PIMAX based on the PIMA and flip2sss methods, the theoretical assumptions can be retrieved directly from [1,3].

Consider clustered data with clusters indexed by $j = 1, \ldots, J$ and observations within clusters indexed by $i = 1, \ldots, n_j$. Let Y_{ij} follow a GLM with exponential family density, mean $\mu_{ij} = \mathbb{E}(Y_{ij})$, and linear predictor $g(\mu_{ij}) = \eta_{ij}$, where $g(\cdot)$ is a known link function. Interest focuses on testing a regression coefficient β associated with a covariate X, in the presence of nuisance parameters.

Two scenarios arise. In the first, X varies within clusters (within-cluster covariate); in the second, X is constant within clusters (between-cluster covariate). Nuisance covariates may be defined at either the cluster or the observation level, reflecting the potential presence of random effects associated with X and/or the nuisance variables, as commonly considered in GLMM settings.

Let $\mathcal{M} = \{m_1, \ldots, m_K\}$ denote a multiverse of K model specifications. For each specification m_k, consider the null hypothesis $H_k : \beta_k = 0$ associated with the covariate of interest X, and define the global null hypothesis as $H_0 = \bigcap_{k=1}^{K} H_k$. Post-selection inference across $\mathcal{M}$ is then conducted using the PIMA framework [3], based on sign-flipping score tests [4][2] and the closed testing principle [5]. At the same time, to accommodate within-cluster dependence, inference within each model m_k relies on the flip2sss procedure [1], based on a two-stage modeling strategy. Without loss of generality, let W_{ij} denote a within-cluster nuisance covariate and Z_j a between-cluster one.

We start by recalling the flip2ss methodology for a generic specification m_k. In the first scenario, suppose that the covariate of interest X_{ij} varies within clusters. In the first stage, for each cluster j a GLM is fitted:

$$g(\mu_{ijk}) = D_{ijk}\delta_{jk}, \tag{1}$$

using only within-cluster covariates, where $D_{ijk} = [\,1 \;\; X_{ijk} \;\; W_{ij}\,] \in \mathbb{R}^{1\times 3}$ and $\delta_{jk} \in \mathbb{R}^3$. This yields the cluster-specific estimator $\hat{\delta}_{jk} = (\hat{\delta}_{0jk}\,, \hat{\delta}_{1jk}, \hat{\delta}_{2jk})^\top \in \mathbb{R}^3$. Collecting the estimates across clusters gives

$$\hat{\Delta}_k = (\hat{\delta}_{jk})_{j=1}^{J} \in \mathbb{R}^{J\times 3}.$$

In the second stage, inference on β_k is based on a standardized sign-flipping score test [2] applied to the working model

$$[\hat{\Delta}_k]_2 = \beta_k \mathbf{1}_J + Z\gamma + E, \tag{2}$$

where $[A]_2$ denotes the second column of a matrix A, $\mathbf{1}_J \in \mathbb{R}^J$ is a vector of ones, and no parametric assumptions are imposed on the covariance structure of the error term E, beyond zero mean and finite variance.

We now turn to the second scenario, where X_j is constant within clusters. In this case, X_j enters directly in the second stage. The first-stage model (1) is therefore fitted using $D_{ijk} = [\,1 \;\; W_{ij}\,] \in \mathbb{R}^{1\times 2}$ leading to $\hat{\delta}_{jk} = (\hat{\delta}_{0jk}, \hat{\delta}_{1jk})^\top \in \mathbb{R}^2$. Inference on β_k is then based on the second-stage working model

$$[\hat{\Delta}_k]_1 = X\beta_k + Z\gamma + E \tag{3}$$

under the same assumptions of Eq. (2).

We proceed with the definition of the sign-flipping score test, valid in both scenarios (within- and between-cluster covariate of interest). Let $S_k = \sum_{j=1}^{J} u_{jk}$ be the score test statistic for H_k, derived from either Eqs. (2) or (3). Here, u_{jk} denotes the cluster-level score contribution. Then take $f_{j1} = 1$, and $f_{jb} \in \{-1, +1\}$ as independent Rademacher variables for $b \in \{2, \ldots, B\}$. For each sign-flip transformation, define the sign-flipped score statistic as

$$S_{kb} = J^{-1/2} \sum_{j=1}^{J} f_{jb}\, u_{jk},$$

and its standardized version

$$\tilde{S}_{kb} = \frac{S_{kb}}{\sqrt{\widehat{\mathrm{Var}}(S_{kb} \mid f_{1b}, \ldots, f_{Jb})}}$$

where $\widehat{\mathrm{Var}}(\cdot)$ is a consistent variance estimator under H_k [2,4]. Assuming that H_k is tested against a two-sided alternative, denote the model-specific (univariate) test statistic by $|\tilde{S}_{k1}|$.

We conclude with the inference in the multiverse of specifications. For each $b = 1, \ldots, B$, combine the model-specific test statistics using a non-decreasing function $\psi : \mathbb{R}^K \to \mathbb{R}$ [6], yielding the global test statistic $T_b = \psi(|\tilde{S}_{1b}|, \ldots, |\tilde{S}_{Kb}|)$. Common choices of the combining function are the mean

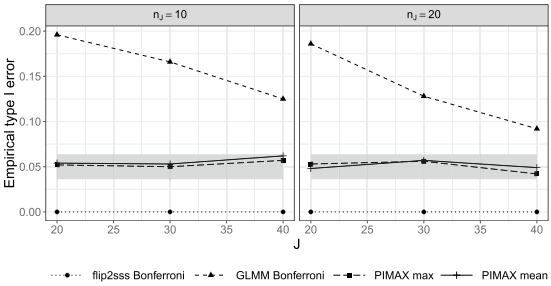

Fig. 1. Empirical type I errors considering $J \in \{20, 30, 40\}$ and $n_J \in \{10, 20\}$ where X is a between-subject variable. Each line corresponds to one method, and the grey area represents the confidence intervals for $\alpha = 0.05$ at level 0.95.

and the maximum. The global p-value is computed as $p = B^{-1} \sum_{b=1}^{B} \mathbb{I}(T_b \geq T_1)$, yielding an asymptotically valid level-α test under heteroscedasticity and within-cluster dependence. A significant global p-value indicates that at least one of the considered models shows a significant effect.

Post-selection inference on an individual hypothesis or a set of hypotheses is obtained via closed testing. For any subset $\mathcal{S} \subseteq \{1, \ldots, K\}$, the intersection hypothesis $\bigcap_{k \in \mathcal{S}} H_k$ is tested using the same construction restricted to the full set of indices, i.e., $T_b(\mathcal{S}) = \psi\big((|Z_{kb}|)_{k \in \mathcal{S}}\big)$. Following the closed testing principle [5], an individual hypothesis H_k is rejected at asymptotic level α if and only if all intersection hypotheses containing H_k are rejected. Using the maximum as combining function yields the maxT shortcut [8], with adjusted p-values

$$p_k^{\text{adj}} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left(\max_{\ell=1,\ldots,K} |\tilde{S}_{\ell b}| \geq |\tilde{S}_{k1}| \right),$$

providing asymptotic strong control of the family-wise error rate. Researchers are able to report the preferred model(s) without inflating the type I error rate.

3 Simulations

We compare the proposed approach with flip2sss and GLMM-based inference, both adjusted for multiplicity using Bonferroni correction. Two scenarios are considered, depending on whether X varies within or between clusters.

Data were generated with number of clusters $J \in \{20, 30, 40\}$, and observations within the clusters $n_J \in \{10, 20\}$. The response variable followed a Bernoulli distribution with parameter $\pi_{ij} \in [0, 1]$, where $\text{logit}(\pi_{ij}) = \eta_{ij}$. Similarly to the simulation setting proposed in [3], a latent variable $L \sim \mathcal{N}(0, 1)$ was introduced to define distinct, yet highly correlated, variables of interest in a multiverse of K specifications. Within each specification, we generated the observed covariate X to have correlation 0.85 with L, and the nuisance covariates W_{ij} and Z_j to have correlation 0.60 with L. Random effects were generated

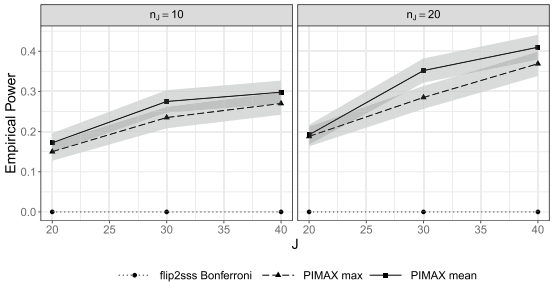

Fig. 2. Estimated power considering $J \in \{20, 30, 40\}$ and $n_J \in \{10, 20\}$ where X is a between-subject variable. Each line corresponds to one method, and the grey areas around the lines represent the confidence intervals at level 0.95.

as $(U_j, G_{1j}, G_{2j})^\top \sim \mathcal{N}(0, \Sigma)$, with $\Sigma = DRD$, $D = \text{diag}(1.0, 0.7, 0.4)$ and R equicorrelation matrix with correlation equal to 0.3. Nuisance parameters were fixed at $\gamma_0 = 0$, $\gamma_1 = 0.5$, and $\gamma_2 = 0.5$.

In the within-cluster scenario, $L = L_{ij}$, $X = X_{ij}$, and

$$\eta_{ij} = \gamma_0 + \gamma_1 W_{ij} + \gamma_2 Z_j + \beta L_{ij} + U_j + G_{1j} L_{ij} + G_{2j} W_{ij}.$$

In the between-cluster scenario, $L = L_j$, $X = X_j$, and

$$\eta_{ij} = \gamma_0 + \gamma_1 Z_j + \gamma_2 W_{ij} + \beta L_j + U_j + G_{2j} W_{ij}.$$

For each simulated dataset, a multiverse of $K = 100$ model specifications was obtained by repeatedly resampling the variable of interest X from the latent variable L, while keeping the remaining data fixed. Inference on β was performed using PIMAX, where the first-stage summary statistics are estimated using Firth-corrected logistic regression, and evidence across specifications was combined using the mean and maxT combining functions. Results were compared with Bonferroni-adjusted p-values from GLMM and original flip2sss. Type I error was evaluated under $\beta = 0$, and power under $\beta = 0.5$, using 1000 replications and $B = 1000$ resampling for each design condition.

Figures 1 and 2 show the empirical type I error and power, respectively, in the case of X_j, while Figs. 3 and 4 in the case of X_{ij}. The GLMM fails to adequately control the type I error rate [1], particularly in small samples, while flip2sss appears markedly conservative. In contrast, PIMAX maintains proper type I error control while retaining competitive power.

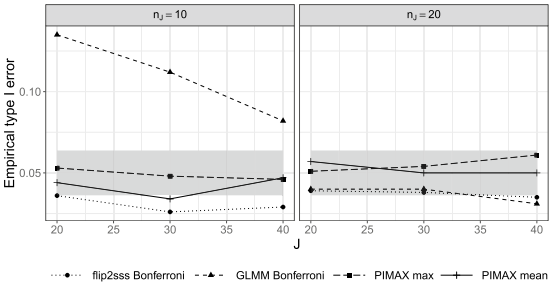

Fig. 3. Empirical type I errors considering $J \in \{20, 30, 40\}$ and $n_J \in \{10, 20\}$ where X is a within-subject variable. Each line corresponds to one method, and the grey area represents the confidence intervals for $\alpha = 0.05$ at level 0.95.

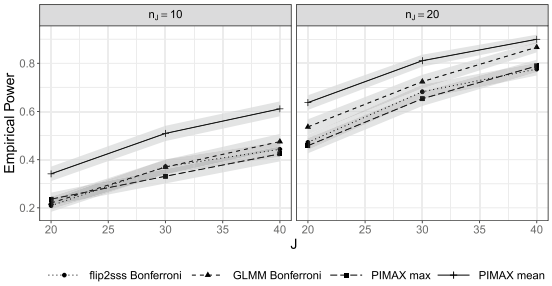

Fig. 4. Estimated power considering $J \in \{20, 30, 40\}$ and $n_J \in \{10, 20\}$ where X is a within-subject variable. Each line corresponds to one method, and the grey areas around the lines represent the confidence intervals at level 0.95.

References

1. Andreella, A., Goeman, J., Hemerik, J., Finos, L.: Robust inference for generalized linear mixed models: a "two-stage summary statistics approach based on score sign flipping. Psychometrika, 1–23 (2025). https://doi.org/10.1017/psy.2024.22
2. De Santis, R., Goeman, J.J., Hemerik, J., Davenport, S., Finos, L.: Inference in generalized linear models with robustness to misspecified variances. J. Am. Stat. Assoc., 1–10 (2025). https://doi.org/10.1080/01621459.2025.2491775
3. Girardi, P., et al.: Post-selection inference in multiverse analysis (PIMA): an inferential framework based on the sign flipping score test. Psychometrika **89**(2), 542–568 (2024). https://doi.org/10.1007/s11336-024-09973-6
4. Hemerik, J., Goeman, J.J., Finos, L.: Robust testing in generalized linear models by sign flipping score contributions. J. R. Stat. Soc. Ser. B Stat Methodol. **82**(3), 841–864 (2020). https://doi.org/10.1111/rssb.12369
5. Marcus, R., Eric, P., Gabriel, K.R.: On closed testing procedures with special reference to ordered analysis of variance. Biometrika **63**(3), 655–660 (1976). https://doi.org/10.1093/biomet/63.3.655
6. Pesarin, F.: Multivariate permutation tests: with applications in biostatistics, vol. 240. Wiley, Chichester (2001)

7. Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W.: Increasing transparency through a multiverse analysis. Perspect. Psychol. Sci. **11**(5), 702–712 (2016). https://doi.org/10.1177/17456916166586
8. Westfall, P.H., Young, S.S.: Resampling-Based Multiple Testing: Examples and Methods for p-value adjustment. John Wiley & Sons (1993)