Empirical likelihood for the analysis of experimental designs

Eunseop Kim[a]*, Steven N. MacEachern[a] and Mario Peruggia[a]

[a]*Department of Statistics, The Ohio State University,*
1958 Neil Ave., Columbus, Ohio 43210, U.S.A.

Empirical likelihood enables a nonparametric, likelihood-driven style of inference without restrictive assumptions routinely made in parametric models. We develop a framework for applying empirical likelihood to the analysis of experimental designs, addressing issues that arise from blocking and multiple hypothesis testing. In addition to popular designs such as balanced incomplete block designs, our approach allows for highly unbalanced, incomplete block designs. We derive an asymptotic multivariate chi-square distribution for a set of empirical likelihood test statistics and propose two single-step multiple testing procedures: asymptotic Monte Carlo and nonparametric bootstrap. Both procedures asymptotically control the generalized family-wise error rate and efficiently construct simultaneous confidence intervals for comparisons of interest without explicitly considering the underlying covariance structure. A simulation study demonstrates that the performance of the procedures is robust to violations of standard assumptions of linear mixed models. We also present an application to experiments on a pesticide.

Keywords: Block design; Bootstrap; Family-wise error rate; Multiple testing; Multivariate chi-square distribution.

AMS Subject Classification: 62G09; 62G10; 62G15; 62G20; 62G30.

1. Introduction

In designed experiments, questions of particular interest frequently involve differences in means of a set of treatments and multiple comparisons. Classical parametric tools for analysis, such as the F-test, Tukey test, or Ryan/Einot–Gabriel/Welsch test, provide efficient ways for testing hypotheses and constructing simultaneous confidence intervals (SCIs) but rely on restrictive assumptions on underlying distributions, variances, and sample sizes. Issues of misspecification and robustness of inference arise when these assumptions are not met. In randomized block designs, rank-based, distribution-free multiple comparison procedures have been suggested, going back to Friedman (1937) and Nemenyi (1963). Mansouri and Shaw (2004) developed a Tukey-type nonparametric pairwise comparison procedure for balanced incomplete block designs. More recently, Eisinga, Heskes, Pelzer, and Te Grotenhuis (2017) proposed an exact test for simultaneous pairwise comparison of Friedman rank sums with a method to quickly calculate the exact p-values and associated statistics. Rank-based approaches, however, have limitations in that they do not fully utilize the available data and have a well-known cycling inconsistency issue (Lehmann and D'Abrera 1975; Fey and Clarke 2012).

Empirical likelihood (Owen 1988) can be helpful as a nonparametric alternative in such situations. With suitably defined estimating functions, empirical likelihood enables nonparametric, likelihood-driven inferences without distributional specifications. It is well established that various forms of empirical likelihood ratio functions admit a nonparametric version of Wilks' theorem

*Corresponding author. Email: kim.7302@osu.edu

under mild conditions, providing a basis for an asymptotic test based on a chi-square null distribution; see, e.g. Qin and Lawless (1994) and Owen (2001). In addition, the empirical distribution of the data determines the shape and orientation of confidence regions. The coverage accuracy of the confidence regions can further be improved by bootstrap or Bartlett-correction (DiCiccio, Hall, and Romano 1991). In the context of the analysis of designed experiments, empirical likelihood has been studied for inference on the median using ranking data by Liu, Yuan, Lin, and Zhang (2012) and Alvo (2015). Inference on the mean is also available by formulating an appropriate estimating function. Designs without a blocking factor, for example, can be analyzed as an analysis of variance problem (Owen 1991). Popular block designs such as randomized complete block designs or balanced incomplete block designs can also be reconfigured as a multivariate mean problem.

The existing literature has mainly focused on establishing limit theorems for a single empirical likelihood (ratio) statistic with a single hypothesis. Wang and Yang (2018) applied F-distribution calibrated empirical likelihood statistics to multiple hypothesis tests, assuming independence between tests. However, these results cannot be directly extended to various dependence scenarios, including the problem of multiple comparisons. Although individual p-values from empirical likelihood tests can be substituted into many existing multiple testing procedures, constructing SCIs for the comparisons based on empirical likelihood has not yet been investigated. This article addresses the challenges of the multiplicity of comparisons by introducing an asymptotic framework for general block designs that leads to manageable inference. In particular, each confidence interval has a variable length that accommodates the underlying covariance structure without explicit studentization, and the SCIs achieve the target coverage probability asymptotically. We also propose empirical likelihood-based multiple testing procedures that rest on this framework. These procedures are generally applicable to other models and estimating functions.

The article is organized as follows. Section 2 introduces some preliminary concepts and conditions used in the rest of the article. Section 3 develops an asymptotic theory for a set of empirical likelihood test statistics. We propose two multiple testing procedures in Section 4 and evaluate the performances of the procedures in Section 5 through a simulation study. An application to pesticide concentration experiments is discussed in Section Section 6. We conclude with a discussion of directions for future research in Section 7. The proofs of the theoretical results are provided in Appendix.

2. Preliminaries

2.1. *General block designs*

A block design is an ordered pair $(\mathcal{T}, \mathcal{B})$ where $\mathcal{T}$ is a set of p points that we call treatments, and $\mathcal{B}$ is a collection of n nonempty subsets of $\mathcal{T}$ called blocks. We consider general block designs where each block size is b_i, with $1 \leq b_i \leq p$, for $i = 1, \ldots, n$. Treatment k is contained in r_k blocks and each pair of distinct treatments k and l is contained in λ_{kl} blocks, for $k, l = 1, \ldots, p$. Then we have the following set of equations:

$$\sum_{i=1}^{n} b_i = \sum_{k=1}^{p} r_k \text{ and } \sum_{l \neq k} \lambda_{kl} = \sum_{i \in \mathcal{B}_k} (b_i - 1) \text{ for each } k,$$

where $\mathcal{B}_k \subseteq \{1, \ldots, n\}$ denotes the index set of the blocks containing treatment k. Let C_n denote the associated $n \times p$ binary incidence matrix with the (i, k) component given by $c_{ik} = 1(i \in \mathcal{B}_k)$, where $1(\cdot)$ is the indicator function of its argument. The ith row sum is then b_i and the kth column sum is r_k.

The n blocks are regarded as random samples from an unknown population. Specifically, we assume independent and identically distributed (i.i.d.) p-dimensional random variables $X_1, \ldots, X_n$ defined on a probability space $(\Omega, \mathcal{F}, P)$ with mean $E(X_1) = \theta_0 \in \text{int}(\theta)$ and positive definite covariance matrix $\text{Var}(X_1) = \Sigma$, where $\Theta \subseteq \mathbb{R}^p$ denotes the parameter space. The parameter of interest is θ_0, the treatment effects. According to the design and C_n, we only observe those X_{ik} from $X_i = (X_{i1}, \ldots, X_{ip})$ for which $c_{ik} = 1$. Since C_n is always available, we do not make a notational distinction between the underlying random variable X_i and its observable components. It will be clear from the context what we are referring to. In order to work with empirical likelihood, we require that $n^{-1} C_n^\top C_n \to D$ as $n \to \infty$ for some matrix D with positive diagonal entries.

2.2. *Empirical likelihood for block designs*

We introduce the general setup for empirical likelihood within the block design framework. The available data are denoted by $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. Inference for θ_0 is based on a p-dimensional estimating function $g(X_i, \theta)$, where $g(X_i, \theta)$ equals $X_i - \theta$ with the unobserved components set to 0. More explicitly, we write

$$g(X_i, \theta) \equiv g(X_i, \theta; c_i) = (X_i - \theta) \circ c_i, \tag{1}$$

where c_i is the ith row of C_n and $\circ$ is the Hadamard product. The (profile) empirical likelihood ratio function, evaluated at θ, is defined as

$$\max_{w_i} \left\{ \prod_{i=1}^n n w_i : w_i > 0, \sum_{i=1}^n w_i = 1, \text{ and } \sum_{i=1}^n w_i g(X_i, \theta) = 0 \right\}.$$

A unique solution exists if the zero vector is contained in $\text{Conv}_n(\theta)$, where $\text{Conv}_n(\theta)$ denotes the interior of the convex hull of $\{g(X_i, \theta) : i = 1, \ldots, n\}$. The Lagrange multipliers $\lambda \equiv \lambda(\theta)$ of the dual optimization problem solve

$$\frac{1}{n} \sum_{i=1}^n \frac{g(X_i, \theta)}{1 + \lambda^\top g(X_i, \theta)} = 0.$$

We denote minus twice the log empirical likelihood ratio function by

$$l_n(\theta) = 2 \sum_{i=1}^n \log \left(1 + \lambda^\top g(X_i, \theta) \right).$$

In the case of $g(X_i, \theta) = X_i - \theta$, Owen (1990) showed that $l_n(\theta_0)$ converges in distribution to χ_p^2, a chi-square distribution with p degrees of freedom. Similar results also hold for other forms of estimating functions (Qin and Lawless 1994), and it can be shown that $l_n(\theta_0) \to \chi_p^2$ in distribution for our general block designs under some regularity conditions. A confidence region for θ_0 can then be constructed as $\{\theta : l_n(\theta) \le \chi_{p,\alpha}^2\}$, where $\chi_{p,\alpha}^2$ is the $(1-\alpha)$th quantile of a χ_p^2 distribution.

For the case of a subset of the parameter vector, let $\theta = (\theta_1, \theta_2)$ for a q-dimensional parameter θ_1 with $q \le p$, and consider testing a hypothesis $H : \theta_1 = \theta_1^*$. Under additional assumptions, a relevant test statistic $l_n(\theta_1^*, \widehat{\theta}_2) \to \chi_q^2$ in distribution, where $\widehat{\theta}_2$ minimizes $l_n(\theta_1^*, \theta_2)$ with respect to θ_2 (Qin and Lawless 1994, Corollary 5). More generally, for a q-dimensional constraint

$h(\theta) = 0$, Qin and Lawless (1995) showed that if $h(\theta_0) = 0$ then $l_n(\widehat{\theta}) \to \chi^2_q$ in distribution, where $\widehat{\theta}$ denotes the minimizer of the problem. Adimari and Guolo (2010) extended hypothesis testing with empirical likelihood to show that the chi-square calibration holds for an even broader class of estimating functions. We can apply these results to general block designs and perform some important tests, including the test of no treatment effect or the interaction between treatments and the blocking variable. The applicability, however, is still restricted to a *single* hypothesis test.

2.3. *Multiple testing*

Consider simultaneously testing m null hypotheses H_j, $j = 1, \ldots, m$. We assume that each H_j corresponds to a nonempty subset of θ through a smooth q_j-dimensional function h_j such that $H_j = \{\theta \in \Theta : h_j(\theta) = 0\}$. We have $\theta_0 \in H_j$ under H_j (when H_j is true). The complete null hypothesis $H_0 = \cap_j H_j$ is also assumed to be nonempty. Then we denote a multiple testing procedure by $\phi = \{\phi_j : j = 1, \ldots, m\}$, where ϕ_j maps the data $\mathcal{X}_n$ into $\{0, 1\}$, and H_j is rejected if and only if $\phi_j = 1$. We restrict our attention to procedures that provide a common cutoff value c_α at a nominal level $\alpha \in (0, 1)$. Given a vector of m test statistics $T_n = (T_{n1}, \ldots, T_{nm})$, we reject H_j if $T_{nj} > c_\alpha$. The total number of false rejections is $V_m = \sum_{j \in \mathcal{I}_0} 1(\phi_j = 1)$, where $\mathcal{I}_0 = \{j : \theta_0 \in H_j\}$ is the index set of true null hypotheses.

Of various Type I error rates for multiple testing, the most common choice in designed experiments is family-wise error rate (FWER). When the number of hypotheses is large, one can consider the generalized family-wise error rate (gFWER) as a less stringent alternative, which is defined as the probability of v or more false rejections for some $v \leq m$. A discussion of procedures for gFWER control can be found in Lehmann and Romano (2005). Formally, a procedure ϕ is said to control gFWER (strongly) at level α if

$$\text{gFWER}_\theta(\phi) = P_\theta(V_m \geq v) \leq \alpha \text{ for all } \theta \in \Theta.$$

When $v = 1$, this reduces to FWER control. We say that ϕ controls gFWER asymptotically if $\limsup_{n\to\infty} \text{gFWER}_\theta(\phi) \leq \alpha$ for all $\theta \in \Theta$. This article addresses single-step procedures for gFWER control with consideration of the *joint* distribution of the empirical likelihood statistics.

3. Asymptotics for multiple testing

3.1. *Multivariate chi-square calibration*

In order to address the multiplicity of our problem and formalize asymptotic multiple testing procedures based on empirical likelihood statistics, we first need a multivariate generalization of chi-square calibration, a multivariate chi-square distribution. The class of multivariate distributions with marginal chi-square distributions is much too broad to be useful in practice, and there is no universal definition of a multivariate chi-square distribution.

In what follows, we adopt a particular type of multivariate chi-square distribution introduced in Dickhaus (2014).

Definition 1 (Dickhaus (2014)) For a vector of positive integers $q = (q_1, \ldots, q_m)$, let $Z_j = (Z_{j1}, \ldots, Z_{jq_j}) \sim N(0, I_{q_j})$, $j = 1, \ldots, m$. Assume that $(Z_1, \ldots, Z_m)$ has a multivariate normal distribution with $\sum_{j=1}^m q_j \times \sum_{j=1}^m q_j$ correlation matrix

$$R = \{\rho(Z_{j_1,l_1}, Z_{j_2,l_2}) : j_1, j_2 = 1, \ldots, m; l_1 = 1, \ldots, q_{j_1}; l_2 = 1, \ldots, q_{j_2}\}.$$

Let $T = (T_1, \ldots, T_m)$, with $T_j = Z_j^\top Z_j \sim \chi^2_{q_j}$. Then T has a multivariate (central) chi-square distribution (of generalized Wishart-type) with parameters m, q, and R. We write $T \sim \chi^2(m, q, R)$.

This distribution naturally arises as a joint limiting distribution of many Wald-type statistics and allows for varying degrees of freedom in each marginal. A comprehensive overview of different types of multivariate chi-square distributions and their applications can be found in Dickhaus and Royen (2015).

We now establish a multivariate extension that covers general block designs as a special case. To this end, we do not require i.i.d. observations X_i and we allow the p-dimensional estimating function $g(X_i, \theta)$ to take forms different from (1). Let θ be a parameter of interest (not necessarily the mean parameter) and define

$$G(\theta) = E\left\{g(X_i, \theta)\right\},\ G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta),\ \text{and}\ S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta) g(X_i, \theta)^\top,$$

with the property that $G(\theta_0) = 0$. Here and throughout, we use $|\cdot|$ to denote the Euclidean norm for vectors. For matrices, $\|\cdot\|$ and $\partial_\theta(\cdot)$ denote the Frobenius norm and the Jacobian matrix, respectively. All limits are taken as $n \to \infty$. We assume the following regularity conditions:

Condition 1 $P\left\{0 \in \text{Conv}_n\left(\theta_0\right)\right\} \to 1$.

Condition 2 $g(X_i, \theta)$ and $G(\theta)$ are continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 almost surely, and $\sup_{\theta \in \mathcal{N}} \|\partial_\theta G_n(\theta) - \partial_\theta G(\theta)\| \to 0$ in probability with nonsingular $\partial_\theta G(\theta_0)$.

Condition 3 There exists a matrix function $V(\theta)$ with positive definite $V(\theta_0)$ such that $\sup_{\theta \in \mathcal{N}} \|S_n(\theta) - V(\theta)\| \to 0$ in probability and $\sup_{|\theta - \theta_0| \le b_n} \|V(\theta) - V(\theta_0)\| \to 0$, for any sequence of positive real numbers $b_n \to 0$.

Condition 4 $a_n G_n(\theta_0) \to U$ in distribution for a sequence of positive real numbers $a_n \to \infty$, where $U \sim N(0, V(\theta_0))$.

Condition 5 $\max_{1 \le i \le n} |g(X_i, \theta_0)| = o_P(a_n)$ and $\max_{1 \le i \le n} \|\partial_\theta g(X_i, \theta_0)\| = O_P(a_n)$.

Condition 6 The function H_j defining the null hypothesis H_j is continuously differentiable on $\mathcal{N}$ with Jacobian matrix $J_j = \partial_\theta h_j(\theta_0)$ of full rank $q_j \le p$, $j = 1, \ldots, m$.

Condition 1 is the basic existence condition for empirical likelihood in the asymptotic setting. Since the computation of $l_n(\theta)$ involves the quadratic forms G_n and S_n, Conditions 2 and 3 are required for $l_n(\theta)$, as a smooth function of θ, to be evaluated in a neighborhood of θ_0. Condition 4 implies that, asymptotically, the quadratic forms have marginal and joint multivariate chi-square distributions. Condition 5 demands that the remainder terms be negligible. Condition 6 completes the statement of the constrained empirical likelihood problems for multiple testing and holds for most practical applications that we consider.

For $j = 1, \ldots, m$, we define the empirical likelihood statistic associated with hypothesis H_j as

$$T_{nj} = \frac{a_n^2}{n} \inf_{\theta \in H_j \cap \overline{K}_n} l_n(\theta), \tag{2}$$

where $\overline{K}_n = \{\theta : |\theta - \theta_0| \le K/a_n\}$ denotes a sequence of closed balls around θ_0 for a given $K > 0$. The m-dimensional test statistic is denoted by $T_n = (T_{n1}, \ldots, T_{nm})$. For notational con-

venience, let $V = V(\theta_0)$, $W = \partial_\theta G(\theta_0)^{-1}$, and $M = WVW^\top$. Then, define $T = (T_1, \ldots, T_m)$, where $T_j = U^\top A_j U$ with $A_j = (J_j W)^\top (J_j M J_j{}^\top)^{-1} (J_j W)$.

THEOREM 3.1 *Under H_0 and Conditions 1–6,*

$$T_n \to T \sim \chi^2(m, q, R)$$

in distribution for some sequence $\overline{K}_n$. Here, $q = (q_1, \ldots, q_m)$ and R is the correlation matrix of $(Z_1, \ldots, Z_m)$, with $Z_j = (J_j M J_j{}^\top)^{-1/2} J_j W U$.

Remark 1 For the general block designs introduced in Section 2.1, it can be shown that l_n is convex in θ so we can find a solution $\widehat{\theta}_c$ of the optimization problem in (2) such that $\widehat{\theta}_c - \theta_0 = O_P(a_n^{-1})$. Thus, the closed ball constraint is not binding asymptotically, i.e. $a_n^2 n^{-1} \inf_{\theta \in H_j} l_n(\theta) = a_n^2 n^{-1} \inf_{\theta : H_j \cap \overline{K}_n} l_n(\theta) + o_P(1)$. In other cases with general estimating functions, identification of θ_0 may require additional conditions, such as compactness of θ, or θ_0 being the unique zero of $G(\theta)$ (see, e.g. Yuan and Jennrich 1998; Jacod and Sørensen 2018).

Remark 2 T_n satisfies the so-called subset pivotality condition (Westfall and Young 1993) asymptotically in the sense that, for any subset $\mathcal{S} \subseteq \{1, \ldots, m\}$ the joint limiting distribution of $\{T_{nj} : j \in \mathcal{S}\}$ remains the same under $\cap_{j \in \mathcal{S}} H_j$ and H_0.

3.2. *Illustration of the theory for general block designs*

We give an illustration of the preceding theory by verifying Conditions 1–5 of Theorem 3.1 for general block designs. Condition 1 holds by applying the Glivenko–Cantelli argument over the half-spaces of Owen (2001, p. 219). Condition 2 is checked by noting that both $g(X_i, \theta)$ and $G(\theta)$ are continuously differentiable with $\partial_\theta G_n(\theta) = -n^{-1}\text{diag}(r_1, \ldots, r_p)$ and $\partial_\theta G(\theta) = -\text{diag}(D)$, where $\text{diag}(\cdot)$ denotes the diagonal matrix of its argument (either a vector or a matrix). The result follows since we have assumed that $n^{-1} C_n^\top C_n \to D$. For Condition 3, observe that

$$\begin{aligned}\|S_n(\theta) - V(\theta)\| \leq \\ \|S_n(\theta) - S_n(\theta_0) - (\theta - \theta_0)(\theta - \theta_0)^\top \circ D\| + \|S_n(\theta_0) - V_n\| + \|V - V_n\|,\end{aligned}$$

where

$$\begin{aligned}S_n(\theta) - S_n(\theta_0) - (\theta - \theta_0)(\theta - \theta_0)^\top \circ D = \frac{1}{n}\sum_{i=1}^n (X_i - \theta_0)(\theta_0 - \theta)^\top \circ c_i c_i^\top \\ + \frac{1}{n}\sum_{i=1}^n (\theta_0 - \theta)(X_i - \theta_0)^\top \circ c_i c_i^\top + (\theta - \theta_0)(\theta - \theta_0)^\top \circ \left(\frac{1}{n} C_n^\top C_n - D\right)\end{aligned}$$

and

$$\|S_n(\theta_0) - V_n\|^2 = \sum_{k=1}^p \sum_{l=1}^p \left| \frac{1}{n} \sum_{i \in \mathcal{B}_k \cap \mathcal{B}_l} (X_{ik} - \theta_{0k})(X_{il} - \theta_{0l}) - \frac{\sigma_{kl}\lambda_{kl}}{n} \right|^2,$$

with the (k, l) component of Σ denoted by σ_{kl}. Then, we have

$$S_n(\theta) - S_n(\theta_0) - (\theta - \theta_0)(\theta - \theta_0)^\top \circ D \to 0 \text{ and } S_n(\theta_0) - V_n \to 0$$

almost surely uniformly in $\theta \in \mathcal{N}$ by the (uniform) law of large numbers, which verifies the first requirement of Condition 3 such that $\sup_{\theta \in \mathcal{N}} |S_n(\theta) - V(\theta)| \to 0$ in probability. For the second requirement,

$$\sup_{|\theta - \theta_0| \le b_n} \|V(\theta) - V\| = \sup_{|\theta - \theta_0| \le b_n} \|(\theta - \theta_0)(\theta - \theta_0)^\top \circ D\| \le p^2 b_n^2 \to 0,$$

establishing Condition 3. For Condition 4, we take $a_n = \sqrt{n}$ and choose any $\epsilon > 0$. Let

$$V_n = \frac{1}{n} \sum_{i=1}^{n} \text{Var}\{g(X_i, \theta_0)\} = E\{S_n(\theta_0)\} = \Sigma \circ \frac{1}{n} C_n^\top C_n.$$

We have $V_n \to V = \Sigma \circ D$ where V is positive definite, and

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} E\left\{|g(X_i, \theta_0)|^2 1\left(|g(X_i, \theta_0)| \ge \epsilon\sqrt{n}\right)\right\} \\ &\qquad \le \frac{1}{n} \sum_{i=1}^{n} E\left\{|X_i - \theta_0|^2 1\left(|X_i - \theta_0| \ge \epsilon\sqrt{n}\right)\right\} \to 0. \end{aligned}$$

It follows from the Lindeberg–Feller central limit theorem that $a_n G_n(\theta_0) \to N(0, V)$ in distribution, and Condition 4 holds with $V(\theta) = V + (\theta - \theta_0)(\theta - \theta_0)^\top \circ D$. Next, a Borel–Cantelli argument (Owen 2001, Lemma 11.2) shows that $\max_{1 \le i \le n} |X_i - \theta_0| = o_P(a_n)$. Since $\max_{1 \le i \le n} \|\partial_\theta g(X_i, \theta_0)\| \le \sqrt{p}$, we have $\max_{1 \le i \le n} \|\partial_\theta g(X_i, \theta_0)\| = o(a_n)$ almost surely and Condition 5 is checked.

4. Empirical likelihood-based multiple testing

4.1. *Asymptotic Monte Carlo*

This section extends the multivariate empirical likelihood theory developed in Section 3 to specific multiple testing procedures for general block designs. We propose two procedures for calibration of the common cutoff value, where the finite-sample null distribution of T_n is approximated by employing appropriate schemes. Both procedures determine cutoffs that provide asymptotic gFWER control (see Remark 2).

As a multivariate analog of chi-square calibration, one may consider relying on multivariate chi-square quantiles of T as a cutoff. In practice, however, the covariance matrix V of U and thus the correlation matrix R of T is rarely known, making it impossible to compute the multivariate quantiles directly. As an alternative, the asymptotic Monte Carlo (AMC) procedure relies on the stochastic representation in Theorem 3.1 to produce a simulation-based approximation to the distribution of T up to any desired precision.

Suppose that we have a consistent estimator $\widehat{\theta}$ of θ_0. It can be shown from Condition 3 that $S_n(\widehat{\theta}) \to V$ in probability; see Hjort, McKeague, and van Keilegom (2009, Remark 2.2). Then, the AMC procedure consists of replacing V with $S_n(\widehat{\theta})$ and simulating samples from the

Data: $\mathcal{X}_n$
Result: cutoff c_α and adjusted p-values $\widetilde{p}_1, \ldots, \widetilde{p}_m$

1. Compute T_n, $S_n(\widehat{\theta})$, and $\widehat{A}_1, \ldots, \widehat{A}_m$
2. Monte Carlo simulation for approximation
   for $b = 1, \ldots, B$ **do**
   Simulate $U_n^{(b)} \sim N(0, S_n(\widehat{\theta}))$
   $\widehat{Q}_{n(v)}^{(b)} \leftarrow$ the vth largest component of $\left({U_n^{(b)}}^\top \widehat{A}_1 U_n^{(b)}, \ldots, {U_n^{(b)}}^\top \widehat{A}_m U_n^{(b)} \right)$
3. $c_\alpha \leftarrow (1-\alpha)$th quantile of $\left\{ \widehat{Q}_{n(v)}^{(1)}, \ldots, \widehat{Q}_{n(v)}^{(B)} \right\}$
4. Adjusted p-values and multiple testing
   for $j = 1, \ldots, m$ **do**
   $$\widetilde{p}_j \leftarrow \frac{1}{B} \sum_{b=1}^{B} 1 \left(\widehat{Q}_{n(v)}^{(b)} \geq T_{nj} \right)$$
   if $T_{nj} > c_\alpha$ or $\widetilde{p}_j < \alpha$ **then** reject H_j

Algorithm 1: AMC

approximate distribution $N(0, S_n(\widehat{\theta}))$. Let $\widehat{A}_j = (J_j W)^\top \{(J_j W) S_n(\widehat{\theta}) (J_j W)^\top\}^{-1} (J_j W)$ and consider random variables $U_n \sim N(0, S_n(\widehat{\theta}))$ and $\widehat{T}_n = ({U_n}^\top \widehat{A}_1 U_n, \ldots, {U_n}^\top \widehat{A}_m U_n)$, defined conditionally on the observed data $\mathcal{X}_n$. With P_n denoting the conditional distribution of $\widehat{T}_n$, the following theorem ensures that the distance between P_n and the distribution of T_n converges to zero in probability.

THEOREM 4.1 *Under H_0 and Conditions 1–6,*

$$\sup_{x \in \mathbb{R}_+^m} \left| P_n(\widehat{T}_n \leq x \mid \mathcal{X}_n) - P\left(T_n \leq x\right) \right| \to 0$$

in probability.

Theorem 4.1 guarantees the asymptotic validity of the AMC procedure described in Algorithm 1. For $v = 1$, the procedure reduces to controlling the asymptotic FWER and the cutoff c_α^{AMC} is computed from the maximum statistics $\{\widehat{Q}_{n(m)}^{(1)}, \ldots, \widehat{Q}_{n(m)}^{(B)}\}$. When the H_js are contrasts of the form $\sum_{k=1}^{p} u_k \theta_k$ with known constants $u_1, \ldots, u_p$, asymptotic $100(1-\alpha)\%$ SCIs for the H_js can be $\{r \in \mathbb{R} : \inf_{h_j(\theta)=r} l_n(\theta) \leq c_\alpha^{\text{AMC}}\}$. The test procedure and the SCIs are compatible, i.e. whenever H_j is rejected, the corresponding interval does not include the null value and vice versa.

Remark 3 Rather than generating draws from an approximate multivariate chi-square distribution, low-dimensional multiplicity-adjusted quantiles can be computed numerically if the underlying correlation matrix fulfills certain structural properties (Stange, Loginova, and Dickhaus 2016).

4.2. *Nonparametric bootstrap*

It has been widely noted that the error rates of tests based on the asymptotic chi-square calibration tend to be higher than the nominal levels, especially in small sample or high-dimensional problems; see, e.g. Qin and Lawless (1994) and Tsao (2004). This issue persists in our setting with multiple empirical likelihood statistics. Moreover, considering the incomplete nature of block designs, convergence to a multivariate chi-square distribution may be slow. As an alternative, Owen (1988) proposed a bootstrap calibration for the mean. Resampling from the original data $\mathcal{X}_n$ yields bootstrap replicates $\mathcal{X}_n^{(b)}$, $b = 1, \ldots, B$. For each $\mathcal{X}_n^{(b)}$, the empirical likelihood statistic $l_n^{(b)}(\overline{X})$ is computed at the sample mean $\overline{X}$ of $\mathcal{X}_n$. The cutoff is obtained as the sample $(1-\alpha)$th quantile of $\{l_n^{(1)}(\overline{X}), \ldots, l_n^{(B)}(\overline{X})\}$.

In our setting, let $\widetilde{\mathcal{X}}_n$ be the null-transformed data with H_0 imposed on the observed data $\mathcal{X}_n$ (see (4) below as an example). Then we denote the (nonparametric) bootstrap samples by $\widetilde{\mathcal{X}}_n^* = \{\widetilde{X}_1^*, \ldots, \widetilde{X}_n^*\}$, where $\widetilde{X}_i^*$, $i = 1, \ldots, n$, are i.i.d. observations from $\widetilde{\mathcal{X}}_n$. Conditional on $\widetilde{\mathcal{X}}_n^*$, we denote the bootstrap empirical likelihood statistic by $l_n^*(\theta)$ and the test statistics by $T_n^* = (T_{n1}^*, \ldots, T_{nm}^*)$, where $T_{nj}^* = a_n^2 n^{-1} \inf_{\theta \in H_j} l_n^*(\theta)$. We establish another consistency result that provides the weak convergence of T_n^* in probability to T. As is customary, we denote the bootstrap distribution conditional on the data by P_n^*.

THEOREM 4.2 *Under H_0 and Conditions 1–6, if $E(|X_i|^4) < \infty$,*

$$\sup_{x \in \mathbb{R}_+^m} \left| P_n^* \left(T_n^* \le x \mid \mathcal{X}_n \right) - P \left(T_n \le x \right) \right| \to 0$$

in probability.

Theorem 4.2 ensures that the conditional distribution of T_n^* approximates the multivariate chi-square distribution of T. Adding the continuity of T implies that the procedures for gFWER control can be asymptotically calibrated by the bootstrap replicates of T_n^*, namely $T_n^{(1)}, \ldots, T_n^{(B)}$.

In Algorithm 2 we describe the nonparametric bootstrap (NB) procedure. It differs from the AMC procedure only in the cutoff c_α^{NB} and the resulting adjusted p-values. Our experience with the procedure shows that the NB procedure is better tuned to the distribution from which the data arise and that c_α^{NB} is typically larger than c_α^{AMC}.

Remark 4 Other bootstrap schemes can be considered as well. The block bootstrap, for instance, may be adapted to produce bootstrap replicates that better preserve the original design structure. This can be of great importance when n is small and the convex hull constraint is of concern. In this regard, it is worth examining the applicability of alternative formulations of empirical likelihood that are free from the constraint (see, e.g. Chen, Variyath, and Abraham 2008; Tsao and Wu 2013).

5. Simulation study

In this section, we carry out a simulation study on a balanced incomplete block design for all pairwise comparisons of treatment effects. The design has five treatments with n blocks, and each block consists of a pair of treatments that appear in $0.1n$ blocks. We have a $(5, n, 0.4n, 2, 0.1n)$-design for short. Finite sample performances of the AMC and the NB procedures are evaluated for controlling FWER and constructing SCIs. We simulate data from the following standard linear

Data: $\mathcal{X}_n$ and $\widetilde{\mathcal{X}}_n$
Result: cutoff c_α and adjusted p-values $\widetilde{p}_1, \ldots, \widetilde{p}_m$

1. Compute T_n
2. Bootstrapping for approximation
   for $b = 1, \ldots, B$ **do**
   Simulate $\widetilde{\mathcal{X}}_n^{(b)}$ from $\widetilde{\mathcal{X}}_n$ and compute $T_n^{(b)}$
   $T_{n(v)}^{(b)} \leftarrow$ the vth largest component of $T_n^{(b)}$
3. $c_\alpha \leftarrow (1-\alpha)$th quantile of $\left\{T_{n(v)}^{(1)}, \ldots, T_{n(v)}^{(B)}\right\}$
4. Adjusted p-values and multiple testing
   for $j = 1, \ldots, m$ **do**
   $$\widetilde{p}_j \leftarrow \frac{1}{B}\sum_{b=1}^{B} 1\left(T_{n(v)}^{(b)} \geq T_{nj}\right)$$
   if $T_{nj} > c_\alpha$ or $\widetilde{p}_j < \alpha$ **then** reject H_j

Algorithm 2: NB

mixed effect model:

$$X_{ik} = \theta_k + \beta_i + \epsilon_{ik} \text{ for } i \in \mathcal{B}_k \text{ and } k = 1, \ldots, 5, \tag{3}$$

where both β_i and ϵ_{ik} are i.i.d. random variables for block effects and errors, respectively. The null hypothesis for treatment pair (k, l) is $H_{kl} : \theta_k - \theta_l = 0$ for $k, l = 1, \ldots, 5$ with $k < l$. We denote the pairwise differences between treatment effects by δ_j for $j = 1, \ldots, 10$, with the corresponding hypothesis H_j and test statistic T_{nj}.

For comparison, we consider the single-step procedure proposed by Hothorn, Bretz, and Westfall (2008) as a benchmark. This procedure (henceforth HBW) is based on an asymptotic multivariate normal distribution for the point estimates and a consistent plug-in estimate of the associated covariance matrix. We apply HBW to restricted maximum likelihood estimates of δ_j, assuming the additive form and compound symmetry that are present in the model. We refer to Hothorn et al. (2008) for technical details.

We fix the level α at 0.05 throughout the simulations. The β_i and ϵ_{ik} are simulated from three different pairs of scenarios:

S1-1. $\beta_i \sim N(0, 1)$ and $\epsilon_{ik} \sim N(0, 1)$;
S1-2. $\beta_i \sim N(0, 0.1), \epsilon_{ik} \sim N(0, 1)$ for $k = 1, 2, 3, 4,$ and $\epsilon_{i5} \sim N(0, 9)$;
S2-1. $\beta_i \sim$ Gamma$(2, 1)$ and $\epsilon_{ik} \sim t(6)$;
S2-2. $\beta_i \sim$ Gamma$(10, 0.1), \epsilon_{ik} \sim t(6)$ for $k = 1, 2, 3, 4,$ and $\epsilon_{i5} \sim U(-5, 5)$;
S3-1. $\beta_i \sim U(-0.5, 0.5)$ and $\epsilon_{ik} \sim U(-0.5, 0.5)$;
S3-2. $\beta_i \sim U(-0.1, 0.1), \epsilon_{ik} \sim U(-0.5, 0.5)$ for $k = 1, 2, 3, 4,$ and $\epsilon_{i5} \sim t(3)$.

where Gamma$(2, 1)$ denotes a Gamma distribution with shape parameter 2 and scale parameter 1. Each pair of scenarios has a distinct distributional specification. In each pair, the first scenario is

of the form (3). The second scenario, however, has negligible block effects and larger variance for the fifth treatment, breaking some assumptions of the model. In each scenario, we consider three different numbers of blocks $n \in \{50, 100, 200\}$ and three different values of θ to vary the number of true null hypotheses. Given specific values of n and θ, simulation results for the AMC procedure are obtained as follows. For $S = 10{,}000$ simulation runs indexed by s:

Step 1 Simulate data from the given scenario and compute $T_n(s)$.

Step 2 With $B = 10{,}000$, apply the AMC procedure in Algorithm 1 to obtain $c_\alpha^{\text{AMC}}(s)$, and compute the SCI $I_j^{\text{AMC}}(s)$ and its length $|I_j^{\text{AMC}}(s)|$ for each j.

The empirical FWER, average length (AL), and coverage probability (CP) of the SCIs are calculated as

$$\text{FWER} = \frac{1}{S}\sum_{s=1}^{S} 1\left(\max\left\{T_{nj}(s) : j \in \mathcal{I}_0\right\} > c_\alpha^{\text{AMC}}(s)\right);$$

$$\text{AL} = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{10}\sum_{j=1}^{10}\left|I_j^{\text{AMC}}(s)\right|\right);$$

$$\text{CP} = \frac{1}{S}\sum_{s=1}^{S} 1\left(\delta_j \in I_j^{\text{AMC}}(s) \text{ for all } j\right).$$

The results for the NB procedure are obtained similarly. Step (i) is modified to set up bootstrap sampling that respects H_0. Before drawing the bootstrap replicates, pass from X_{ik} to

$$\widetilde{X}_{ik} = X_{ik} - \overline{X}_k, \tag{4}$$

where $\overline{X}_k = r_k^{-1}\sum_{i\in\mathcal{B}_k} X_{ik}$ is the maximum empirical likelihood estimate for θ_k. Applying Algorithm 2 in Step (ii), we obtain the same T_n but the cutoff c_α^{NB} is different from c_α^{AMC}, which produces different SCIs, FWER, AL, and CP. All simulations are performed in R (Team 2023). We implement AMC and NB with the **melt** package (Kim 2022). For HBW, we fit (3) via the **lme4** package (Bates, Mächler, Bolker, and Walker 2015) and then pass the result to the **multcomp** package (Hothorn et al. 2008).

Tables 1–3 summarize the simulation results. In all scenarios and procedures, the FWER is largest for all n when H_0 holds and decreases as the number of false hypotheses increases since there are fewer opportunities to reject the true null hypotheses. By construction, AL and CP are not related to θ. The intervals are shorter when the model generates less variation in the data. FWER and CP approach their respective targets, 0.05 and 0.95, under H_0 as n increases. For AMC, FWER and CP are quite far from the targets when $n = 50$ and are also sensitive to the distribution of the data. As can be seen from Table 2, the FWER and CP tend to be worse when the distribution is highly skewed and has a thick tail. The estimates computed with only 20 observations per treatment can be inaccurate in the presence of skewness and outliers. On the other hand, NB provides FWER and CP close to target even when $n = 50$. NB outperforms AMC in FWER and CP but has larger AL in all scenarios. This finding is consistent with our experience that $c_\alpha^{\text{AMC}} < c_\alpha^{\text{NB}}$ holds in most cases and in keeping with the slow convergence of Wald-type statistics to chi-square distributions (see, e.g. Pauly, Brunner, and Konietschke 2015). The performances of AMC and NB are similar when $n = 200$ or when the data range is restricted (Table 3). Interestingly, NB is more conservative when $n = 50$ than when $n = 100$ or $n = 200$.

Table 1. Simulation results under scenario S1.

			AMC			NB			HBW	
n	(θ_1, θ_2)	FWER	AL	CP (%)	FWER	AL	CP (%)	FWER	AL	CP (%)
					S1-1					
	$(0, 0)$	0.077	2.284	92.4	0.043	2.498	95.7	0.065	1.973	93.5
50	$(1, 0)$	0.057	2.283	91.4	0.034	2.496	94.7	0.044	1.974	93.3
	$(2, 1)$	0.030	2.282	91.8	0.018	2.491	95.0	0.026	1.970	93.4
	$(0, 0)$	0.061	1.628	93.9	0.051	1.674	94.9	0.058	1.401	94.2
100	$(1, 0)$	0.036	1.629	94.3	0.030	1.674	95.4	0.036	1.403	94.6
	$(2, 1)$	0.021	1.628	93.8	0.017	1.673	94.9	0.020	1.401	94.3
	$(0, 0)$	0.053	1.148	94.7	0.049	1.160	95.1	0.054	0.993	94.6
200	$(1, 0)$	0.037	1.149	94.3	0.033	1.161	94.9	0.035	0.994	94.5
	$(2, 1)$	0.017	1.149	95.3	0.017	1.158	95.4	0.018	0.994	94.9
					S1-2					
	$(0, 0)$	0.078	2.574	92.2	0.041	2.859	95.9	0.101	2.783	89.9
50	$(1, 0)$	0.058	2.569	91.7	0.028	2.854	96.0	0.097	2.775	89.1
	$(2, 1)$	0.031	2.572	91.7	0.018	2.829	94.8	0.078	2.778	89.5
	$(0, 0)$	0.061	1.845	94.0	0.052	1.905	94.8	0.101	1.979	89.9
100	$(1, 0)$	0.041	1.846	93.7	0.033	1.906	94.8	0.092	1.979	89.7
	$(2, 1)$	0.021	1.847	93.9	0.017	1.897	94.5	0.075	1.980	89.7
	$(0, 0)$	0.053	1.305	94.7	0.051	1.320	95.0	0.094	1.405	90.6
200	$(1, 0)$	0.036	1.304	94.4	0.035	1.320	94.8	0.084	1.404	90.7
	$(2, 1)$	0.020	1.305	94.4	0.019	1.307	94.5	0.071	1.405	90.4

θ_3, θ_4, and θ_5 are set to zero. The largest standard error of the results is 0.003 when $n = 50$.

Table 2. Simulation results under scenario S2.

			AMC			NB			HBW	
n	(θ_1, θ_2)	FWER	AL	CP (%)	FWER	AL	CP (%)	FWER	AL	CP (%)
					S2-1					
	$(0, 0)$	0.091	3.032	90.1	0.047	3.435	95.2	0.064	2.444	93.6
50	$(1, 0)$	0.064	3.035	90.2	0.031	3.439	95.1	0.044	2.445	93.3
	$(2, 1)$	0.032	3.029	90.9	0.015	3.431	95.1	0.023	2.439	93.6
	$(0, 0)$	0.072	2.167	92.8	0.053	2.270	94.7	0.057	1.739	94.3
100	$(1, 0)$	0.041	2.168	93.4	0.031	2.271	95.1	0.036	1.741	94.3
	$(2, 1)$	0.023	2.171	93.2	0.017	2.275	95.1	0.021	1.742	94.6
	$(0, 0)$	0.056	1.524	94.5	0.049	1.552	95.1	0.053	1.234	94.7
200	$(1, 0)$	0.039	1.526	94.0	0.035	1.553	94.7	0.035	1.234	94.7
	$(2, 1)$	0.020	1.526	94.2	0.018	1.554	94.7	0.019	1.236	94.4
					S2-2					
	$(0, 0)$	0.092	2.761	90.8	0.047	3.091	95.3	0.093	2.917	90.7
50	$(1, 0)$	0.055	2.765	91.2	0.027	3.097	95.6	0.082	2.922	90.4
	$(2, 1)$	0.029	2.765	90.7	0.014	3.079	94.8	0.069	2.923	90.3
	$(0, 0)$	0.066	1.992	93.4	0.051	2.074	94.9	0.089	2.077	91.1
100	$(1, 0)$	0.040	1.991	93.7	0.031	2.072	95.1	0.078	2.075	91.1
	$(2, 1)$	0.022	1.993	93.5	0.017	2.072	94.8	0.064	2.077	90.9
	$(0, 0)$	0.056	1.412	94.4	0.050	1.435	95.1	0.086	1.472	91.5
200	$(1, 0)$	0.036	1.414	94.6	0.032	1.438	95.2	0.072	1.474	91.9
	$(2, 1)$	0.019	1.415	94.4	0.017	1.432	94.8	0.059	1.474	91.6

θ_3, θ_4, and θ_5 are set to zero. The largest standard error of the results is 0.003 when $n = 50$.

This is partly due to the higher chance that a bootstrap sample may not satisfy the convex hull constraint, contributing to the large cutoff of NB (see Remark 4).

The HBW procedure, contrary to the empirical likelihood-based procedures, depends heavily on the assumptions in (3). HBW performs well in scenarios where the compound symmetry assumption is met (S1-1, S2-1, and S3-1). FWER and CP are robust across different distributions for the block effects and the errors. Except for $n = 200$, HBW outperforms AMC and comes

Table 3. Simulation results under scenario S3.

		AMC			NB			HBW		
n	(θ_1, θ_2)	FWER	AL	CP (%)	FWER	AL	CP (%)	FWER	AL	CP (%)
					S3-1					
	$(0, 0)$	0.078	0.652	92.3	0.046	0.699	95.4	0.070	0.571	93.1
50	$(1/8, 0)$	0.054	0.652	92.2	0.034	0.699	94.9	0.046	0.571	93.3
	$(1/4, 1/8)$	0.029	0.651	92.2	0.018	0.699	94.9	0.024	0.571	93.3
	$(0, 0)$	0.060	0.465	94.0	0.051	0.474	94.9	0.058	0.405	94.2
100	$(1/8, 0)$	0.039	0.465	93.8	0.035	0.474	94.4	0.041	0.405	93.7
	$(1/4, 1/8)$	0.021	0.465	94.2	0.019	0.475	95.0	0.021	0.406	94.4
	$(0, 0)$	0.053	0.329	94.8	0.050	0.332	95.0	0.055	0.287	94.5
200	$(1/8, 0)$	0.034	0.329	94.9	0.032	0.332	95.1	0.033	0.287	94.9
	$(1/4, 1/8)$	0.018	0.329	94.9	0.017	0.332	95.3	0.017	0.287	94.6
					S3-2					
	$(0, 0)$	0.108	1.062	90.9	0.033	1.536	96.6	0.117	1.277	88.3
50	$(1/8, 0)$	0.079	1.077	91.1	0.024	1.567	96.6	0.109	1.304	88.3
	$(1/4, 1/8)$	0.056	1.071	91.2	0.015	1.547	96.7	0.097	1.292	88.5
	$(0, 0)$	0.070	0.807	93.3	0.045	0.886	95.6	0.105	0.940	89.5
100	$(1/8, 0)$	0.049	0.803	93.2	0.032	0.877	95.3	0.096	0.934	89.7
	$(1/4, 1/8)$	0.032	0.809	93.2	0.022	0.886	95.1	0.087	0.943	89.7
	$(0, 0)$	0.061	0.580	93.9	0.051	0.603	94.9	0.106	0.673	89.4
200	$(1/8, 0)$	0.041	0.583	94.0	0.035	0.608	94.9	0.101	0.678	89.3
	$(1/4, 1/8)$	0.025	0.583	94.2	0.020	0.607	95.3	0.092	0.678	89.3

θ_3, θ_4, and θ_5 are set to zero. The largest standard error of the results is 0.007 when $n = 50$.

close to NB with considerably shorter AL. In scenarios where compound symmetry is violated (S1-2, S2-2, and S3-2), however, HBW shows a substantial performance deterioration. The AL of HBW is larger than those of AMC and NB when n is 100 or 200. FWER and CP are far from their target values, and the rate at which they improve is much slower. Figure 1 further shows the impact of the violation on AL and CP by gradually decreasing the degrees of freedom for the distribution of ϵ_{i5} in S3-2 when $n = 200$ and $\theta = (0, 0, 0, 0, 0)$. The AL of AMC and NB is much larger for the intervals with θ_5 than the rest, and only the AL of these intervals increases as the degrees of freedom decrease to 2 (infinite variance). As a result of this adjustment, the CP of the individual interval is maintained above 0.95 for AMC and NB. In contrast, all intervals of HBW have the same length. This implies that the intervals with θ_5 are not wide enough as SCIs, causing the under-coverage shown in Figure 1. For AMC and NB, additional simulation results for gFWER control are also shown in Table 4.

In summary, AMC and NB show robust performance without relying on restrictive model assumptions. Notably, NB successfully approximates the target error rate and CP even with small sample sizes. The performance gap between AMC and NB is modest for larger n, where the computational burden of the bootstrap and optimization involved in NB gives the advantage to AMC.

6. Application to pesticide concentration experiments

We apply the methodology developed in Sections 3 and 4 to analyze the clothianidin concentration data in Alford and Krupke (2017). Clothianidin, a neonicotinoid pesticide, is a potent agonist of the nicotinic acetylcholine receptor in insects and is extensively applied in the United States to maize seeds before planting. Quantifying the amount of clothianidin translocated into plant tissue, coupled with its potential for environmental accumulation via runoff or leaching, provides information on the costs and benefits of this delivery method.

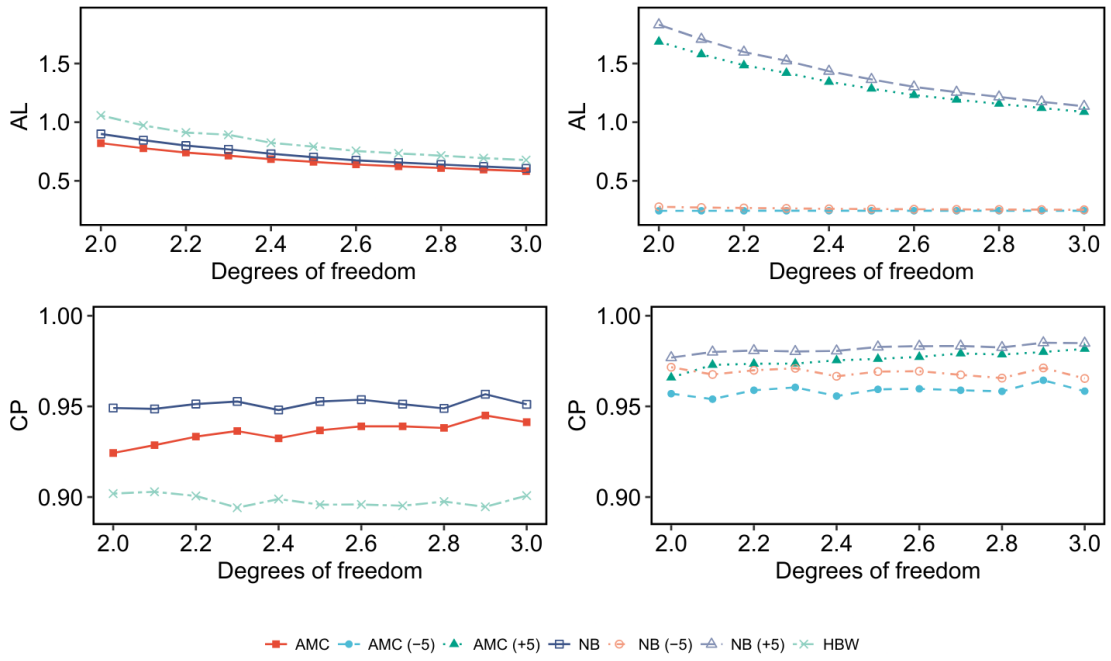


Figure 1. AL and CP of S3-2 with varying degrees of freedom for the distribution of ϵ_{i5} when $n = 200$ and $\theta = (0, 0, 0, 0, 0)$. For AMC and NB, separate results are presented for the comparisons that involve θ_5 $(+5)$ and those that do not (-5). AL and CP are computed group-wise.

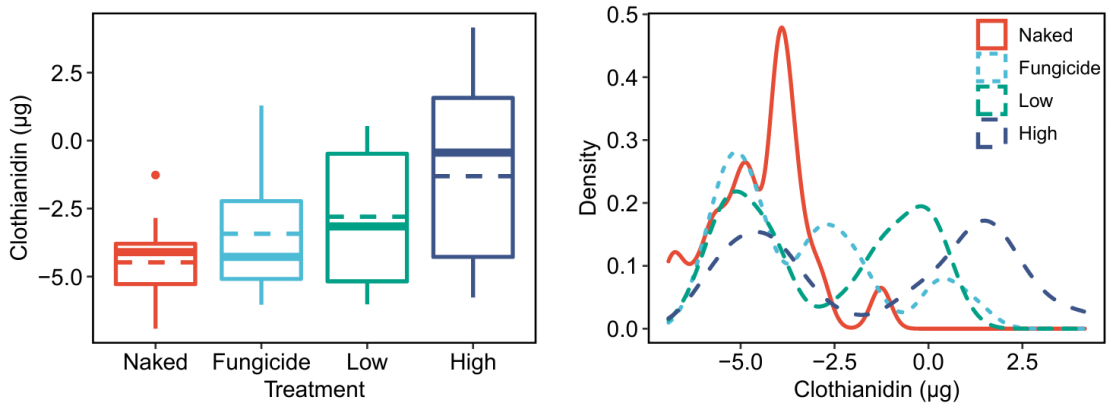
Alford and Krupke (2017) investigated clothianidin concentration in three regions (seed, root, and shoot) of maize plants in the early growing season. Two field experiments were conducted in 2014 and 2015 with four seed treatments: untreated (Naked), fungicide only (Fungicide), low rate of 0.25 mg clothianidin/kernel (Low), and high rate of 1.25 mg clothianidin/kernel (High). In a randomized complete block design with four blocks for each year, the treatments were applied to four plots in each block. Clothianidin contamination of the untreated plots was expected due to subsurface flow and proximity between plots. Sampling was carried out at 6, 8, 10, 13, 15, 17, 20, and 34 days post-planting in 2014, and at 5, 7, 9, 12, 14, 16, 19, 47, and 61 days post-planting in 2015. On each sampling day, up to ten plants were randomly sampled from each plot, and three or five of them were processed for chemical analysis. Some plant observations were lost before the analysis. Root and shoot regions of the remaining observations were scored as "complete" ($> 80\%$ present) or "incomplete" ($< 80\%$ present). In this way, the experimental design had a hierarchical structure of sampling (year/block/days post-planting/plot) that is unbalanced and incomplete in several layers. The clothianidin concentration data were log-transformed to conform more closely to a normal distribution. Alford and Krupke (2017) fit a linear mixed model to test two contrasts: Naked + Fungicide vs. Low and Naked + Fungicide vs. High. Jensen, Schaarschmidt, Onofri, and Ritz (2018) used the same data and performed a variety of post-hoc pairwise comparisons with various linear mixed models.

We subdivide the original blocks into 68 new blocks according to the days post-planting, considering that plant tissue clothianidin dissipates over time following an exponential decay pattern (Alford and Krupke 2017, Fig. 2). For illustration, we analyze only the "incomplete" shoot region observations. This results in 32 blocks. Plot level replicates, if any, are averaged over the treatments within these blocks, resulting in 102 observations in total. The decay pattern gives

Table 4. Simulation results for the gFWER control. For the three simulation scenarios, gFWER of each procedure is computed below for $v = 2, 3, 4, 5$ with $\theta = (0, 0, 0, 0, 0)$.

	AMC				NB			
n	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$
				S1-1				
50	0.065	0.057	0.052	0.054	0.041	0.039	0.038	0.040
100	0.055	0.054	0.054	0.054	0.048	0.049	0.048	0.046
200	0.056	0.054	0.049	0.053	0.053	0.050	0.046	0.050
				S1-2				
50	0.067	0.058	0.057	0.058	0.042	0.040	0.042	0.043
100	0.056	0.057	0.056	0.057	0.048	0.051	0.051	0.048
200	0.055	0.051	0.051	0.051	0.051	0.048	0.048	0.048
				S2-1				
50	0.074	0.064	0.058	0.058	0.044	0.043	0.040	0.041
100	0.065	0.063	0.058	0.056	0.054	0.050	0.050	0.046
200	0.055	0.053	0.050	0.051	0.051	0.049	0.047	0.046
				S2-2				
50	0.076	0.067	0.062	0.060	0.044	0.045	0.044	0.043
100	0.061	0.055	0.057	0.058	0.051	0.048	0.050	0.050
200	0.055	0.053	0.051	0.049	0.050	0.049	0.049	0.046
				S3-1				
50	0.067	0.064	0.061	0.056	0.049	0.049	0.047	0.044
100	0.053	0.051	0.051	0.055	0.046	0.047	0.046	0.048
200	0.052	0.047	0.051	0.054	0.049	0.046	0.048	0.050
				S3-2				
50	0.097	0.091	0.087	0.071	0.039	0.044	0.047	0.046
100	0.070	0.072	0.073	0.063	0.048	0.050	0.051	0.048
200	0.062	0.060	0.061	0.058	0.049	0.046	0.047	0.049

The largest standard error of the results is 0.003 when $n = 50$.



(a) (b)

Figure 2. Summary of data for each treatment: (a) box plot of log transformed clothianidin concentration with median (solid line) and mean (dashed line); (b) density plot of clothianidin concentration.

rise to skewed or multimodal marginal distributions for the treatments with different variances, as shown in Figure 2. Regardless of the treatments, the clothianidin concentration is close to zero roughly 20 days post-planting. Furthermore, the pairwise plots in Figure 3 indicate that no pairs of treatments follow a bivariate normal distribution. Salient features of the data are heteroscedasticity, non-normality, and a violation of block-treatment additivity.

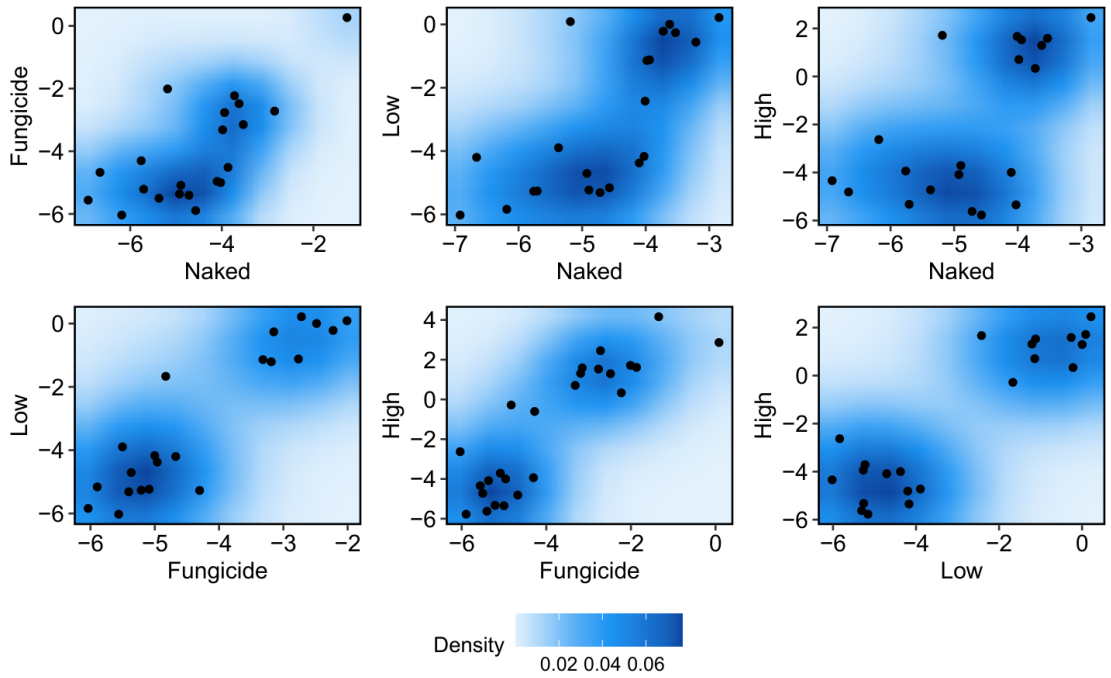


Figure 3. Pairwise scatter plots of observations. Each dot represents a pair of observations in a block; incomplete pairs are discarded in each plot. The overlaid heat maps show densities. Many dots are located either in the bottom left or upper right corners.

Table 5. Pairwise comparisons between treatments: Naked(N), Fungicide(F), Low(L), and High(H). The estimates are obtained from empirical likelihood (EL) and the mixed effect model.

	Estimate		p-value		
Comparison	EL	Mixed model	AMC	NB	HBW
$\theta_F - \theta_N$	1.053	0.443	0.001	0.008	0.601
$\theta_L - \theta_N$	1.679	1.615	< 0.001	0.005	< 0.001
$\theta_H - \theta_N$	3.173	2.883	< 0.001	0.001	< 0.001
$\theta_L - \theta_F$	0.627	1.172	0.503	0.532	0.006
$\theta_H - \theta_F$	2.120	2.434	0.001	0.007	< 0.001
$\theta_H - \theta_L$	1.493	1.268	0.003	0.015	0.002

The p-values are based on 10,000 Monte Carlo samples for AMC and 10,000 bootstrap replicates for NB.

We follow the procedures outlined in Section 5 to perform pairwise comparisons. We obtain estimates from the empirical likelihood and linear mixed model in (3); then, adjusted p-values and confidence intervals are constructed using AMC, NB, and HBW. Table 5 reports the estimates and p-values, where the treatment effects are denoted by θ_N, θ_F, θ_L, and θ_H. Despite the small number of blocks ($n = 32$), AMC and NB reach similar conclusions: the clothianidin concentration does not differ significantly between the fungicide and low rate treated plants ($\theta_L - \theta_F$), but significant differences are observed in all the other comparisons. In Figure 4, the lengths of the SCIs are similar for the two procedures, with AMC intervals being slightly shorter. The intervals are wider for those comparisons involving some clothianidin treatment.

HBW produces a different conclusion. The violations of the assumptions for HBW lead to different estimates for treatment effects than under empirical likelihood (except for θ_F). This leads to substantially different estimates for comparisons $\theta_F - \theta_N$ and $\theta_L - \theta_F$. The equal variance assumption distorts the standard errors for the comparisons. This contributes to the large p-value for $\theta_F - \theta_N$ produced by HBW.

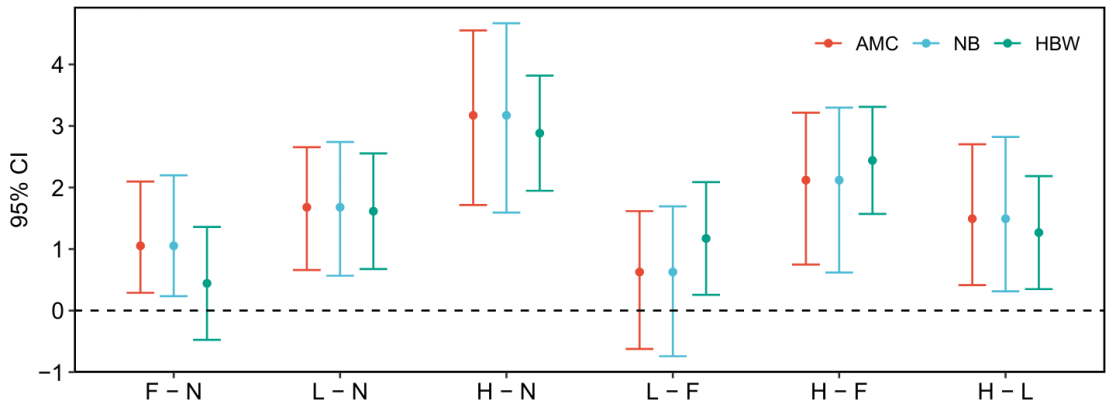


Figure 4. Asymptotic 95% simultaneous confidence intervals for pairwise comparisons. The estimates are given as dots inside the error bars. As a result of the larger cutoff, the NB interval contains the corresponding AMC interval for all comparisons.

To summarize, empirical likelihood has the advantage of avoiding clearly inappropriate assumptions. For these data, imposing these assumptions and performing a standard analysis leads to different conclusions.

7. Discussion

Several extensions remain open for future research. We are primarily interested in producing common cutoffs for pairwise comparisons and SCIs, but the AMC and NB procedures can be modified to yield common quantiles. Common quantile procedures are appropriate when the asymptotic multivariate chi-square distribution in Theorem 3.1 has different degrees of freedom for each marginal. While preserving the gFWER control, improvement in power can be achieved by adapting to stepwise procedures in both approaches.

As previously mentioned in Section 3.1, it is also possible to study other parameters and estimating functions in a similar fashion. AMC and NB would need minor adjustment once an asymptotic multivariate distribution is established, though it can be challenging to specify the null transformation for NB. Due to nonconvexity, a major challenge would be the computation of the statistics in (2). See Tang and Wu (2014) for general strategies for computing constrained empirical likelihood problems.

Finally, another interesting topic concerning multiple testing is the use of high-dimensional estimating functions with a growing number of parameters. Hjort et al. (2009) and Tang and Leng (2010) investigated the feasibility of empirical likelihood methods when p, the dimension of the parameter, is allowed to increase with n. In such high-dimensional settings, however, the typical objective is often to control other types of error rates, such as the false discovery rate, which is outside the scope of this article.

Acknowledgments

We thank the associate editor and three reviewers for their valuable feedback and suggestions, which greatly improved the quality of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the US National Science Foundation under Grants No. SES-1921523 and DMS-2015552.

References

Adimari, G., and Guolo, A. (2010), 'A Note on the Asymptotic Behaviour of Empirical Likelihood Statistics', *Statistical Methods & Applications*, 19, 463–476.

Alford, A., and Krupke, C.H. (2017), 'Translocation of the Neonicotinoid Seed Treatment Clothianidin in Maize', *PLOS ONE*, 12, e0173836.

Alvo, M. (2015), 'Empirical Likelihood and Ranking Methods', in *Asymptotic Laws and Methods in Stochastics: A Volume in Honour of Miklós Csörgő*, Springer-Verlag, pp. 367–377.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015), 'Fitting Linear Mixed-Effects Models Using lme4', *Journal of Statistical Software*, 67, 1–48.

Bücher, A., and Kojadinovic, I. (2019), 'A Note on Conditional Versus Joint Unconditional Weak Convergence in Bootstrap Consistency Results', *Journal of Theoretical Probability*, 32, 1145–1165.

Chaudhuri, S., Mondal, D., and Yin, T. (2017), 'Hamiltonian Monte Carlo Sampling in Bayesian Empirical Likelihood Computation', *Journal of the Royal Statistical Society, Series B*, 79, 293–320.

Chen, J., Variyath, A.M., and Abraham, B. (2008), 'Adjusted Empirical Likelihood and its Properties', *Journal of Computational and Graphical Statistics*, 17, 426–443.

DiCiccio, T., Hall, P., and Romano, J. (1991), 'Empirical Likelihood is Bartlett-Correctable', *The Annals of Statistics*, 19, 1053–1061.

Dickhaus, T. (2014), *Simultaneous Statistical Inference: With Applications in the Life Sciences*, Springer-Verlag.

Dickhaus, T., and Royen, T. (2015), 'A Survey on Multivariate Chi-Square Distributions and Their Applications in Testing Multiple Hypotheses', *Statistics*, 49, 427–454.

Dudley, R.M. (2002), *Real Analysis and Probability*, Cambridge University Press.

Eisinga, R., Heskes, T., Pelzer, B., and Te Grotenhuis, M. (2017), 'Exact p-Values for Pairwise Comparison of Friedman Rank Sums, with Application to Comparing Classifiers', *BMC Bioinformatics*, 18, 1–18.

Fey, M., and Clarke, K.A. (2012), 'Consistency of Choice in Nonparametric Multiple Comparisons', *Journal of Nonparametric Statistics*, 24, 531–541.

Friedman, M. (1937), 'The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance', *Journal of the American Statistical Association*, 32, 675–701.

Hjort, N.L., McKeague, I.W., and van Keilegom, I. (2009), 'Extending the Scope of Empirical Likelihood', *The Annals of Statistics*, 37, 1079–1111.

Hothorn, T., Bretz, F., and Westfall, P. (2008), 'Simultaneous Inference in General Parametric Models', *Biometrical Journal*, 50, 346–363.

Jacod, J., and Sørensen, M. (2018), 'A Review of Asymptotic Theory of Estimating Functions', *Statistical Inference for Stochastic Processes*, 21, 415–434.

Jensen, S.M., Schaarschmidt, F., Onofri, A., and Ritz, C. (2018), 'Experimental Design Matters for Statistical Analysis: How to Handle Blocking', *Pest Management Science*, 74, 523–534.

Kim, E. (2022), *melt: Multiple Empirical Likelihood Tests*, https://CRAN.R-project.org/package=melt. R package version 1.9.0.

Lehmann, E.L., and D'Abrera, H.J. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day.

Lehmann, E.L., and Romano, J.P. (2005), 'Generalizations of the Familywise Error Rate', *The Annals of Statistics*, 33, 1138–1154.
Liu, T., Yuan, X., Lin, N., and Zhang, B. (2012), 'Rank-Based Empirical Likelihood Inference on Medians of k Populations', *Journal of Statistical Planning and Inference*, 142, 1009–1026.
Mansouri, H., and Shaw, C. (2004), 'Nonparametric Multiple Comparison Procedures for Ordered Parameters in Balanced Incomplete Blocks', *Computational Statistics & Data Analysis*, 46, 593–604.
Nemenyi, P.B. (1963), *Distribution-Free Multiple Comparisons*, Princeton University.
Owen, A.B. (1988), 'Empirical Likelihood Ratio Confidence Intervals for a Single Functional', *Biometrika*, 75, 237–249.
Owen, A.B. (1990), 'Empirical Likelihood Ratio Confidence Regions', *The Annals of Statistics*, 18, 90–120.
Owen, A.B. (1991), 'Empirical Likelihood for Linear Models', *The Annals of Statistics*, 19, 1725–1747.
Owen, A.B. (2001), *Empirical Likelihood*, Chapman and Hall/CRC.
Pauly, M., Brunner, E., and Konietschke, F. (2015), 'Asymptotic Permutation Tests in General Factorial Designs', *Journal of the Royal Statistical Society, Series B*, 77, 461–473.
Qin, J., and Lawless, J. (1994), 'Empirical Likelihood and General Estimating Equations', *The Annals of Statistics*, 22, 300–325.
Qin, J., and Lawless, J. (1995), 'Estimating Equations, Empirical Likelihood and Constraints on Parameters', *Canadian Journal of Statistics*, 23, 145–159.
Singh, K. (1981), 'On the Asymptotic Accuracy of Efron's Bootstrap', *The Annals of Statistics*, 9, 1187–1195.
Stange, J., Loginova, N., and Dickhaus, T. (2016), 'Computing and Approximating Multivariate Chi-Square Probabilities', *Journal of Statistical Computation and Simulation*, 86, 1233–1247.
Tang, C.Y., and Leng, C. (2010), 'Penalized High-Dimensional Empirical Likelihood', *Biometrika*, 97, 905–920.
Tang, C.Y., and Wu, T.T. (2014), 'Nested Coordinate Descent Algorithms for Empirical Likelihood', *Journal of Statistical Computation and Simulation*, 84, 1917–1930.
Team, R.C. (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.
Tsao, M. (2004), 'Bounds on Coverage Probabilities of the Empirical Likelihood Ratio Confidence Regions', *The Annals of Statistics*, 32, 1215–1221.
Tsao, M., and Wu, F. (2013), 'Empirical Likelihood on the Full Parameter Space', *The Annals of Statistics*, 41, 2176–2196.
Wang, L., and Yang, D. (2018), 'F-Distribution Calibrated Empirical Likelihood Ratio Tests for Multiple Hypothesis Testing', *Journal of Nonparametric Statistics*, 30, 662–679.
Westfall, P.H., and Young, S.S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Wiley.
Yuan, K.H., and Jennrich, R.I. (1998), 'Asymptotics of Estimating Equations under Natural Conditions', *Journal of Multivariate Analysis*, 65, 245–260.

Appendix

Proof of Theorem 3.1. By Condition 3, $S_n(\theta_0) \to V$ in probability and $S_n(\theta_0)$ has full rank with high probability for large n. Then, Proposition 1 of Chaudhuri, Mondal, and Yin (2017) applies and there exists an open ball around θ_0 where $l_n(\theta)$ is defined. Adjusting $\mathcal{N}$ if necessary, it follows from Condition 1 that

$$P\{0 \in \text{Conv}_n(\theta) \text{ for all } \theta \in \mathcal{N}\} \to 1. \tag{5}$$

Moreover, the implicit function theorem implies that $l_n(\theta)$ is continuously differentiable on $\mathcal{N}$.

Consider any consistent estimator $\widehat{\theta}$ of θ_0 such that $\widehat{\theta} - \theta_0 = O_P(a_n^{-1})$. From (5), we assume that the convex hull constraint is satisfied for $\widehat{\theta}$ throughout the proof. Let $\widehat{g}_i = g(X_i, \widehat{\theta})$ and

$Z_n = \max_{1\le i\le n} |\widehat{g}_i|$. Following standard arguments as in Owen (2001, pp. 219–222), write $\widehat{\lambda} \equiv \widehat{\lambda}(\widehat{\theta}) = |\widehat{\lambda}|\widehat{\mu}$ for $|\widehat{\mu}| = 1$, where $\widehat{\lambda}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{g}_i}{1+\widehat{\lambda}^\top \widehat{g}_i} = 0. \tag{6}$$

Substituting $1/(1+\widehat{\lambda}^\top\widehat{g}_i) = 1 - \widehat{\lambda}^\top\widehat{g}_i/(1+\widehat{\lambda}^\top\widehat{g}_i)$ into (6), we obtain

$$0 = \widehat{\mu}^\top G_n(\widehat{\theta}) - |\widehat{\lambda}|\widehat{\mu}^\top \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{g}_i\widehat{g}_i^\top}{1+\widehat{\lambda}^\top\widehat{g}_i}\right)\widehat{\mu}.$$

It follows that $|\widehat{\lambda}|\widehat{\mu}^\top S_n(\widehat{\theta})\widehat{\mu} \le \widehat{\mu}^\top G_n(\widehat{\theta})(1+|\widehat{\lambda}|Z_n)$, and we have

$$|\widehat{\lambda}|\left(\widehat{\mu}^\top S_n(\widehat{\theta})\widehat{\mu} - Z_n\widehat{\mu}^\top G_n(\widehat{\theta})\right) \le \widehat{\mu}^\top G_n(\widehat{\theta}).$$

From Condition 2, a Taylor expansion of $G_n(\widehat{\theta})$ around θ_0 yields

$$G_n(\widehat{\theta}) = G_n(\theta_0) + \partial_\theta G_n(\theta_0)(\widehat{\theta}-\theta_0) + o_P(a_n^{-1}),$$

and we see that $\widehat{\mu}^\top G_n(\widehat{\theta}) = O_P(a_n^{-1})$ from Condition 4. Similarly, from Condition 5 we have

$$Z_n \le \max_{1\le i\le n} |g(X_i,\theta_0)| + \left(\max_{1\le i\le n} \|\partial_\theta g(X_i,\theta_0)\|\right)|\widehat{\theta}-\theta_0| + o_P(|\widehat{\theta}-\theta_0|) = o_P(a_n).$$

Since $\widehat{\theta} \to \theta_0$ in probability, there exists a sequence $\epsilon_n \to 0$ such that $P(|\widehat{\theta}-\theta_0| > \epsilon_n) \to 0$, and $\sup_{|\theta-\theta_0|\le\epsilon_n} \|S_n(\theta) - S_n(\theta_0)\| \to 0$ in probability from Condition 3. Then for any $\epsilon > 0$,

$$P\left(\|S_n(\widehat{\theta}) - S_n(\theta_0)\| > \epsilon\right) \le P\left(\sup_{|\theta-\theta_0|\le\epsilon_n} \|S_n(\theta) - S_n(\theta_0)\| > \epsilon\right) + P\left(|\widehat{\theta}-\theta_0| > \epsilon_n\right),$$

and it follows that $S_n(\widehat{\theta}) - S_n(\theta_0) \to 0$ and $S_n(\widehat{\theta}) \to V$ in probability. We write $\sigma_{\min} + o_P(1) \le \widehat{\mu}^\top S_n(\widehat{\theta})\widehat{\mu} \le \sigma_{\max} + o_P(1)$, with $0 < \sigma_{\min} \le \sigma_{\max}$ denoting the smallest and largest eigenvalues of v. This shows that $\widehat{\lambda} = O_P(a_n^{-1})$. Iterating the substitution after (6) gives $1/(1+\widehat{\lambda}^\top\widehat{g}_i) = 1 - \widehat{\lambda}^\top\widehat{g}_i + (\widehat{\lambda}^\top\widehat{g}_i)^2/(1+\widehat{\lambda}^\top\widehat{g}_i)$, leading to $0 = G_n(\widehat{\theta}) - S_n(\widehat{\theta})\widehat{\lambda} + r_n(\widehat{\theta})$, where

$$|r_n(\widehat{\theta})| = \left|\frac{1}{n}\sum_{i=1}^{n} \frac{(\widehat{\lambda}^\top\widehat{g}_i)^2\widehat{g}_i}{1+\widehat{\lambda}^\top\widehat{g}_i}\right| \le \max_{1\le i\le n} |1+\widehat{\lambda}^\top\widehat{g}_i|^{-1} Z_n|\widehat{\lambda}|^2\|S_n(\widehat{\theta})\|.$$

With $\max_{1\le i\le n} |1+\widehat{\lambda}^\top\widehat{g}_i|^{-1} = O_P(1)$, it follows that $r_n(\widehat{\theta}) = o_P(a_n^{-1})$ and

$$\widehat{\lambda} = S_n(\widehat{\theta})^{-1} G_n(\widehat{\theta}) + o_P(a_n^{-1}), \tag{7}$$

where $S_n(\widehat{\theta})$ is invertible with probability tending to 1.

Define the empirical likelihood statistic $l_n(\widehat{\theta}) = 2a_n^2 n^{-1} \sum_{i=1}^n \log(1 + \widehat{\lambda}^\top \widehat{g}_i)$ and apply a Taylor expansion of $\log(1+x)$ to write

$$\begin{aligned} l_n(\widehat{\theta}) &= \frac{2a_n^2}{n} \left(\widehat{\lambda}^\top \sum_{i=1}^n \widehat{g}_i - \frac{1}{2} \widehat{\lambda}^\top \sum_{i=1}^n \widehat{g}_i \widehat{g}_i^\top \widehat{\lambda} + \frac{1}{3} \sum_{i=1}^n \frac{(\widehat{\lambda}^\top \widehat{g}_i)^3}{(1+\eta_i)^3} \right) \\ &= 2a_n^2 \widehat{\lambda}^\top G_n(\widehat{\theta}) - a_n^2 \widehat{\lambda}^\top S_n(\widehat{\theta}) \widehat{\lambda} + \frac{2}{3} R_n(\widehat{\theta}), \end{aligned}$$

where $|\eta_i| < |\widehat{\lambda}^\top \widehat{g}_i|$ for all i and

$$|R_n(\widehat{\theta})| = \left| \frac{a_n^2}{n} \sum_{i=1}^n \frac{(\widehat{\lambda}^\top \widehat{g}_i)^3}{(1+\eta_i)^3} \right| \leq a_n^2 |\widehat{\lambda}| Z_n (1 + |\widehat{\lambda}| Z_n)^{-3} |\widehat{\lambda}^\top S_n(\widehat{\theta}) \widehat{\lambda}| = o_P(1).$$

From (7), we have $l_n(\widehat{\theta}) = a_n^2 G_n(\widehat{\theta})^\top V^{-1} G_n(\widehat{\theta}) + o_P(1)$, and it can also be shown that

$$a_n^2 G_n(\widehat{\theta})^\top V^{-1} G_n(\widehat{\theta}) = a_n^2 (H_n + \widehat{\theta} - \theta_0)^\top M^{-1} (H_n + \widehat{\theta} - \theta_0) + o_P(1),$$

where $H_n = W G_n(\theta_0)$. Thus, we can write

$$l_n(\widehat{\theta}) = Q_n(\widehat{\theta}) + o_P(1), \tag{8}$$

where $Q_n(\widehat{\theta}) = a_n^2 (H_n + \widehat{\theta} - \theta_0)^\top M^{-1} (H_n + \widehat{\theta} - \theta_0)$ is a quadratic approximation to $l_n(\widehat{\theta})$.

We now introduce a generic q-dimensional function $h(\theta)$ with Jacobian matrix J from Condition 6. Since (8) holds for any $\widehat{\theta}$ such that $\widehat{\theta} - \theta_0 = O_P(a_n^{-1})$, it follows that

$$\inf_{\theta \in H \cap \overline{K}_n} l_n(\theta) = \min_{\theta \in H \cap \overline{K}_n} Q_n(\theta) + o_P(1), \tag{9}$$

where $H = \{\theta \in \Theta : h(\theta) = 0\}$ is the constraint set and $\overline{K}_n = \{\theta : |\theta - \theta_0| \leq K/a_n\}$ denotes any sequence of closed balls around θ_0 for $K > 0$; see Adimari and Guolo (2010, p. 470). Thus, we can consider minimizing $Q_n(\theta)$, instead of $l_n(\theta)$, under the constraint by constructing

$$L = Q_n(\theta) + 2a_n^2 h(\theta)^\top \nu,$$

where ν is a q-dimensional Lagrange multiplier. Differentiating L with respect to θ and ν, we obtain

$$\theta - \theta_0 + H_n + M J^\top \nu = 0 \text{ and } h(\theta) = h(\theta_0) + J(\theta - \theta_0) + o(|\theta - \theta_0|) = 0. \tag{10}$$

Since we only consider solutions $\tilde{\theta}$ such that $\tilde{\theta} - \theta_0 = O_P(a_n^{-1})$ from H_n and Condition 4, it follows from (10) that $\nu = P^{-1} J H_n + o_P(a_n^{-1})$, where $P = J M J^\top$. Consequently, we have $H_n + \tilde{\theta} - \theta_0 = M J^\top P^{-1} J H_n + o_P(a_n^{-1})$, and

$$Q_n(\tilde{\theta}) = a_n^2 G_n(\theta_0)^\top A G_n(\theta_0) + o_P(1), \tag{11}$$

where $A = (JW)^\top P^{-1} (JW)$. It follows from (9) and (11) that

$$\inf_{\theta \in H \cap \overline{K}_n} l_n(\theta) = a_n^2 G_n(\theta_0)^\top A G_n(\theta_0) + o_P(1)$$

for some $K > 0$ and $\overline{K}_n$.

Now consider hypotheses H_j for $j = 1, \ldots, m$. For each j, there exist $K_j > 0$ and $\overline{K}_n^j = \{\theta : |\theta - \theta_0| \leq K_j/a_n\}$ such that $\inf_{\theta \in H_j \cap \overline{K}_n^j} l_n(\theta) = a_n^2 G_n(\theta_0)^\top A_j G_n(\theta_0) + o_P(1)$. We may take $K = \max\{K_1, \ldots, K_m\}$ and define $T_{nj} = \inf_{\theta \in H_j \cap \overline{K}_n} l_n(\theta)$. Then,

$$T_{nj} = U_{nj}^\top \Sigma_j^{-1} U_{nj} + o_P(1),$$

where $U_{nj} = a_n J_j W G_n(\theta_0) \to U_j \sim N(0, J_j M J_j^\top)$ in distribution with $\Sigma_j = J_j M J_j^\top$. Applying the Crámer–Wold device, under H_0 we have

$$T_n = (T_{n1}, \ldots, T_{nm}) \to T \equiv (U_1^\top \Sigma_1^{-1} U_1, \ldots, U_m^\top \Sigma_m^{-1} U_m)$$

in distribution. Finally, $(\Sigma_1^{-1/2} U_1, \ldots, \Sigma_m^{-1/2} U_m)$ has a multivariate normal distribution with mean 0 and correlation matrix R, where R is a $\sum_{j=1}^m q_j \times \sum_{j=1}^m q_j$ block matrix whose kth diagonal matrix is I_{q_k} and (k, l) off-diagonal matrix is $\Sigma_k^{-1/2} J_k M J_l^\top \Sigma_l^{-1/2}$. Then, T follows a multivariate chi-square distribution with parameters m, $q = (q_1, \ldots, q_m)$, and R, i.e. $T \sim \chi^2(m, q, R)$. ■

Proof of Theorem 4.2. We present the proof of Theorem 4.2 first, followed by the proof of Theorem 4.1 since it readily follows from Theorem 4.2. We set up some notation and terminology. Recall that the bounded Lipschitz metric d_{BL} between two probability measures p and Q on $\mathbb{R}^m$ for $m \in \mathbb{N}$ metrizes weak convergence and is defined by

$$d_{\text{BL}}(P, Q) = \sup_{f \in \text{BL}_1} \left| \int f dP - \int f dQ \right|,$$

where BL_1 denotes the set of functions $f : \mathbb{R}^m \mapsto [-1, 1]$ such that $|f(x) - f(y)| \leq |x - y|$ for all $x, y \in \mathbb{R}^m$. As a mapping from the common probability space $(\Omega, \mathcal{F}, P)$ into the set of probability measures on $\mathbb{R}^m$, let $\widehat{P}_n$ be a sequence of random probability measures such that $\int f d\widehat{P}_n$ is measurable for any bounded and Lipschitz continuous function f. Then we say that $\widehat{P}_n$ converges weakly to p in probability if

$$\int f d\widehat{P}_n \to \int f dP \tag{12}$$

in probability for all $f \in \text{BL}_1$. Also, (12) holds if and only if $d_{\text{BL}}(\widehat{P}_n, P) \to 0$ in probability. Moreover, if the distribution function of p is continuous, it is equivalent to $d_K(\widehat{P}_n, P) \to 0$ in probability (see, e.g. Bücher and Kojadinovic 2019, Lemma 2.5), where

$$d_K(P, Q) = \sup_{x \in \mathbb{R}^m} |P(\{(-\infty, x)\}) - Q(\{(-\infty, x)\})|$$

denotes the Kolmogorov distance between P and Q.

Let X_i^* be an independent observation from $\mathcal{X}_n$ for $i = 1, \ldots, n$. It can be shown that, for each j, $T_{nj}^* = n G_n^*(\overline{X})^\top A_j^* G_n^*(\overline{X}) + o_P(1)$, where $G_n^*(\overline{X}) = n^{-1} \sum_{i=1}^n g(X_i^*, \overline{X})$ and

$$A_j^* = (J_j W^{-1})^\top \left(J_j (W^\top S_n^{*-1} W)^{-1} J_j^\top \right)^{-1} (J_j W^{-1}).$$

We first establish a bootstrap central limit theorem for $\sqrt{n}G_n^*(\overline{X})$ as in Singh (1981). Observe that

$$E^*\left\{g(X_i^*, \overline{X})\right\} = \frac{1}{n}\sum_{i=1}^n g(X_i, \overline{X}) = 0,$$

and

$$\frac{1}{n}\sum_{i=1}^n \text{Var}^*\left\{g(X_i^*, \overline{X})\right\} = \frac{1}{n}\sum_{i=1}^n g(X_i, \overline{X})g(X_i, \overline{X})^\top = S_n(\overline{X}) = S_n(\theta_0) + o(1)$$

almost surely. Then $S_n(\overline{X}) \to V$ almost surely by the law of large numbers. Applying the Lindeberg–Feller central limit theorem, for any $\epsilon > 0$ we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n E^*\left\{|g(X_i^*, \overline{X})|^2 1\left(|g(X_i^*, \overline{X})| \geq \epsilon\sqrt{n}\right)\right\} \\
&\quad \leq \frac{1}{n}\sum_{i=1}^n \left\{|X_i - \overline{X}|^2 1\left(|X_i - \overline{X}| \geq \epsilon\sqrt{n}\right)\right\} \to 0
\end{aligned}$$

almost surely. It follows that $\sqrt{n}G_n^*(\overline{X})$ converges weakly to a $N(0, V)$ distribution almost surely, i.e.

$$d_K\left(\mathcal{L}\left(\sqrt{n}G_n^*(\overline{X}) \,\middle|\, \mathcal{X}_n\right),\ N(0, V)\right) \to 0 \tag{13}$$

almost surely. Next, and let s_{jk}^* denote the (j, k) component of S_n^*. Then $E^*(S_n^*) = S_n(\overline{X})$ and, by the law of iterated expectation,

$$E\left(\|S_n^* - S_n(\overline{X})\|^2\right) \leq \sum_{j=1}^p \sum_{k=1}^p \frac{1}{n^2}\sum_{i=1}^n E\left\{(X_{ij} - \overline{X}_j)^2(X_{ik} - \overline{X}_k)^2\right\} \to 0.$$

This implies that $S_n^* \to V$ in probability.

It follows from the continuous mapping theorem and (13) that $d_K(\mathcal{L}^*(T_{nj}^*), \mathcal{L}(T_j)) \to 0$ in probability for each j. Then for every fixed $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$, an application of the continuous mapping theorem implies that

$$d_K\left(\mathcal{L}^*\left(\lambda^\top T_n^* \,\middle|\, \mathcal{X}_n\right),\ \mathcal{L}(\lambda^\top T)\right) \to 0 \tag{14}$$

in probability. From the subsequential property of convergence in probability (Dudley 2002, Theorem 9.2.1), there exists a subsequence such that (14) holds almost surely along the subsequence. Then the Crámer–Wold device implies that T_n^* converges weakly to T almost surely along the subsequence. Another application of the subsequential argument shows that

$$d_K\left(\mathcal{L}\left(T_n^* \mid \mathcal{X}_n\right),\ \mathcal{L}(T)\right) \to 0$$

in probability. Finally, $T_n \to T$ in distribution under H_0 and the result follows from the continuity of T. ■

Proof of Theorem 4.1. Since $S_n(\widehat{\theta}) \to V$ in probability, we have $\widehat{A}_j \to A_j$ in probability for $j = 1, \ldots, m$ and the continuity of the characteristic function of the normal distribution implies that

$$d_K\left(\mathcal{L}\left(U_n \mid \mathcal{X}_n\right), N(0, V)\right) \to 0$$

in probability. For any $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$, it follows from the continuous mapping theorem that

$$d_K\left(\mathcal{L}\left(U_n^\top \left(\sum_{j=1}^m \lambda_j \widehat{A}_j\right) U_n \,\middle|\, \mathcal{X}_n\right),\ \mathcal{L}\left(U^\top \left(\sum_{j=1}^m \lambda_j A_j\right) U\right)\right) \to 0$$

in probability. Then the Crámer–Wold device and the subsequential argument applied in (14) complete the proof. ■