Lag Order Selection for Granger Causality Estimation

Joshua H. Sylvester[1], Micah A. Thornton[2], Eric C. Larson[1], and Mitchell A. Thornton[1]

[1] Southern Methodist University, Darwin Deason Institute for Cybersecurity, Dallas, TX, USA
jsylvester@smu.edu, eclarson@smu.edu, mitch@smu.edu
[2] Texas Woman's University, Department of Mathematics, Denton, TX, USA
mthornton11@twu.edu

Abstract. Granger causality testing remains widely used for assessing relationships amongst time series, despite being limited to traditional bivariate time series analysis settings. The method entails selecting a lag order for vector autoregressive models, followed by statistical testing. The purpose of this work is to develop lag order selection methods that are explicitly aligned with the objectives of Granger causality testing, rather than general model fit. We propose two approaches for selecting lag order specifically tailored for Granger causality testing. Traditional methods, such as information criteria (AIC, BIC, HQIC) and sequential likelihood ratio tests, prioritize model fit but do not directly optimize for Granger causal significance. In contrast, our Granger-based methods determine lag order by measuring the impact of an independent variable's lagged values on a dependent variable, aligning with the core tenets of Granger causality. The first approach, Granger minimum p-value, is similar in structure to an information criterion but differs by optimizing Granger causal significance rather than model fit. The second, Granger-based likelihood ratio, refines likelihood ratio testing by isolating the effect of lagged predictors. Using datasets with known lag structures, we demonstrate these methods can outperform conventional lag order selection approaches and excel under different noise conditions. Our findings highlight both methods' potential for improving traditional Granger causality analysis in time series.

Keywords: Granger causality, bivariate, time series, lag order

1 Introduction

Modeling influence among time series is a critical aspect in areas such as forecasting [5], anomaly detection [16], self-supervised classification [23], clustering [11], causal structural learning [24], and many other applications. Granger causality, a test introduced by Clive Granger [7], leverages vector autoregressive (VAR) models to determine whether past values of an independent time series variable

x influence, or have predictive power for, future values of a dependent time series variable y. A reliable Granger causality result depends on the selection of a lag order that accurately captures the time-lagged relationship between x and y. Therefore, selecting the correct lag order is a critical component of applying Granger causality to time series data and can impact downstream tasks such as forecasting, anomaly detection, and causal modeling, where underestimating temporal structure can lead to model misspecification or misleading conclusions.

Several approaches exist for lag order selection in VAR modeling. One common method involves calculating a tradeoff between goodness of fit and model simplicity, referred to as information criterion tests (ICTs). Popular ICTs include the Akaike Information Criterion (AIC) [1, 2], Bayesian Information Criterion (BIC) [17], and Hannan-Quinn Information Criterion (HQIC) [8]. These ICTs fit time series models across all candidate lags and select the lag that minimizes the criterion and goodness of fit. An alternative approach involves performing sequential likelihood ratio (LR) tests, which iteratively test the null hypothesis that simpler models offer comparable explanatory power. Sequential LR tests proceed by starting with the largest lag and descending through the candidate lags, selecting the most recent lag for which the null hypothesis is not rejected. While sequential LR tests provide better Type I error control than ICTs due to the nested nature of the tested models, both methods focus primarily on model fit and do not directly leverage the principles underlying Granger causality. Therefore, we hypothesize that lag selection, when used as a preprocessing step for Granger causality estimation, can be improved.

For a time series x influencing a time series y, Granger causality models their relationship at varying time lags. However, existing lag order selection methods do not explicitly account for causal mechanisms. To address this, we propose two Granger-based approaches that integrate causality testing into lag selection. The first, Granger Minimum p-value (GMP), tests Granger causality across candidate lags and selects the lag with the lowest p-value. While not an information criterion specifically, GMP accounts for model complexity via degrees of freedom, paralleling the penalization of complexity in AIC, BIC, or HQIC. The second approach, termed Granger-based likelihood ratio (GLR), adapts sequential LR testing by isolating the influence x exerts on y: it varies only the number of lags of x, while keeping the number of lags of y fixed. This ensures that the lag order selection process isolates the temporal influence of x on y, avoiding the confounding effects of changes in model fit that arise from varying the number of y-lags. By remaining closely tied to the well-established LR testing framework, the GLR method is more theoretically grounded and thus better equipped to control Type I error.

To evaluate the efficacy of these Granger-based approaches, we conduct experiments using simulated time series datasets with known lag structures. Performance is assessed by comparing predicted lag orders against ground truth lags and benchmarking GMP and GLR against traditional methods, including information criteria (AIC, BIC, HQIC) and sequential LR tests. Our evaluation is guided by three hypotheses:

- The GMP method will outperform traditional information criteria approaches (AIC, BIC, HQIC) in identifying true lag structures by optimizing for Granger causal significance rather than overall model fit.
- The GLR method will outperform traditional and small-sample-corrected LR tests in identifying true lag structures due to its focus on Granger causality principles, while still maintaining the Type I error control benefits of the sequential LR testing procedure.
- The GLR method will outperform the GMP method by isolating variation in x-lags while holding y-lags constant, thereby reducing confounding effects from y in the estimation of lag order under noisy conditions.

The remainder of this paper is organized as follows. In Section 2, we review background material on Granger causality and VAR lag order selection. Section 3 surveys prior work on lag order selection in Granger causality, including comparisons between information criteria and sequential likelihood ratio testing. In Section 4, we present the proposed GMP and GLR methods and introduce the simulated datasets used for evaluation. Section 5 reports experimental analyses comparing GMP and GLR against information criteria and traditional sequential likelihood ratio tests, demonstrating improved lag selection performance and robustness across varying noise levels.

2 Background

2.1 Granger Causality

Given two time series variables, x and y, Clive Granger introduced the concept of Granger causality, a test which evaluates whether knowledge of past values of x increases the predictability of y [7]. To test Granger causality, two vector autoregressive (VAR) models are fit. The first, the restricted model, is shown in eq. (1) and includes only past values of y. The second, the unrestricted model, is shown in eq. (2) and includes past values of both x and y. In eq. (1) and eq. (2), m denotes the number of past observations (lags) included in the model.

$$\mathrm{R}_m: \quad y_t = b_0 + \sum_{i=1}^{m} b_i \, y_{t-i} + \varepsilon_t \tag{1}$$

$$\mathrm{U}_m: \quad y_t = b_0 + \sum_{i=1}^{m} b_i \, y_{t-i} + \sum_{j=1}^{m} a_j \, x_{t-j} + \eta_t \tag{2}$$

The null hypothesis, shown in eq. (3), is that there is no difference between error of the restricted and unrestricted models. The alternative hypothesis, shown in eq. (4), is that the variance of the error of the unrestricted model is significantly less than that of restricted. If the null hypothesis is rejected, this provides evidence in support of the alternative hypothesis and it is said that x Granger-causes y. Nevertheless, because Granger causality tests whether previous values

of x have a temporal relationship with y, Granger causality should be understood as the presence of predictability rather than a statement about definitive causation [19].

$$H_0 : \mathrm{Var}(\eta) = \mathrm{Var}(\varepsilon) \tag{3}$$

$$H_1 : \mathrm{Var}(\eta) < \mathrm{Var}(\varepsilon) \tag{4}$$

To perform Granger causality testing there are three common approaches. The first is block exogeneity testing, the second is an F-test between the sum of squared errors of the unrestricted and restricted models, and the third is a likelihood ratio test between the unrestricted and restricted models [19, 6]. In this work, we implement Granger causality testing using the likelihood ratio test. Since we will be comparing our Granger-based lag order selection approaches to methods that frequently rely on log-likelihoods or likelihood ratio tests, this choice ensures more direct and meaningful comparisons.

When applying ICTs or LR tests for lag order selection in Granger settings, the unrestricted model U_m is typically used, since it captures lagged effects of both x and y. However, these criteria remain oriented toward model fit, rather than explicitly aligning with the causal objective of determining whether x Granger-causes y. This gap motivates the development of lag selection methods tailored to Granger causality, such as GMP and GLR.

2.2 Lag Order Selection

A crucial consideration in Granger causality testing is determining the appropriate number of lags to include in the model. In eq. (1) and eq. (2), m represents the lag order of the model, a key parameter that determines how many past observations are included in the analysis. The selection of m is therefore a key concern in Granger causality testing and serves as the primary focus of this paper.

When fitting VAR models, choosing too small of a lag can lead to misspecified models, while choosing too large of a lag can lead to overfitting and poor forecasting performance [19, 4, 14]. Lag order selection for VARs is a widely investigated topic that generally involves the use of a criterion to determine what lag order should be used. We provide background on two categories of lag selection that employ (1) information criteria techniques (ICTs) and (2) sequential likelihood ratio (LR) testing.

Information Criteria Techniques Akaike Information Criterion (AIC) [1, 2], Schwarz Bayesian Information Criterion (BIC) [17], and Hannan-Quinn Information Criterion (HQIC) [8] are ICTs which are commonly applied to lag order selection. Defined below, these criteria all incorporate the number of model parameters ($|\theta|$) and the maximum value of the likelihood function ($\hat{L}$) evaluated at the estimated parameters.

$$\text{AIC} = 2 \cdot |\theta| - 2\ln(\hat{L}) \tag{5}$$

$$\text{BIC} = |\theta| \cdot \ln(n) - 2\ln(\hat{L}) \tag{6}$$

$$\text{HQIC} = 2 \cdot |\theta| \cdot \ln(\ln(n)) - 2\ln(\hat{L}) \tag{7}$$

AIC, BIC, and HQIC all aim to strike a balance between two components: the complexity penalty (first term in each equation) and the goodness of fit (last term in each equation). The complexity penalty discourages overfitting by penalizing models with more parameters, while the goodness of fit reflects how well the model performs on the training data. For each candidate lag $m \in \{1, \ldots, M\}$, an information criterion value is computed, and the lag with the minimum criterion value is selected as $\hat{m}$, as smaller criteria values indicate a model which is fit well without being overspecified. When applying these criteria in the context of Granger causality, we compute all information criteria using the unrestricted Granger model (U_m, shown in eq. (2)), excluding the restricted model since U_m captures the interactions between the dependent and independent variables.

Sequential Likelihood Ratio (LR) Tests Another approach to lag order selection involves a sequential likelihood ratio (LR) test [9, 14]. This test compares sequential models in a descending order starting from the maximum lag. It aims to identify a significant drop in performance from one model to the next. Such a difference indicates that removing the previous lag significantly impedes the model's performance, thereby identifying the optimal lag to characterize the time-lagged relationship.

Specifically, considering a maximum lag M, and a counter $i = 1, 2, \ldots, M$, the null hypothesis is that the model at $M - i$ provides an adequate fit to the data, and is conditioned on all previous null hypotheses being true. The alternative hypothesis is that the model at $M - i + 1$ provides a significantly better fit. When an individual likelihood test yields a statistically significant result, the lag of $M - i + 1$ is chosen as the optimal lag, $\hat{m}$. The test statistic for the LR test is shown in eq. (8), where $\hat{L}_m$ is the maximum likelihood of a VAR at lag m and N is the effective sample size.

$$\lambda_{\text{LR}}(i) = N \left[\ln(\hat{L}_{M-i}) - \ln(\hat{L}_{M-i+1}) \right], \quad i = 1, 2, \ldots, M \tag{8}$$

The LR test is typically utilized with K-dimensional VARs where all variables are jointly modeled. That is, every variable depends on previous values of itself and all other variables. In such settings, each successive lag introduces K^2 additional parameters. This differs from the context of bivariate Granger causality testing, where the unrestricted model at each lag (U_m, shown in eq. (2)) includes only one dependent variable modeled on its own past values and the past values of the independent variable. In this case, each additional lag adds only two parameters, one for the lagged dependent variable and one for the lagged independent

variable, as opposed to K^2 parameters in multivariate VARs. Accordingly, when applying sequential LR tests in the context of Granger causality, we do so using the unrestricted Granger model U_m, comparing sequential instances across lags. Each test statistic follows a $\chi^2(K)$ distribution, where $K = 2$ in our case, since two additional parameters are introduced at each lag.

In contexts where N is not sufficiently large compared to the number of parameters, a small-sample correction can be applied to create the small-sample correction likelihood ratio (SLR) test statistic, as shown in eq. (9) [9, 21]. This correction is represented by the term S, defined in eq. (10).

$$\lambda_{\mathrm{SLR}}(i) = S\left[\ln(\hat{L}_{M-i}) - \ln(\hat{L}_{M-i+1})\right] \tag{9}$$

$$S = N - (M - i + 1)K \tag{10}$$

It is important to distinguish the overall Type I error probability of the sequential LR testing procedure from the significance threshold of the individual tests [14]. As the number of individual tests increases, the cumulative probability of making at least one Type I error also increases. However, it is difficult to vary the individual significance thresholds at different lags in a way that appropriately minimizes Type I error [14]. Therefore, we rely on a fixed significance threshold when conducting each individual likelihood ratio test. This approach is summarized by Lütkepohl who also defines the approximate probability of Type I error as is shown in eq. (11), where ε_i represents the cumulative probability of a Type I error after i tests, and γ_i denotes the significance level of the i-th individual test. Note that one may also use eq. (11) to control the overall Type I error probability by solving for a common individual significance level corresponding to a desired cumulative error rate; however, this method still requires that the significance level remain constant across all lags.

$$\varepsilon_i = 1 - (1 - \gamma_1) \cdots (1 - \gamma_i), \quad i = 1, 2, \ldots, M. \tag{11}$$

When applying the sequential LR and small-sample correction (SLR) tests, we follow Ivanov and Kilian's convention by using a 0.01 significance level for individual tests [9]. This threshold is commonly adopted in statistical analyses and aligns with established norms. While Ivanov and Kilian also report results using a 0.05 threshold, they limit their maximum lag to 12. In contrast, our investigated dataset (described in Section 4.3) employs a maximum lag of 21. In this case, utilizing a significance threshold of 0.05 would substantially increase overall probability of Type I error, as indicated by eq. (11). For this reason, we maintain the more conservative 0.01 threshold throughout our experiments.

3 Related Work

Several works have considered different lag selection methods in the context of VAR models. In practice, there is never one criterion that consistently outperforms the others, suggesting that the optimum criterion choice is often a matter

of application and dataset [9]. AIC has been shown to perform well at small sample sizes [12]. However, AIC can be prone to overfitting, with BIC often being more robust [10]. Additional work has shown that BIC and HQIC consistently perform well at true lag order prediction and are a reasonable choice of criteria when there is limited selection information [13]. Practically, the use of AIC, BIC, and HQIC is similar to the use of sequential LR tests, with the penalty term having a similar effect on the criterion values as the critical value on the outcome of the sequential LR tests [9]. Some experimental evidence shows that, in certain circumstances, the information criteria dominate the sequential tests [9]. Works also note that when dealing with small samples, the LR test is a particularly poor performer while the SLR test is more adaptable in such circumstances [9, 13].

Investigation into lag order selection techniques is typically limited to VARs and extensive investigations are lacking in Granger causality contexts specifically. A search of the lag space is often required to ensure robust Granger causality testing [22] and many works rely on choosing lag based on the minimum of an ICT [4, 22]. In cases where time delays are not fixed but variable, previous research has explored adaptations of Granger causality to account for these variations [3]. In high-dimensional and multivariate settings, regularization-based approaches have been proposed to shrink coefficients and determine lag order [15, 20, 18]. In contrast, our work focuses on the traditional bivariate Granger causality framework, where a single predictor (x) and response (y) are considered, and the goal is to identify a single optimal lag order.

4 Methodology

We present two methods for selecting lag order: the Granger Minimum p-value method (GMP) and the Granger sequential Likelihood Ratio test (GLR). We describe each method in turn and then describe a simulated dataset designed to test the efficacy of each.

4.1 Granger Minimum p-value (GMP)

When using ICTs (such as AIC, BIC, and HQIC) to select the lag order in a VAR model, one typically chooses the lag that minimizes the chosen criterion. Each information criterion balances model fit (i.e., how well the model explains the data) against model complexity (i.e., the number of parameters), penalizing models that increase complexity without substantially improving fit. However, in a Granger causality context, we can exploit more than just fit-based information regarding a single model. Specifically, because we are interested in whether one time series "Granger-causes" another, it can be effective to select the lag at which the strongest apparent causal influence occurs. To incorporate this notion of influence directly into the selection process, we replace the concept of "maximizing goodness-of-fit" used in standard ICTs with the concept of "maximizing

inferred causal influence." In practice, this can be done by identifying the lag for which the strongest evidence for Granger causality is present.

To execute this, we perform Granger causality testing by conducting likelihood ratio tests at each lag between a restricted model (R_m) and unrestricted model (U_m). That is, for each lag m, we compare the likelihood of R_m (which only includes m-lags of y (eq. (1))) to the likelihood of U_m (which contains m-lags for both x and y (eq. (2))) [3]. We choose an LR-based approach to Granger causality testing because it closely aligns with the traditional ICTs use of log-likelihoods in the calculation of goodness-of-fit. Except instead of leveraging the goodness-of-fit of just one model, we leverage the difference in fit between restricted and unrestricted models in line with the principles of Granger causality. The calculation of the test statistic (λ_{GC}) is defined in eq. (12), where $\hat{L}$ denotes the maximum value of the likelihood function evaluated at the estimated parameters.

$$\lambda_{\mathrm{GC}}(m) = -2(\ln(\hat{L}_{\mathrm{R}_m}) - \ln(\hat{L}_{\mathrm{U}_m})), \quad m = 1, 2, 3, \ldots, M. \tag{12}$$

To arrive at a set of p-values we use a chi-square distribution where the degrees of freedom is the difference in parameters between the restricted and unrestricted models. In this context, the difference is always equal to the number of lags used, i.e., m. Thus, given $\lambda_{\mathrm{GC}}(m)$, we use a $\chi^2(m)$ distribution to arrive at a final p-value for each lag m. These p-values offer insight into evidence of Granger causality at each lag.

Intuitively, smaller p-values indicate greater confidence that the hypothesized causal influence is present at the given lag. Thus we select the lag which yields the smallest (i.e., the most significant) Granger causality p-value, as is shown in eq. (13) where p_m is the p-value at lag m and $\hat{m}$ is the selected lag. We refer to this new method as Granger Minimum p-value (GMP). Importantly, because we select the minimum p-value, any adjustments via Type I error rate control or false discovery rate methods would not alter the ranking; hence, additional multiple testing corrections are unnecessary.

$$\hat{m} = \arg \min_{m \in \{1,\ldots,M\}} p_m \tag{13}$$

By identifying the lag with the strongest evidence of Granger causality, GMP bases its decision of lag on "maximizing inferred causal influence" rather than on "maximizing goodness-of-fit," as in traditional ICTs. Further, similar to how ICTs directly compute a complexity penalty, the p-value lookup in GMP inherently penalizes additional complexity: as m increases, the degrees of freedom

[3] This process should not be confused with that of sequential LR tests, which compare model fit *at one lag to the next*. This application of a likelihood ratio test specifically for Granger causality limits testing to a restricted and unrestricted model *at the same lag*.

also increase, and consequently the p-value is penalized unless the added complexity substantially improves observed causal influence. A visual illustration of this behavior is provided in Appendix B.

4.2 Granger Sequential Likelihood Ratio Test (GLR)

In Granger causality analyses, the central objective is to determine whether x Granger-causes y. From this perspective, selecting the lag where x appears to exert the strongest influence on y (as in the GMP approach) is intuitively appealing. However, while GMP offers a direct way to capture maximal influence, it is best viewed as a heuristic; it does not provide a predictable or theoretically grounded mechanism for controlling Type I error.

In contrast, sequential likelihood ratio tests (LR tests) offer a more formalized procedure for controlling Type I error because they test nested models in a descending manner. Even if the null hypothesis is rejected prematurely, the selected model still retains higher-order terms that nest the true lag structure. Note that Type I errors are still possible (i.e., a late rejection where a lag below the true order is chosen); however, early rejections do not eliminate the correct lag from the model space. As a result, this approach naturally guards against spurious findings in a way that GMP, by design, cannot.

The LR and SLR test sequential models where each model is specified in accordance with eq. (2). However, as the number of lags changes in the traditional unrestricted Granger model, shown in eq. (2), both the number of y-lags and the number of x-lags change. To isolate the varying x-lags, we adapt eq. (2) for the application of a sequential Granger-based LR test (GLR). Equation (14) shows this adapted unrestricted model estimated at lag m, where the autoregressive order of y_t is fixed at M lags while the number of x_t lags is set to m. We fit models for $m = 1, 2, \ldots, M$, such that the only difference between models is the varying number of x-lags, which vary in accordance with m. Then these models can be compared using the traditional sequential LR or SLR frameworks. Thus, the key difference between the GLR and traditional LR testing lies in the specification of the underlying model used.

In this work, we implement the GLR with a small-sample correction in accordance with eq. (9) and use a significance level of 0.01. The GLR employs a chi-square distribution with one degree of freedom, as each subsequent model varies by only one term. This differs from the LR and SLR, which use two degrees of freedom because both x- and y-lags vary between models.

$$y_t = b_0 + \sum_{i=1}^{M} b_i \, y_{t-i} + \sum_{j=1}^{m} a_j \, x_{t-j} + \eta_t, \quad m = 1, 2, \ldots, M. \tag{14}$$

By adapting the sequential likelihood ratio framework so that each nested model differs only in the number of x-lags included, while keeping the number of y-lags fixed, we effectively merge the principles of Granger causality with the safeguards of sequential LR tests. The GLR approach ensures that the selected

lag reflects x's direct temporal influence on y, aligning naturally with the fundamental principle of Granger causality. While this modification is straightforward in implementation, it could significantly impact lag selection and therefore Granger causality estimates in causal inference.

4.3 Dataset

We utilize simulated datasets with known causal influences and lag labels to enable ground-truth evaluation, which is infeasible with real-world data. Each dataset consists of two sets of time series, $\mathbf{X}$ and $\mathbf{Y}$, where subsets of $\mathbf{X}$ influence subsets of $\mathbf{Y}$ through controlled time-lagged dependencies.

Specifically, 24 $\mathbf{X}$ series are created by inserting impulse spikes, applying smoothing, and adding Gaussian noise. Then, 18 $\mathbf{Y}$ series are generated in groups, where each group of $\mathbf{Y}$ is derived from the pointwise sum of a corresponding group of $\mathbf{X}$ series. Within each group, the resulting summed signal is shifted by a random temporal lag and perturbed with additional Gaussian noise to create individual $\mathbf{Y}$ series. Consequently, only those $\mathbf{X}$–$\mathbf{Y}$ pairs belonging to the same group share a ground-truth causal relationship, while all other cross-group pairs are non-causal. The applied temporal shift (i.e., true lag) for each causal $\mathbf{X}$–$\mathbf{Y}$ pair is randomly selected from a uniform range between 7 and 21. We explicitly record which $\mathbf{X}$ series contribute to each $\mathbf{Y}$ and the corresponding lag, enabling ground-truth evaluation of both whether a method can detect a time-lagged relationship and whether it can correctly identify the lag that characterizes it.

We generate 100 independent $\mathbf{X}$ datasets. For each $\mathbf{X}$ dataset, three corresponding $\mathbf{Y}$ datasets are created, one for each noise level, with added Gaussian noise of standard deviations $\sigma = 1$, $\sigma = 2$, and $\sigma = 3$, corresponding to approximate SNRs of -5, -10, and -13 dB, respectively.[4] This replication allows us to report averaged performance metrics and assess the consistency of each method under repeated random variation. For completeness, pseudocode for generating $\mathbf{X}$ and $\mathbf{Y}$ is provided in Appendix A.

5 Results

5.1 Causality Identification

To assess the ability of the sequential LR tests to correctly identify the presence or absence of time-lagged relationships, we report a range of performance metrics[5] in Table 1. The reported values represent averages across all 100 experiments. For each experiment, these metrics are calculated at each noise level across all time series pairs between $\mathbf{X}$ and $\mathbf{Y}$, so that some pairs contain true

[4] In Tables 1 to 4, noise levels are denoted as $\mathcal{N}$, and $\sigma = 1$, $\sigma = 2$, and $\sigma = 3$ are abbreviated as σ_1, σ_2, and σ_3, respectively, to conserve horizontal space.

[5] All sequential LR-based tests (LR, SLR, and GLR) are implemented with a conservative significance threshold of 0.01.

time-lagged relationships while others do not. In this evaluation, lag identification is treated as a positive outcome and non-identification as a negative outcome. We limit our analysis to the sequential LR tests and exclude the ICTs, which always select a lag based on the minimum criterion and, consequently, prevent a meaningful analysis of false positives and true negatives. That is, ICTs always report a false positive rate of 100%. Because averages can obscure variability across trials, we also include boxplots in Figure 1 to show the distribution of the various metrics across the 100 experiments for each noise level.

In low-noise conditions, GLR outperforms both LR and SLR across all metrics (except recall, where all methods tie), as shown in Table 1. As noise increases, GLR maintains notably higher average recall than LR and SLR. This trend is reinforced by the recall boxplots in Figure 1, where GLR consistently achieves higher recall, indicating that it is better at identifying true positives and avoiding false negatives.

GLR generally reports higher average accuracy and F1 scores than LR and SLR, with the exceptions of F1 at medium noise and accuracy at high noise

Table 1. Average metrics for causality identification using sequential LR Methods. The highest value in each row is shown in **bold**.

$\mathcal{N}$	SLR			LR			GLR (ours)		
	Acc	Prec / Rec	F1	Acc	Prec / Rec	F1	Acc	Prec / Rec	F1
σ_1	0.801	0.464 / **1.000**	0.631	0.786	0.445 / **1.000**	0.614	**0.841**	**0.514** / **1.000**	**0.678**
σ_2	0.842	**0.523** / 0.997	**0.683**	0.827	0.500 / 0.998	0.663	**0.843**	0.516 / **0.999**	0.680
σ_3	**0.831**	**0.505** / 0.889	0.641	0.818	0.483 / 0.900	0.626	**0.831**	0.497 / **0.934**	**0.649**

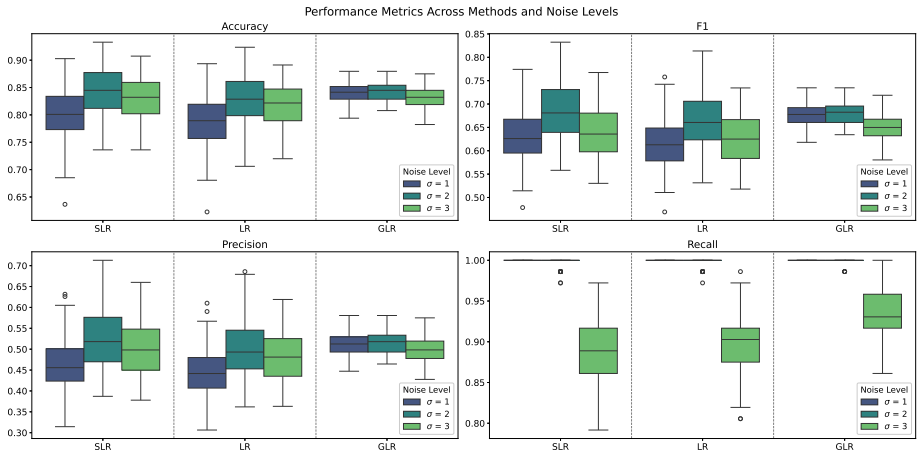

Fig. 1. Distributions of accuracy, precision, recall, and F1 across 100 experiments for each sequential LR method and noise level. In the recall plots (bottom right), both low- and medium-noise conditions yielded near-perfect performance, with quartiles and whiskers collapsing to 1.0 except for a few outliers.

(where it ties SLR). Moreover, across all metrics in Figure 1, GLR exhibits a tighter interquartile range (IQR) and shorter whiskers than the other methods, reflecting less variability across experiments.

We also report average false positive rates (FPR) and false negative rates (FNR) across the 100 experiments in Table 2, along with boxplots showing their distributions in Figure 2. These results provide two key insights. First, as noise increases, both SLR and LR exhibit progressively lower FPRs but higher FNRs. In contrast, GLR maintains a relatively stable FPR while its FNR increases more slowly than SLR or LR. Combined with the recall results, this suggests that as noise grows, LR and SLR increasingly bias toward non-identification of lag, whereas GLR achieves a better balance between identification and non-identification. Second, regarding Type I error, Equation (11) indicates that a sequential LR test with a 0.01 significance threshold and a maximum lag of

Table 2. Average false positive (FPR) and false negative rates (FNR) across 100 experiments for each sequential LR method.

$\mathcal{N}$	SLR		LR		GLR (ours)	
	FPR	FNR	FPR	FNR	FPR	FNR
σ_1	0.239	0.000	0.256	0.000	0.191	0.000
σ_2	0.189	0.003	0.207	0.002	0.188	0.001
σ_3	0.180	0.111	0.199	0.100	0.190	0.066

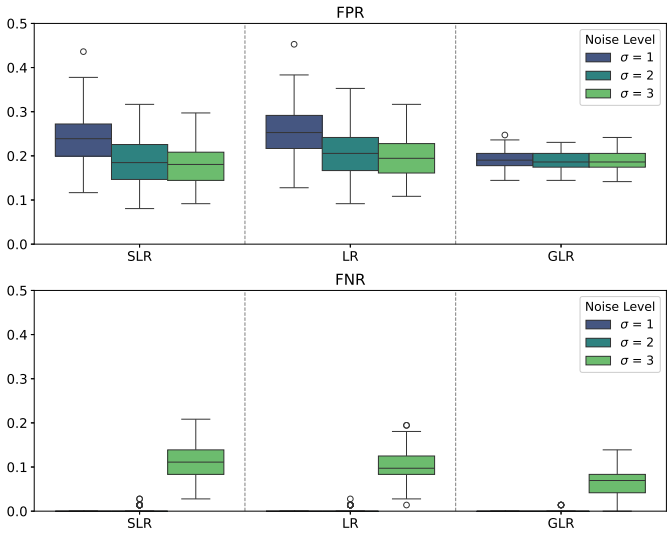

Fig. 2. Distribution of FPR and FNR across experiments for each method. In the FNR plots (bottom), both low- and medium-noise conditions yielded near-perfect performance, with quartiles and whiskers collapsing to zero except for a few outliers.

21 should yield an approximate Type I error rate of 19.02%. The GLR's FPR distribution is more tightly centered around this expected value than that of LR or SLR, suggesting that it more faithfully adheres to theoretical error control.

5.2 Evaluation of Lag Estimation Accuracy

Next, we evaluate GMP and GLR's ability to recover the correct lag order in time series pairs where a causal relationship exists (true positives). In Table 3, we report the average mean absolute error (MAE) for each method, computed by comparing predicted lag values against the ground truth lags recorded during dataset generation. Because the metric represents error, lower values indicate better performance. Within ICTs, GMP significantly outperforms other criteria at medium- and high-noise. Among LRs, GLR achieves the best performance under medium- and high-noise conditions, although SLR has an advantage under low noise.

Table 3. Average mean absolute errors (MAEs) for lag order prediction across 100 experiments, restricted to true positive causal pairs. Results are shown separately for ICT and LR groups, with the best-performing method within each group and noise level shown in **bold**.

$\mathcal{N}$	ICTs				Sequential LR Tests		
	AIC	BIC	HQIC	GMP (ours)	LR	SLR	GLR (ours)
σ_1	**0.430**	10.838	2.890	0.465	0.432	**0.369**	0.415
σ_2	4.152	12.489	11.709	**1.035**	0.469	0.417	**0.372**
σ_3	9.749	12.584	12.524	**2.197**	0.623	0.556	**0.508**

Figure 3 shows boxplots of the raw lag prediction errors for the ICT methods, where error is defined as the ground truth lag minus the predicted lag. Thus, positive values correspond to underpredictions (choosing a lag smaller than the ground truth), while negative values correspond to overpredictions (choosing a lag larger than the ground truth). Underpredictions are undesirable because they do not nest the true lag structure, whereas overpredictions preserve it. The plot highlights GMP's robustness to noise relative to AIC, BIC, and HQIC. Both BIC and HQIC exhibit a consistent bias toward underprediction across all noise levels, and AIC shows a similar tendency at medium and high noise. By comparison, when GMP does err, it tends to overpredict, which is more tolerable since the true lag remains nested within the larger structure.

We omit raw error boxplots for the likelihood ratio (LR) tests, as all three variants yield $Q_1 = Q_3 = 0$—indicating that at least 75% of signed errors are exactly zero. This concentration of values collapses the interquartile range and undermines the interpretability of a boxplot visualization. Instead, we report raw counts of exact predictions, overpredictions, and underpredictions of lag in Table 4.

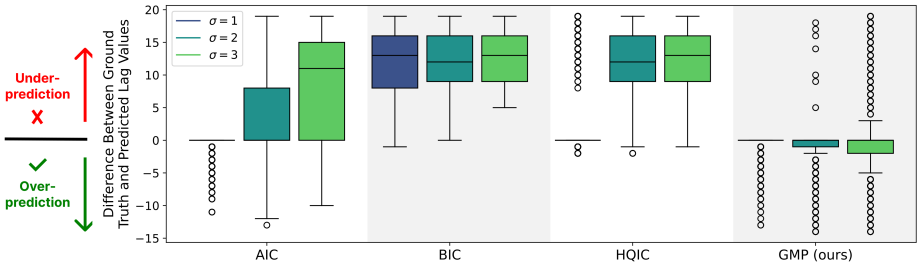

Fig. 3. Raw lag prediction errors for ICT methods (AIC, BIC, HQIC, and GMP). Each boxplot summarizes 7,200 values (72 causal relationships across 100 experiments). Error is defined as ground truth lag minus predicted lag.

Table 4. Lag prediction outcome counts for sequential LR methods. For each noise condition, N_{TP} denotes the number of true positives identified; the remaining columns report whether the selected lag was exact, an overprediction, or an underprediction.

$\mathcal{N}$	SLR				LR				GLR (ours)			
	N_{TP}	Exact	Over	Under	N_{TP}	Exact	Over	Under	N_{TP}	Exact	Over	Under
σ_1	7200	6688	512	0	7200	6611	589	0	7200	6637	563	0
σ_2	7180	6648	531	1	7183	6575	607	1	7195	6657	537	1
σ_3	6399	5747	537	115	6478	5744	620	114	6724	6125	529	70

N_{TP} denotes the number of true positives, cases where a ground-truth causal relationship existed and was correctly identified. However, even when a ground-truth causal relationship is identified for a time series pair, an underprediction of the actual lag that characterizes that relationship can still occur. Such an underprediction is inherently a false positive. Of particular interest in Table 4 is the high-noise condition (bottom row). Under this condition, GLR detects substantially more true positives than either SLR or LR ($N_{\mathrm{TP}} = 6724$ vs. 6399 and 6478, respectively), consistent with its higher recall observed in Figure 1; moreover, GLR achieves more exact predictions of lag (6125 vs. 5747 for SLR and 5744 for LR). Furthermore, despite its larger effective sample size, GLR reports fewer underpredictions (70 vs. 115 and 114). These results indicate that GLR not only recovers more causal pairs under challenging noise but also does so with less risk of underestimation.

6 Conclusions

Our results show that lag selection approaches guided by Granger-specific principles can outperform traditional approaches such as information criteria tests (ICTs) and sequential likelihood ratio (LR) tests. Both the Granger Minimum p-value (GMP) and sequential Granger-Likelihood Ratio (GLR) methods align lag selection directly with the objectives of Granger causality testing, yielding

improvements in identifying true causal relationships and maintaining robustness to noise. The GMP method optimizes lag selection by maximizing Granger causal significance rather than overall model fit, allowing it to outperform traditional ICTs, particularly under medium- and high-noise conditions. The GLR method extends the sequential LR framework to isolate the influence of x-lags while holding y-lags fixed, thereby preserving Type I error control while improving sensitivity to causal structure. Across experiments, GLR demonstrated consistent performance across noise levels and greater recall of true causal influences compared to standard LR and SLR methods.

While both methods improve lag order selection for Granger causality estimation, they are suited to different analytical goals. When a definitive lag order is required for downstream applications or modeling tasks, GMP may be preferable since, like traditional ICTs, it always selects a lag. Conversely, when statistical rigor and Type I error control are of greater concern, GLR provides a principled alternative that integrates the advantages of sequential LR testing with Granger-specific considerations. By improving the reliability of lag order selection, these methods have direct implications for downstream time series tasks, including forecasting, anomaly detection, and causal inference, where accurate temporal characterization is critical for robust modeling and interpretation. Overall, these findings demonstrate that aligning lag order selection directly with the principles of Granger causality improves both interpretability and reliability in causal time series analysis, particularly under noisy conditions.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (1974). DOI 10.1109/TAC.1974.1100705. Conference Name: IEEE Transactions on Automatic Control
2. Akaike, H.: On the Likelihood of a Time Series Model. Journal of the Royal Statistical Society. Series D (The Statistician) **27**(3), 217–235 (1978). DOI 10.2307/2988185. URL https://www.jstor.org/stable/2988185. Publisher: [Royal Statistical Society, Wiley]
3. Amornbunchornvej, C., Zheleva, E., Berger-Wolf, T.Y.: Variable-Lag Granger Causality for Time Series Analysis. In: 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 21–30 (2019). DOI 10.1109/DSAA.2019.00016. URL https://ieeexplore.ieee.org/document/8964168
4. Bose, E., Hravnak, M., Sereika, S.M.: Vector Autoregressive (VAR) Models and Granger Causality in Time Series Analysis in Nursing Research: Dynamic Changes Among Vital Signs Prior to Cardiorespiratory Instability Events as an Example. Nursing research **66**(1), 12–19 (2017). DOI 10.1097/NNR.0000000000000193
5. Chen, Y., Yang, K., An, Z., Holder, B., Paloutzian, L., Bali, K.M., Du, W.: MARLP: Time-series Forecasting Control for Agricultural Managed Aquifer Recharge. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, pp. 4862–4872. Association for Computing Machinery, New York, NY, USA (2024). DOI 10.1145/3637528.3671533. URL https://dl.acm.org/doi/10.1145/3637528.3671533
6. Cromwell, J., Hannan, M., Labys, W., Terraza, M.: Multivariate Tests for Time Series Models. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America (1994). DOI 10.4135/9781412985239. URL https://methods.sagepub.com/book/multivariate-tests-for-time-series-models
7. Granger, C.W.J.: Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica **37**(3), 424–438 (1969). DOI 10.2307/1912791. Publisher: [Wiley, Econometric Society]
8. Hannan, E.J., Quinn, B.G.: The Determination of the Order of an Autoregression. Journal of the Royal Statistical Society. Series B (Methodological) **41**(2), 190–195 (1979). DOI https://doi.org/10.1111/j.2517-6161.1979.tb01072.x. URL https://www.jstor.org/stable/2985032. Publisher: [Royal Statistical Society, Oxford University Press]
9. Ivanov, V., Kilian, L.: A Practitioner's Guide to Lag Order Selection For VAR Impulse Response Analysis. Studies in Nonlinear Dynamics & Econometrics **9**(1), 1–36 (2005). DOI 10.2202/1558-3708.1219
10. Koehler, A.B., Murphree, E.S.: A Comparison of the Akaike and Schwarz Criteria for Selecting Model Order. Journal of the Royal Statistical Society. Series C (Applied Statistics) **37**(2), 187–195 (1988). DOI 10.2307/2347338. URL https://www.jstor.org/stable/2347338. Publisher: [Royal Statistical Society, Oxford University Press]
11. Lee, M., Sylvester, J., Aggarwal, S., Sinha, A., Taylor, M., Srirama, N., Larson, E., Thornton, M.: Side Channel Identification using Granger Time Series Clustering with Applications to Control Systems:. In: Proceedings of the 8th International Conference on Information Systems Security and Privacy, pp. 290–298. SCITEPRESS - Science and Technology Publications (2022). DOI 10.5220/0010781600003120. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010781600003120

12. Liew, V.K.S.: Which Lag Length Selection Criteria Should We Employ? Economics Bulletin **3**(33), 1–9 (2004). URL https://ssrn.com/abstract=885505
13. Lütkepohl, H.: Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process. Journal of Time Series Analysis **6**(1), 35–52 (1985). DOI 10.1111/j.1467-9892.1985.tb00396.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.1985.tb00396.x
14. Lütkepohl, H.: Introduction to multiple time series analysis. Springer (2005). URL https://link.springer.com/content/pdf/10.1007/978-3-540-27752-1.pdf
15. Nicholson, W.B., Matteson, D.S., Bien, J.: VARX-L: Structured regularization for large vector autoregressions with exogenous variables. International Journal of Forecasting **33**(3), 627–651 (2017). DOI 10.1016/j.ijforecast.2017.01.003
16. Qiu, H., Liu, Y., Subrahmanya, N.A., Li, W.: Granger Causality for Time-Series Anomaly Detection. In: 2012 IEEE 12th International Conference on Data Mining, pp. 1074–1079. IEEE (2012). DOI 10.1109/ICDM.2012.73. URL https://ieeexplore.ieee.org/document/6413806. ISSN: 2374-8486
17. Schwarz, G.: Estimating the Dimension of a Model. The Annals of Statistics **6**(2), 461–464 (1978). DOI 10.1214/aos/1176344136. Publisher: Institute of Mathematical Statistics
18. Shojaie, A., Basu, S., Michailidis, G.: Adaptive Thresholding for Reconstructing Regulatory Networks from Time-Course Gene Expression Data. Statistics in Biosciences **4**(1), 66–83 (2012). DOI 10.1007/s12561-011-9050-5
19. Shojaie, A., Fox, E.B.: Granger Causality: A Review and Recent Advances. Annual Review of Statistics and Its Application **9**(Volume 9, 2022), 289–319 (2022). DOI 10.1146/annurev-statistics-040120-010930. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-040120-010930. Publisher: Annual Reviews
20. Shojaie, A., Michailidis, G.: Discovering graphical Granger causality using the truncating lasso penalty. Bioinformatics (Oxford, England) **26**(18), i517–523 (2010). DOI 10.1093/bioinformatics/btq377
21. Sims, C.A.: Macroeconomics and Reality. Econometrica **48**(1), 1–48 (1980). DOI 10.2307/1912017. URL https://www.jstor.org/stable/1912017. Publisher: [Wiley, Econometric Society]
22. Thornton, D.L., Batten, D.S.: Lag-Length Selection and Tests of Granger Causality Between Money and Income. Journal of Money, Credit and Banking **17**(2), 164–178 (1985). DOI 10.2307/1992331. Publisher: [Wiley, Ohio State University Press]
23. Yehuda, Y., Freedman, D., Radinsky, K.: Self-supervised Classification of Clinical Multivariate Time Series using Time Series Dynamics. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, pp. 5416–5427. Association for Computing Machinery, New York, NY, USA (2023). DOI 10.1145/3580305.3599954. URL https://dl.acm.org/doi/10.1145/3580305.3599954
24. Zhang, Z., Ren, S., Qian, X., Duffield, N.: Learning Flexible Time-windowed Granger Causality Integrating Heterogeneous Interventional Time Series Data. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, pp. 4408–4418. Association for Computing Machinery, New York, NY, USA (2024). DOI 10.1145/3637528.3672023. URL https://dl.acm.org/doi/10.1145/3637528.3672023

A Simulated Datasets Pseudocode

In algorithms 1 and 2 we detail the pseudocode for creating the simulated datasets $\mathbf{X}$ and $\mathbf{Y}$. For algorithm 1, we use $N_{\mathrm{X}} = 24$ to generate 24 independent $\mathbf{X}$ time series. For algorithm 2, we use $\mathrm{gsize}_{\mathrm{X}} = 4$, $\mathrm{gsize}_{\mathrm{Y}} = 3$, and $N_{\mathrm{groups}} = 6$; this configuration produces 18 time series in each $\mathbf{Y}$ dataset when $N_{\mathrm{X}} = 24$. Each group of four $\mathbf{X}$ series jointly influences three corresponding $\mathbf{Y}$ series, such that each causal Y is generated from a subset of four $\mathbf{X}$ series via a pointwise sum and a random temporal shift (lag) selected from 7 to 21. This results in six groups of related series and a total of 72 causal $\mathbf{X}$–$\mathbf{Y}$ pairs per dataset. In algorithm 2, the noise variable controls the standard deviation of added Gaussian noise, which is varied from 1 to 3 to produce the three different $\mathbf{Y}$ datasets generated for each $\mathbf{X}$. We conduct 100 experiments, generating 100 independent $\mathbf{X}$ datasets, each paired with three corresponding $\mathbf{Y}$ datasets at the three different noise levels ($n = 1$, $n = 2$, $n = 3$).

Algorithm 1: Simulated data generation algorithm for creating the $\mathbf{X}$ time series dataset.

Input: Number of time series in $\mathbf{X}$ (N_{X})
Output: Dataset $\mathbf{X}$ containing N_{X} time series
1. Initialize N_{X} time series with zeroes, each time series with 1000 points;
2. Randomly add 1s (spikes) to each time series according to $\mathcal{N}(0, 1) > 1.5$;
3. Smooth each time series using a sinc function (to create structured signals from the added spikes) ;
4. Add Gaussian noise to each with $\mathcal{N}(0, 0.3)$;

Algorithm 2: Given $\mathbf{X}$ time series, generate related $\mathbf{Y}$ time series.

Input: Dataset $\mathbf{X}$, noise level (n), X group size ($\mathrm{gsize}_{\mathrm{X}}$), Y group size ($\mathrm{gsize}_{\mathrm{Y}}$), number of groups ($N_{\mathrm{groups}}$)
Output: Dataset $\mathbf{Y}$
$c = 0$;
while $c < N_{\mathrm{groups}}$ **do**
  $X_{\mathrm{temp}} = \mathbf{X}[\mathrm{gsize}_{\mathrm{X}} \cdot c : \mathrm{gsize}_{\mathrm{X}} \cdot c + \mathrm{gsize}_{\mathrm{X}}]$;
  $i = 0$;
  while $i < \mathrm{gsize}_{\mathrm{Y}}$ **do**
    Take a pointwise sum of the members of X_{temp};
    Shift lag, m, to the right by a random amount (7 to 21);
    Add Gaussian noise $\mathcal{N}(0, n)$;
    Append to $\mathbf{Y}$ dataset;
    Save m as ground truth lag for time series pair;
    $i = i + 1$;
  end
end

B Visual Example for Single Time Series Pair

The plots in Figure 4 illustrate the motivation behind the GMP approach by breaking down the components of the ICTs across lags for a low-noise time series pair with a true lag of 10. In the leftmost plot, the true lag is visually evident in the AIC, BIC, and HQIC curves. However, only AIC and HQIC report a minimum criterion value at a lag of 10. BIC's minimum value occurs at a lag of 1, suggesting that model fit and additional complexity are not effectively balanced. This imbalance is further demonstrated in the middle plot, which separates the complexity penalty and goodness-of-fit components. In the case of BIC, the complexity penalty is too strong, overpowering the goodness of fit term and leading to a significant underprediction of the lag.

In contrast, the rightmost plot shows that the Granger causality p-values drop markedly at a lag of 10, but do not descend further. The inset plot shows that, for lags greater than 10, the additional degrees of freedom impose an effective penalty, preventing the p-value from decreasing further. Ultimately, this results in a global minimum occurring at the optimal lag of 10.

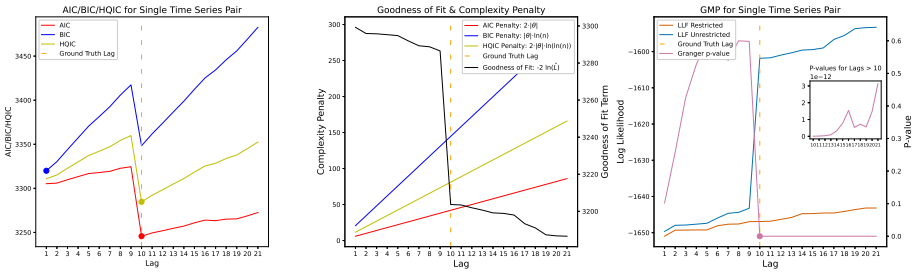

Fig. 4. ICT for one low noise ($\sigma = 1$) time series pair with a true time-lagged relationship characterized by a lag of 10. Left: Plotting the traditional ICT (AIC, BIC, HQIC) across 21 lags with IC minimums indicated by dots. Middle: Plotting the goodness of fit and complexity penalty terms for AIC, BIC, HQIC. Right: Plotting the log likelihood functions across 21 lags for the restricted and unrestricted Granger causality models. The associated Granger causality p-values are also plotted with the minimum indicated by a dot. The inset figure decreases the y-axis's scale and shows the Granger p-values past a lag of 10.