The EAS approach for graphical selection consistency in vector autoregression models

Jonathan P. WILLIAMS¹*, Yuying XIE², and Jan HANNIG³

¹Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA
²Department of Statistics and Probability, Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, Michigan, USA
³Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina, USA

Key words and phrases: Empirical Bayes; generalized fiducial inference; graph selection; high-dimensional model selection; large-sample properties.

MSC 2020: Primary 62A01; Secondary 62M10.

Abstract: As evidenced by various recent and significant papers within the frequentist literature, along with numerous applications in macroeconomics, genomics, and neuroscience, there continues to be substantial interest in understanding the theoretical estimation properties of high-dimensional vector autoregression (VAR) models. To date, however, while Bayesian VAR (BVAR) models have been developed and studied empirically (primarily in the econometrics literature), there exist very few theoretical investigations of the repeated-sampling properties for BVAR models in the literature, and there exist no generalized fiducial investigations of VAR models. In this direction, we construct methodology via the \( \epsilon \)-admissible subsets (EAS) approach for inference based on a generalized fiducial distribution of relative model probabilities over all sets of active/inactive components (graphs) of the VAR transition matrix. We provide a mathematical proof of pairwise and strong graphical selection consistency for the EAS approach for stable VAR(1) models, and demonstrate empirically that it is an effective strategy in high-dimensional settings.

Résumé: Comme en témoignent plusieurs articles récents et importants dans la littérature fréquentiste, ainsi que de nombreuses applications en macroéconomie, génomique et neurosciences, l’intérêt pour la compréhension des propriétés théoriques d’estimation de modèles auto-régressifs vectoriels de grandes dimensions (VAR) demeure important. Toutefois, à ce jour, bien que des modèles bayésiens de VAR (BVAR) aient été développés et étudiés empiriquement (principalement dans la littérature économétrique), la littérature renferme très peu d’études théoriques des propriétés d’échantillonnage répété pour les modèles BVAR et il n’existe pas d’enquêtes fiduciaires généralisées sur les modèles VAR non plus. À cet égard, en utilisant l’approche des sous-ensembles \( \epsilon \)-admissibles (EAS), nous construisons la méthodologie pour une inférence fondée sur une distribution fiduciaire généralisée des probabilités relatives du modèle sur tous les ensembles de composantes actives/inactives (graphiques) de la matrice de transition VAR. Nous établissons la convergence de la sélection graphique par paires et forte pour l’approche EAS dans le cadre de modèles VAR(1) stables, et démontrons de façon empirique qu’il s’agit d’une stratégie efficace dans des contextes de grande dimension.

Additional Supporting Information may be found in the online version of this article at the publisher’s website.

* Corresponding author: jwilli27@ncsu.edu

© 2022 The Authors. The Canadian Journal of Statistics/La revue canadienne de statistique published by Wiley Periodicals LLC on behalf of Statistical Society of Canada / Société statistique du Canada. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
1. INTRODUCTION

Despite the lack of theoretical investigations of the repeated-sampling properties for Bayesian vector autoregression (BVAR) models, Bayesian methodology can surely offer important contributions to the high-dimensional vector autoregression (VAR) model literature, beyond what could be developed in a frequentist framework. One notable such contribution is the construction of posterior distributions over the set of all relative model probabilities. This framework of posterior inference has been widely exploited over the last decade in the high-dimensional linear regression literature, and we anticipate it will see comparable success for high-dimensional VAR models in the near future. Moreover, the generalized fiducial inference (GFI) approach (Hannig et al., 2016) provides these same advantages of the Bayesian approach but without requiring the specification of prior distributions.

Our constructed \( \epsilon \)-admissible subsets (EAS) methodology allows for GF inference of relative model probabilities for all graphs, and additionally we provide an algorithm that is self-tuning (i.e., no cross-validation is needed for calibration to data sets). Such model selection approaches are very useful for learning important relationships among the various components (univariate time series) in the VAR model. The EAS methodology is an entirely new perspective on model selection that was originally developed to effectively account for linear dependencies among subsets of covariates in the high-dimensional linear regression setting in Williams & Hannig (2019).

In the present article, we address the open question of whether the EAS procedure can be extended to the high-dimensional VAR model setting. To the best of our knowledge, our established pairwise and strong model selection consistency results are the first of their kind in the VAR model literature. This type of result is sure to be followed by similar results in the high-dimensional BVAR literature, analogous to the emergence of strong model selection consistency results in the high-dimensional Bayesian/GFI linear regression literature such as Johnson & Rossell (2012), Narisetty & He (2014), and Williams & Hannig (2019). Additionally, we validate our methods empirically in low and high-dimensional settings on both synthetic and real data, and provide Python code for implementing our algorithm. The website https://jonathanpw.github.io/research contains the code/workflow for reproducing all numerical results.

Fiducial inference has a long history, but in the last decade, there has been a renewed interest in the topic with a large number of authors contributing fundamental insights (Berger, Bernardo & Sun, 2009; Edlefsen, Liu & Dempster, 2009; Taraldsen & Lindqvist, 2013; Xie & Singh, 2013; Martin & Liu, 2015; Veronese & Melilli, 2015; Schweder & Hjort, 2016; Fraser, 2019; Jacob et al., 2021). A gentle introduction to technical aspects of GFI is provided in Section 2.

Recent theoretical work on VAR models is largely comprised of considerations of regularized estimation procedures, most notably Basu & Michailidis (2015). The Bayesian literature has not yet caught up. There do exist numerous papers on BVAR methodology, especially in the econometric literature, but on predominantly empirical investigations; see, for example, Bańbura, Giannone & Reichlin (2010), Korobilis (2013), Giannone, Lenza & Primiceri (2015), and Ahelegbey, Billio & Casarin (2016). The primary tool of the BVAR literature has been implementations of the Minnesota (shrinkage) prior and its variants (Litterman, 1986). It has been found that BVAR with shrinkage priors is effective for large VAR models of economic time series, but little has been provided in the way of theoretical guarantees (a notable exception is Ghosh, Khare & Michailidis, 2019) or even uncertainty quantification of competing model choices (notable exceptions are George, Sun & Ni (2008) and Korobilis (2013)). To the best of our knowledge, Ghosh, Khare & Michailidis (2019) are the first in the literature to establish posterior parameter estimation consistency in the “large p, large n” BVAR setting with \( p = o(n) \), where \( p \) is the dimension of the VAR model and \( n \) is the number of observed time
instances. While their consistency results are about the posterior behaviour of the transition matrix coefficients under various prior specifications, our consistency results are about the GF behaviour of all relative model probabilities (akin to Bayes factors) under the prior-free GFI framework.

For completeness, we also provide an overview of the literature on high-dimensional considerations of VAR models from the frequentist perspective since Basu & Michailidis (2015). Guo, Wang & Yao (2016) established convergence rates of \((p/n)^{1/2}\) and \((\log p/n)^{1/2}\) for least squares estimators of a banded coefficient matrix VAR model with respect to two matrix norms. Moment bounds on the VAR model innovations are required for the case when the number of time series in the VAR model, \(p\), is on the order of some polynomial of the number of observed time instances, \(n\), and this is referred to as the “high-dimensional” scenario, whereas exponential moment bounds on the innovations must be assumed for \(p\) not to exceed some subexponential rate of \(n\), termed the “ultrahigh-dimensional” setting. The banded coefficient matrix assumption is a sparsity assumption that the number of unknown coefficient parameters is \(O(p)\) rather than \(p^2\), and it is assumed that the bandwidth is fixed to establish bandwidth selection consistency (unless further assumptions are imposed). Giurcanu (2017) demonstrates oracle consistency properties for M-estimators in multivariate AR time series based on various block bootstrap designs. Lin & Michailidis (2017) provide consistency theorems for maximum likelihood estimators for multi-block VAR models with Gaussian innovations, largely building on the results in Basu & Michailidis (2015). However, their estimation consistency for multi-block VAR models requires \(n > p\) (in our notation). Hall, Raskutti & Willett (2019) extend the analysis of high-dimensional VAR models to the generalized linear VAR setting, and in particular for count-valued data. Assuming sparsity, complexity bounds are provided for consistency of regularized maximum likelihood estimators in this non-Gaussian setting. From the high-dimensional spectral analysis perspective, theoretical guarantees for sparse inverse periodogram estimation are investigated in Fiecas et al. (2019). Wong, Li & Tewari (2020) demonstrate consistency in the high-dimensional setting for the coefficient matrix in a LASSO objective function for a strictly stationary multivariate time series consisting of sub-Weibull random vectors.

Within the frequentist literature since Basu & Michailidis (2015), attention has been focusing on studying model selection and estimation for VAR models with non-Gaussian innovations; see, e.g., Guo, Wang & Yao (2016), and in the univariate setting, Medeiros & Mendes (2016). Nonetheless, in the recent work of Basu, Li & Michailidis (2019), high-dimensional estimation consistency is shown for regularized estimators of VAR(1) model parameters with alternative low-rank and structured sparsity assumptions, and Gaussian innovations are assumed to guarantee that their sufficient conditions for consistency are satisfied with high probability. See also Raskutti, Yuan & Chen (2019) for derived theory for high-dimensional VAR models with Gaussian innovations and low-dimensional structural assumptions. From a different angle, Safikhani & Shojaie (2022) establish high-dimensional estimation consistency for VAR models that are only piecewise stationary, but again the result relies on an assumption of Gaussian innovations. Nicholson et al. (2020) provide a concentration bound on the in-sample prediction error for a high-dimensional VAR model, with a Gaussian innovation assumption. Notably, they afford special emphasis to the VAR lag selection.

Moreover, within the scope of Bayesian theory, VAR models with Gaussian innovations are still under development; see, e.g., Ghosh, Khare & Michailidis (2019) and Ghosh, Khare & Michailidis (2021). In future work, we hope to extend our study to non-Gaussian innovations, but that will require careful consideration of a misspecified data-generating equation that will accommodate the type of non-Gaussianity considered. GFI is an alternative paradigm for statistical inference with well-established principles, and our work is the first study of GFI in the multivariate time series setting, so it is natural that we consider Gaussian innovations. Moreover, establishing theoretical guarantees within the paradigm of GFI is a more ambitious endeavour.
than in frequentist and Bayesian approaches. Theoretical guarantees within the frequentist paradigm are obtained by starting with some conveniently posed optimization problem suited to facilitate well-established proof techniques (such as for regularized regression objective functions applied in a plethora of statistical models for the past two decades). Theoretical guarantees within the Bayesian paradigm are arrived at by constructing contrived prior densities with the strategy of mimicking the frequentist-based proof techniques, a practice that technically violates fundamental Bayesian principles. On the contrary, GFI is a statistically principled approach to inference with appeals to Bernstein–von Mises theory, and we do not relax any of its principles to develop our theoretical results. This is particularly meaningful as a contribution to the literature for a field such as statistics that lacks unifying foundational principles, and every new GFI investigation provides more clarity about the fundamental mechanics of GFI.

We loosely adopt notation for multivariate time series from Lütkepohl (2005). The time series \( X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p \) is taken to denote data from a VAR(1) model with no serial correlation, and so is generated as

\[
Y = AX + \Sigma^\frac{1}{2}U, \tag{1}
\]

where \( Y := (X^{(1)} \cdots X^{(n)}) \) and \( X := (X^{(0)} \cdots X^{(n-1)}) \) are \( p \times n \) matrices, \( U := (U^{(1)} \cdots U^{(n)}) \) is a \( p \times n \) matrix with \( U^{(t)} \overset{iid}{\sim} N_p(0, I_p) \) for \( t \in \{1, \ldots, n\} \), \( A \) is a \( p \times p \) matrix of coefficients, and \( \Sigma := \text{diag}(\sigma_1^2, \ldots, \sigma_p^2) \). Note that, based on Remark 1.1, it is without loss of generality that we assume \( \Sigma \) is a diagonal matrix. Assume \( X^{(0)} \) is the \( p \)-dimensional zero vector. Further, let \( G \subseteq \{1, \ldots, p^2\} \) be a set of indices denoting a graph of active (i.e., nonzero) components of \( \text{vec}(A) \), where the \( \text{vec}() \) operator transforms an \( n \times p \) matrix into an \( np \times 1 \) vector by stacking columns in descending order, from left to right. Lastly, take \( A_g \) to be the \( p \times p \) matrix \( A \) with active components corresponding to the graph \( G \) (all other components are zero).

**Remark 1.1.** Suppose that \( Y \) is an arbitrary positive-definite covariance matrix, and \( \tilde{A} \) is a \( p \times p \) coefficient matrix. By the spectral theorem, there exists an orthogonal matrix \( Q \) such that \( Y = Q\Sigma Q' \) for some positive-definite diagonal matrix \( \Sigma \). That being so, pre-multiply the data-generating equation \( \tilde{Y} = \tilde{A}\tilde{X} + Y^\frac{1}{2}\tilde{U} \) by \( Q' \) to obtain

\[
Y := Q'\tilde{Y} = Q'\tilde{A}\tilde{X} + Q'Y^\frac{1}{2}U = Q'\tilde{A}QQ'\tilde{X} + Q'Q\Sigma^\frac{1}{2}Q'U = AX + \Sigma^\frac{1}{2}U,
\]

where \( A := Q'\tilde{A}Q, X := Q'\tilde{X}, \) and \( U := Q'\tilde{U} \). Assuming \( \tilde{U} \) is a \( p \times n \) matrix with iid \( N_p(0, I_p) \) distributed columns, it follows that \( Q'\tilde{U} \) also has iid \( N_p(0, I_p) \) distributed columns. In summary, for any error covariance matrix, there will always exist a change-of-basis transformation of the data such that the error covariance is diagonal or nondiagonal, and so the diagonal representation is most meaningful (or least arbitrary) for specifying the data-generating equation that defines the VAR(1) transition matrix, \( A \).

We extend the high-dimensional linear regression EAS methodology developed in Williams & Hannig (2019) to this VAR(1) setting. This extension, particularly in the context of GFI, is characterized by a variety of additional difficulties. Namely, the mathematical details of the GF distribution for a VAR model are nontrivial and have never been worked out before. In particular, in a linear regression model, uncorrelated predictors can be omitted from the model with no effect on the data-generating system. However, if components in a VAR model are omitted, then the data-generating system is altered. These details make our article substantially different from Williams & Hannig (2019).

The idea behind the EAS procedure is to efficiently make inference on the set of \( 2^p \) graphs, \( G \), by discriminating on graphs that contain redundant active components. Our notion of redundancy
is defined rigorously by the “$h$-function” given later in (4). However, the basic intuition is to assign negligible GF probability to all $A_g$ that can be closely approximated, predictively, by a graph containing fewer active components. This can occur for a variety of reasons, such as correlated time series in the VAR system of equations, and too small signal-to-noise coefficient magnitudes.

The remainder of the article is organized as follows. Section 2 defines the notion of $\varepsilon$-admissibility, constructs the GF distribution for the EAS approach, and describes the Markov chain Monte Carlo (MCMC)-based computations. The main theoretical results are presented in Section 3, and numerical results are provided in Sections 4 and 5. The majority of the proofs are moved to the Supplementary Material.

2. METHODOLOGY

To adapt ideas more smoothly from the linear regression setting of Williams & Hannig (2019), re-express the VAR(1) model in (1) in the form

$$Y = Z_{G_o} \alpha_{G_o}^0 + \left(W^0\right)^{1/2} \text{vec}(U),$$

where $Y := \text{vec}(Y)$, $Z := \lambda' \otimes I_p$, $W^0 := I_n \otimes \Sigma^0$, $\alpha := \text{vec}(A)$, and $G_o$ (as well as $g_o$ seen later) denotes the oracle graph. Here and throughout, the superscript-zero notation denotes the true fixed values of the corresponding quantities. The subscript notation, $Z_{G_o}$ (or $\alpha_{G_o}$), refers to the submatrix (or subvector) with columns (or components) corresponding to the active components given by the index set $G_o$. This linear model representation is also more convenient for expressing the likelihood function,

$$f \left(Y | \alpha_{G_o}, \{\sigma_j\}\right) = \frac{1}{(2\pi)^{np/2} \prod_{j=1}^{p} \sigma_j^2} e^{\frac{1}{2} \left(Y - Z_{G_o} \alpha_{G_o}\right)' \left(Y - Z_{G_o} \alpha_{G_o}\right)},$$

which will be needed later on. For conciseness, the notation $\{\sigma_j\}$ is used as shorthand for $\{\sigma_1, \ldots, \sigma_p\}$.

Additional notation used for the remainder of the article includes the following. For a scalar-valued argument $|\cdot|$ represents the absolute value, but for a set-valued argument it represents the cardinality. The norms $\|\cdot\|$ and $\|\cdot\|_0$ denote the vector $L_2$ and $L_0$ norms, respectively, while for a matrix $A$, $\|A\|_2 := \sqrt{\lambda_{\max}(A'A)}$ and $\|A\|_F := \sqrt{\text{tr}(A'A)}$ represent the matrix spectral and Frobenius norms, respectively. Additionally, the quantities $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a given matrix, $A$, respectively. The notations $P(\cdot)$ and $E(\cdot)$ refer, respectively, to the probability measure and expectation with respect to the joint GF distribution of $A_g$ and $\Sigma$. Conversely, the notations $P_x(\cdot)$ and $E_x(\cdot)$ refer, respectively, to the probability measure and expectation associated with the uncertainty from the VAR(1) process, rather than the probability measure for the GF distribution of the unknown parameters.

The centrepiece of the EAS model selection approach is a definition of model redundancy, as made rigorous by our notion of $\varepsilon$-admissibility and the $h$-function, presented next. As described in Section 1, the main intuition is that $\alpha_G$ is considered nonredundant, or $\varepsilon$-admissible, if and only if there does not exist a close fitting graph with strictly fewer active components. However, there are also two additional constraints embedded in the $h$-function for $\varepsilon$-admissibility.

**Definition 2.1.** Assume $\varepsilon, d > 0$ and $c \in (0, 1)$. A given coefficient matrix $A_g$, equivalently $\alpha_G$, for some graph $G$ is said to be $\varepsilon$-admissible if and only if $h(\alpha_G, \{\sigma_j\}) = 1$, where

$$h(\alpha_G, \{\sigma_j\}) := \{\frac{1}{2} \|Z^t_G Z_G^{-1} Z_G (\alpha_G - b_{\min})\|^2 \geq \varepsilon, \min_{1 \leq j \leq p} \{m_j^G\} \geq d, \|A_g\|_2 \leq c\}.$$
in which $b_{\text{min}}$ solves $\min_{b\in\mathbb{R}^{|G|}} \frac{1}{2}||Z_G'W^{-1}Z_G(b - \alpha_G)||^2$ subject to $||b||_0 \leq |G| - 1$,

$$\{m_1^g, \ldots, m_p^g\} = \text{diag}\left\{ \left( Y - \hat{A}_g x\right) \left( Y - \hat{A}_g x\right)' \right\},$$

and $\hat{A}_g := YZ_G'(Z_GZ_G')^{-1}$ is the least squares estimator for graph $G$.

To begin to understand the behaviour of the $h$-function, first note that

$$||Z_G'W^{-1}Z_G(\alpha_G - b_{\text{min}})||^2 = ||Z_G'W^{-1}(Z_G\alpha_G - Z_Gb_{\text{min}})||^2,$$

is analogous to a noiseless version of the Dantzig selector (Candes & Tao, 2007) where $Z_G$ is the design matrix for the linear model representation (2). One reason to use $Z_G'W^{-1}Z_G$ versus simply $Z_G$ is that the former is scale-invariant to the $\{\sigma_j\}$ and invariant to orthogonal transformations of the data. Second, note that if $Z_G$ contains linearly dependent columns, then for any coefficients $\alpha_G$, the linear prediction $Z_G\alpha_G$ can be exactly recovered by $Z_Gb_{\text{min}}$ (since $||b_{\text{min}}||_0 \leq |G| - 1$). This immediately implies that since $Z_G$ is an $np \times |G|$ matrix, for all $G$ with $|G| > np$, $h(\alpha_G, \{\sigma_j\}) = 0$ by definition. Then by construction, for high-dimensional settings where $p > n$, considering only $\text{admissible}$ graphs reduces the model selection problem from $2^p$ candidate graphs to only $2^{np}$. This fact makes the EAS methodology inherently scalable.

The quantities $c$, $d$, and $\varepsilon$ will now be described in alphabetical order. The component in the $h$-function, $||A_g||_2 \leq c$, concentrates the distribution of $A_g$ to only allow for stable VAR(1) models with $c \in (0, 1)$ (Negahban & Wainwright, 2011; Loh & Wainwright, 2012; Han, Lu & Liu, 2015). In practice, since $||A^0||_2$ is typically not known, the constraint $||A_g||_2 \leq c$ is replaced by $||A_g||_2 < 1$. The second component in the $h$-function is the expression $\min_{1 \leq j \leq p} \{m_j^g\} \geq d$, where $m_j^g$ for $j \in \{1, \ldots, p\}$ is understood as the residual sum-of-squares (RSS) for the $j$th component of the VAR system. The basic idea is that the data-dependent quantity $d = d(Y, x', G_g)$ should be calibrated to $\min_{1 \leq j \leq p} \{m_j^g\}$, which corresponds to the oracle graph, and so any graphs that have a better fit than the oracle will be excluded from consideration via the $h$-function. Accordingly, this device is designed to eliminate graphs that over-fit the data, and is important for establishing our asymptotic consistency results. However, in practice $d$ can be set to a small value and left alone; more will be said about this in Section 4 with the numerical results.

For $Z_G$, which have full column rank, the degree to which the features associated with graph $G$ are redundant depends on the correlations between the $p$ components of the VAR model, the distribution of the coefficients $\alpha_G$ (i.e., the transition matrix $A_g$), scale matrix components $\{\sigma_j\}$, and the specified level of precision, $\varepsilon$. Our proposed default choice of $\varepsilon$, formulated from theoretical investigations (based on the Gaussian contemporaneous errors assumption), is for some $\rho \in (0, \frac{1}{2})$,

$$\varepsilon = \Lambda_g \cdot \max \left\{ 1, n^{1-\rho}p^2 \left( .5 \log(\log(n))|G| - |G_o| \right) \right\}.$$  

(6)

There are predominantly two components to $\varepsilon$; the quantity $\Lambda_g := ||W^{-\frac{1}{2}}Z_G||_F$ is particularly calibrated to the observed data since it originates from a tight concentration inequality for the transition matrix $A_g$, and the term $n^{1-\rho}p^2\log(\log(n))|G|$ is necessary asymptotically for managing the accumulating data and rapidly growing number of candidate graphs as $n, p \to \infty$. The basic idea is that $\Lambda_g$ will always contribute, and the remaining terms will contribute for sufficiently large $n$ or for $|G|$ that exceeds the number of active components in the oracle model. However, as is demonstrated in Section 4, for observed data $\Lambda_g$ is so well calibrated that it suffices to set $\varepsilon = \Lambda_g$, and thus also eliminating the need for a tuning parameter. More details about $\Lambda_g$ are given in Section 3.2, particularly its expectation in (12).
With the EAS methodology now developed, a framework of statistical inference is required for implementing it. GFI is a suitable such framework because it will allow us to construct GF inference over the 2^2 candidate graphs without having to specify any prior distributions. The intuition for GFI is to begin with a data-generating equation such as (2) and invert the equation on the data to solve for the unknown parameters. The resulting quantity is defined as the GF distribution of the unknown parameters. An introduction to the details for GFI is described next.

In general, let \( \Theta \) represent the support of the collection of unknown parameter values to learn from a data-generating equation (e.g., Eq. 2). Suppose \( y_1, \ldots, y_n \) is an observed collection of iid random variables, \( V \) is the deterministic data-generating function, and \( U_1, \ldots, U_n \) are independent auxiliary random variables with a common known distribution. The motivating idea for GFI is to construct a distribution of parameter values that are most consistent with the observed data arising from the data-generation equation \( V \), given that the auxiliary variables \( U_1, \ldots, U_n \) admit parameter values such that \( \sum_{i=1}^n ||y_i - V(U_i, \theta)||^2 \) is “small”. More precisely, define, as a pseudo-inverse of the data-generating equation, \( Q_y(u_1, \ldots, u_n) := \arg\min_{\theta \in \Theta} \sum_{i=1}^n ||y_i - V(u_i, \theta)||^2 \), and for any \( \epsilon > 0 \) let \( \theta_{\epsilon} := Q_y(U_1^*, \ldots, U_n^*) \), where \( U_1^*, \ldots, U_n^* \) are iid with the same distribution as \( U_1, \ldots, U_n \) but restricted to the set

\[
\left\{ U_1^*, \ldots, U_n^* : \min_{\theta \in \Theta} \sum_{i=1}^n ||y_i - V(U_i^*, \theta)||^2 \leq \epsilon \right\}.
\]

Then, as defined in Hannig et al. (2016), a GF distribution on \( \Theta \) is the weak limit of \( \theta_{\epsilon} \) as \( \epsilon \to 0 \). Taking \( \epsilon = 0 \) is reminiscent of earlier attempts made in the mid-twentieth century at constructing fiducial distributions, and leads to conditioning on zero-measure sets for continuous data. To rectify these complications, GFI instead considers the limit as \( \epsilon > 0 \) is taken to zero.

Sampling from the GF distribution directly from this definition is similar in spirit to approximate Bayesian computations (see, e.g., Sisson, Fan & Beaumont, 2018). However, it is proven in Hannig et al. (2016) that for continuous data, and under conditions mostly relating to the smoothness of \( V \), the GF distribution as defined above can be described by the probability density function,

\[
r(\theta|y_1, \ldots, y_n) := \frac{f(y_1, \ldots, y_n; \theta) J(y_1, \ldots, y_n, \theta)}{\int_\Theta f(y_1, \ldots, y_n; \theta) J(y_1, \ldots, y_n, \theta) \, d\theta},
\]

where \( f(y_1, \ldots, y_n; \theta) \) is the likelihood of the data, and the Jacobian term \( J(y_1, \ldots, y_n, \theta) := \sqrt{\det(C'C)} \), with

\[
C := \nabla_\theta V(u^n, \theta)|_{u^n = V^{-1}(y_1, \ldots, y_n, \theta)}.
\]

The Jacobian term results from inverting the data-generating equation on the unknown parameters. Full details for GFI are provided in Hannig et al. (2016).

In the case of the VAR model (1) with data-generating Equation (2),

\[
r(\alpha, \sigma|Y) = \frac{f(Y|\alpha, \sigma) \cdot J(Y, (\alpha, \sigma)) \cdot h(\alpha, \sigma)}{\int \int f(Y|\alpha, \sigma) \cdot J(Y, (\alpha, \sigma)) \cdot h(\alpha, \sigma) \, d\alpha \, d\sigma},
\]

where the multiplication by the \( h \)-function appears as an infusion of the EAS methodology into the GFI framework, and the Jacobian term

\[
J(Y, (\alpha, \sigma)) := D\left(\nabla_{(\alpha, \sigma)} V(u, (\alpha, \sigma)) \bigg|_{u = V^{-1}(Y, (\alpha, \sigma))}\right).
\]
Note that the \( \{ \sigma_j \} \) are also dependent on the particular graph \( G \), but this dependence is suppressed in the notation for conciseness. The likelihood function in (7) is given by (3), the \( h \)-function is given by (4), and the derivation of the Jacobian term is presented in the Supplementary Material. From the GF density of \( \alpha_G \) and \( \{ \sigma_j \} \), the GF mass function for a graph \( G \) is proportional to the normalizing constant in (7). In Bayesian theory, this constant of proportionality is understood as the marginal density of the data. Evaluating the integral in the denominator of (7) gives

\[
E \left( h(\alpha_G, \{ \sigma_j \}) | D_g \right) \prod_{j=1}^{\mathbf{1}} \frac{m_j^2}{2} \left( -\frac{n-r_j^\delta}{2} \right) \Gamma \left( \frac{n-r_j^\delta}{2} \right),
\]

(8)

where \( r_j^\delta \) is the set of active row indices of \( A_g \) for column \( j \in \{ 1, \ldots, p \} \), and \( D_g \) is a data-dependent and parameter-free \((np) \times (|G| + p)\) matrix defined in the Supplementary Material as part of the Jacobian term. Note that the inner expectation is with respect to the \( N_{|G|} \left( \hat{\alpha}_g, (Z_G'W^{-1}Z_G)^{-1} \right) \) distribution, conditional on \( \{ \sigma_j^\delta \} \), and for each \( \sigma_j^\delta \), is taken with respect to the inv-gamma \( \left( \frac{1}{2} (n - |r_j^\delta|), \frac{1}{2} m_j^\delta \right) \) distribution. To ensure that \( r(G|Y) \) defines a proper probability mass function, the normalizing constant in (8) is scaled so that \( \sum_{|G|=\alpha} r(G|Y) = 1 \).

Lastly, the relative model probabilities given by (8) can be computed via pseudo-marginal MCMC algorithms. Traditional MCMC is not feasible because the expected value appearing in (8) is not available in closed form. We implement the grouped independence Metropolis–Hastings (GIMH) algorithm described in Andrieu & Roberts (2009), which replaces the expected value with the empirical mean of importance samples at each step of the MCMC algorithm. In the case of (8), efficient importance samples are easily drawn from the \( N_{|G|} \left( \hat{\alpha}_g, (Z_G'W^{-1}Z_G)^{-1} \right) \) and inv-gamma \( \left( \frac{1}{2} (n - |r_j^\delta|), \frac{1}{2} m_j^\delta \right) \) distributions for \( \alpha_G \) and \( \sigma_j \), respectively. The GIMH algorithm we construct is a Markov chain on the set of graphs \( G \subseteq \{ 1, \ldots, p^2 \} \), and proposals are made by either adding, removing, or replacing a component index in the current iteration of \( G \) in the chain.

A point of caution about the GIMH algorithm is that the mixing conditions are usually particularly sensitive to the number of importance samples taken to estimate an expectation at each step of the algorithm. However, the algorithm mixed well enough to yield very encouraging numerical results for the high-dimensional linear regression setting in Williams & Hannig (2019) and in Sections 4 and 5 of the present article, which serve to demonstrate that the algorithm is not only computationally feasible but also favourable for graph selection in the VAR(1) model setting. Further discussion of the algorithm is provided in Williams & Hannig (2019), and a detailed pseudo-code description of the algorithm is provided at https://jonathanpw.github.io/research.

2.1. Toy Example

In finite samples, and particularly in high-dimensional settings with highly correlated data, the EAS framework has the intuition that the oracle graph itself may not be \( \epsilon \)-admissible. In these settings, the EAS methodology re-defines the notion of the “true” graph to be some nonredundant subgraph of the oracle graph, at least nonasymptotically. This idea is important because it suggests that to develop inherently scalable methodology, the key may be to re-define the notion of what one should hope to recover from a “true” data-generating model in high-dimensional settings. Next, we provide a short example to illustrate these ideas.

Suppose \( p = 2 \) and the true graph of active components is \( G_o = \{ 1, 2, 3, 4 \} \) with \( A_{g_o} = \begin{pmatrix} 0.6 & 0.05 \\ 0.55 & 0.3 \end{pmatrix} \). The true coefficient matrix \( A_{g_o} \) is not \( \epsilon \)-admissible if, for instance, for some
well-calibrated precision, $\varepsilon > 0$,

$$\left\|\begin{pmatrix} 0.6 & 0.05 \\ 0.55 & 0.3 \end{pmatrix}X - \begin{pmatrix} 0.6 & 0 \\ 0.55 & 0.3 \end{pmatrix}X'\right\| < \varepsilon,$$

where $\| \cdot \|$ is some measure of distance. In this case, predictions from the graph $\{1, 2, 4\}$ approximate that of $A_{\mathcal{G}_o}$ within $\varepsilon$ precision, and so $A_{\mathcal{G}_o}$ is said to contain redundant information.

To illustrate the methodology we develop in this article, using the notion of redundancy that we will define and study, we estimate the distribution over graphs $G \subset \{1, 2, 3, 4\}$ for this simple example using this specific $A_{\mathcal{G}_o}$ to generate a time series from model (1) with $n = 70$ and $\Sigma = I_2$. See Table 1 and Figure 1. The computations involved are described in the sections that follow. The basic idea is that while the true graph $G_o$ includes all four components as active, the graph $\{1, 2, 4\}$ is less complex yet predicts future out-of-sample instances of the time series with virtually the same level of accuracy (observe the prediction errors in Table 1). Additional intuition for the EAS methodology is provided in Williams & Hannig (2019) in the context of linear regression.

The GFI construction that we employ allows for parametric inference analogous to Bayesian inference. With respect to uncertainty quantification of the unknown parameters, conditional on a given model/graph $G$ (i.e., set of active coefficients), the standard deviation of the estimated GF (i.e., posterior) distribution for each coefficient would serve a similar role as a standard error, if the posterior mean is taken as the point estimate. Instead of using $P$-values to determine statistical significance, we use the marginal inclusion probability of a particular coefficient over all models/graphs in the GF distribution of $G$ to determine the significance of the coefficient. This measure of significance is relevant to the familiar notion of frequentist statistical significance in so far as our main theoretical finding says that the GF distribution of models/graphs $G$ will converge to the true distribution, asymptotically. Moreover, this notion is akin to repeated-sampling Bayesian investigations where it is shown that certain credible sets achieve their nominal frequentist coverage level of significance.

**Table 1:** Estimated GF mass function over graphs $G \subset \{1, 2, 3, 4\}$.

| $G$          | $\{1, 2, 4\}$ | $\{1, 4\}$ | $\{1, 2, 3, 4\}$ | $\{1, 2\}$ | $\{1, 3, 4\}$ |
|-------------|----------------|------------|------------------|------------|----------------|
| $\hat{\pi}(G|\mathcal{Y})$ | 0.877         | 0.091      | 0.017            | 0.013      | 0.002          |
| Prediction error | 1.2384        | 1.9197     | 1.235            | 1.5586     | 1.9089         |

**Note:** Estimated graph probabilities are given by $\hat{\pi}(G|\mathcal{Y})$, and graphs not shown had $\hat{\pi}(G|\mathcal{Y}) = 0$. On an out-of-sample test set of 70 time instances, prediction error is defined as $\frac{1}{n} \| Y - \hat{A}_x X \|_2$, where $\hat{A}_x$ represents the estimated least squares transition matrix on the first $n = 70$ in-sample time instances.

**Figure 1:** Directed graph of inclusion probabilities of components of the transition matrix, $A$. Each edge label represents the marginal GF inclusion probability of a particular component of $A$. That is, the proportion of graphs (over all MCMC-sampled graphs) in which each component (i.e., edge) of $A$ is active. Line widths are proportional to inclusion probabilities.
3. THEORETICAL RESULTS

The problem of graphical selection is difficult because the number of candidate graphs to choose among grows super-exponentially in the dimension of the VAR(1) model, $2^{p^2}$. Accordingly, the utility of the EAS procedure is its inherent ability to effectively manage a very large number of candidate graphs by assigning negligible GF probability to redundant graphs. The meaning of this assertion is made precise in Theorem 3.12, which states that the GF distribution obtained from the EAS methodology exhibits pairwise graph selection consistency as both $n$ and $p$ are taken to infinity, and as a corollary, strong selection consistency for fixed $p$. In what follows, the notation $\lim_{p \to \infty}$ is used to denote convergence with respect to the data-generating mechanism.

There are three major added difficulties in the extension from the high-dimensional univariate linear regression setting. First, in autoregressive models, data appear on both sides of the data-generating equation, and so probabilistic guarantees must be established for conditions that otherwise would only involve deterministic quantities (e.g., bounds on eigenvalues of the design matrix). Probabilistic guarantees for a “restricted eigenvalue condition” and a “deviation condition”, different but analogous to some of our results, are provided in Basu & Michailidis (2015) in the frequentist setting. Second, added complexity arises simply by extending from a univariate-response setting to a vector-response setting. For instance, concentration bounds on ratios of residual sums of squares for different models in the univariate setting become concentration bounds on ratios of determinants (or eigenvalues) of error covariance matrices for different models in the multivariate setting, for which much less standard theory exists. The third added difficulty, specific to GFI, is that the Jacobian term for the VAR(1) model (derived in Section S1 of the Supplementary Material) is mathematically complicated and exerts substantial influence on the resulting model probabilities in Equation (8), whereas the Jacobian term in the univariate linear regression setting (Williams & Hannig, 2019) is concise and involves components readily relatable to the likelihood function.

3.1. Conditions

The first two conditions presented are related to the identifiability of the true data-generating graph, $G_0$. The expected ramifications of these conditions are verified empirically on data in Sections 4 and 5. Since this article focuses on developing (i) novel methodology for graphical selection in the context of VAR models, (ii) the theory for GFI in the context of VAR models, and (iii) asymptotic theory related to BVAR models, we consider only a stable VAR(1) model for simplicity. We understand that, out of context, restriction to the stable VAR(1) setting is limiting. However, with respect to the three theoretical milestones we address, it would be overly ambitious to consider a more general setting in this first investigation. We adopt our notion of stability from Negahban & Wainwright (2011), Loh & Wainwright (2012), and Han, Lu & Liu (2015) in that the true transition matrix satisfies $\|A_0\|_2 \leq c$ for some $c \in (0, 1)$. It is assumed throughout that a valid $c$ has been fixed a priori. However, we expect the EAS procedure to perform meaningful graph selection even if the stability condition is violated. The whole idea about the EAS approach is to find a meaningful model that is nonredundant as defined by our $h$-function. So if $h$ includes a stability constraint, then we are finding the “best fitting” distribution over models that are nonredundant. We are intentionally dismissing the notion that we hope to find the “true model” in finite data sets.

Condition 3.1 arises in the proof of Lemma S2.4, which is a necessary result for Theorem 3.10. It guarantees that the Jacobian term for the oracle graph in (8) will be lower-bounded away from zero in probability. The quantity $\delta$ represents an approximation to $\lambda_{\min}(\Omega - E_x(\Omega))$ (via Lemma A.1) that manages the uncertainty resulting from the minimum eigenvalue of the Jacobian.
matrix $\tilde{D}_\omega \tilde{D}_{k\omega}$, where $\Omega := \frac{1}{\nu^2} \begin{pmatrix} \mathcal{X} \mathcal{X}' & \mathcal{X} \mathcal{U}' \\ \mathcal{U} \mathcal{X}' & \mathcal{U} \mathcal{U}' \end{pmatrix}$ and $E_x(\Omega) = \begin{pmatrix} \Gamma_n(0) \\ I_p \end{pmatrix}$. It is also assumed that a valid $\delta > 0$ has been fixed a priori.

**Condition 3.1.** The true transition matrix satisfies $\|A^0\|_2 \leq c < 1$, $\lambda_{\max}(\Gamma_n(0))$ is bounded from above by a fixed constant, and

$$\sqrt{n} \left[ \lambda_{\min} \left( \begin{pmatrix} \Gamma_n(0) \\ I_p \end{pmatrix} \right) - \delta \right] > 4(1 + c^2),$$

where $\delta > 0$, and

$$\Gamma_n(0) := \frac{1}{n} E_x(\mathcal{X} \mathcal{X}') = \frac{1}{n^2} \sum_{t=1}^{n} \sum_{k=0}^{t-2} (A^0)^k \Sigma^0 (A^0)^k'.$$

Observe that this condition also implies that $\lambda_{\min}(\Gamma_n(0)) > \delta$.

Note that Lemma A.1 guarantees $\lambda_{\min}(\Omega - E_x(\Omega)) \xrightarrow{P_x} 0$ as $n \to \infty$, assuming the $p$ versus $n$ relationship given by Condition 3.4. Thus, the condition can reasonably be verified on real data by assuming $\delta > 0$ is arbitrarily small and comparing the value of $\sqrt{n} \lambda_{\min} \left( \begin{pmatrix} \frac{1}{n} \mathcal{X} \mathcal{X}' \\ I_p \end{pmatrix} \right)$ to

$$4(1 + c^2),$$

where $\frac{1}{n} \mathcal{X} \mathcal{X}'$ is the obvious sample analogue to the population quantity considered in Condition 3.1. Since $\nu$ is unknown in practice, for the purposes of checking this condition on real data, evaluate $4(1 + c^2) = 8$ for the worst case with $c$ replaced by 1. We demonstrate on synthetic data in Section 4 that this verifiable condition is indeed meaningful for practical applications.

Condition 3.2, which originates from the proof of Theorem 3.10, is also well calibrated to real data. This condition states the maximum rate at which $\epsilon$ can be allowed to grow as a function of $n, p$, and $\Lambda_{k\omega}$, while the oracle model $G_o$, remains identifiable (i.e., no faster than $n^{1-p} p^2 \Lambda_{k\omega}$).

The fixed quantity $\rho \in (0, \frac{1}{2})$ represents the “gap” between how fast $\epsilon$ must grow (stated in Condition 3.4) to effectively manage the set of all $2^p$ candidate graphs under consideration, and how slow it must grow to not eliminate the oracle graph from consideration. Specifically, $\epsilon \propto n^{1-p} p^2 \Lambda_{k\omega}$ simultaneously satisfies Conditions 3.2 and 3.4 for any $\rho \in (0, \frac{1}{2})$. It is assumed throughout that a valid $\rho$ has been fixed a priori. The quantities on the left side of the inequality in Condition 3.2 are expected values of the corresponding quantities on the left side of the first constraint in the $h$-function (4).

**Condition 3.2.** The oracle graph, $G_o$, satisfies $\min_{\rho \leq \rho} \{ m_{G_o}^\rho \} \geq d$ and

$$\frac{1}{2} \left\| (\Gamma_n(0) \otimes (\Sigma^0)^{-1})_{G_o,G_o} (\alpha_{G_o}^0 - \tilde{b}) \right\|^2 \geq \frac{9\epsilon}{n^{1-p} p^2 \Lambda_{k\omega}},$$

where $\rho \in (0, \frac{1}{2})$, $\tilde{b}$ solves $\min_{b \in [G_o]} \left\| (\Gamma_n(0) \otimes (\Sigma^0)^{-1})_{G_o,G_o} (\alpha_{G_o}^0 - b) \right\|^2$ subject to $\|b\|_0 \leq |G_o| - 1$, and $\epsilon = \Lambda_{k\omega} \cdot \bar{\epsilon}$ for some $\bar{\epsilon}$ not depending on $\Sigma$ or $A_{k\omega}$.

Unless the oracle model is known, Condition 3.2 is not verifiable on real data, but in Section 4 we are able to demonstrate the varying performance of the EAS procedure on simulated data when this condition is satisfied compared with when it is not.
The next condition is a component in the proof of Theorem 3.9 for guaranteeing that the \( h \)-function will drive the EAS procedure to assign negligible GF probability to non-\( \varepsilon \)-admissible graphs via the mass function \( r(G|Y) \) in (8).

**Condition 3.3.** For any \( G \) with \( G \not\subseteq G_o \),

\[
\frac{1}{2} \left\| (E_x(\tilde{Z}_G^t Z_G))^{-1} E_x(\tilde{Z}_G^t Y) - \tilde{b} \right\|_2^2 < \frac{\varepsilon}{9n^{1+\delta}p^3\Lambda_g},
\]

where \( \tilde{b} \) solves \( \min_{b \in \mathbb{R}^{|G|}} \left\| (E_x(Z_G^t Z_G))^{-1} E_x(Z_G^t Y) - b \right\|_2^2 \) subject to \( ||b||_0 \leq |G| - 1 \), and \( \varepsilon = \Lambda_g \cdot \tilde{\varepsilon} \) for some \( \tilde{\varepsilon} \) not depending on \( \Sigma \) or \( A_g \).

The intuition for Condition 3.3 is that for graphs containing redundant active components, the central tendency of the least squares estimator \( \hat{\alpha}_g \) can be closely approximated by a vector of fewer active components. Notice that \( (E_x(Z_G^t Z_G))^{-1} E_x(Z_G^t Y) \) is an approximation to \( E_x(\hat{\alpha}_g) \). Since the least squares estimator is asymptotically well behaved for Gaussian VAR models, this condition is not particularly interesting and is easily satisfied in numerical experiments. Furthermore, it will hold trivially if, for instance, the columns \( Z_G \) are linearly dependent.

The final condition in this section is Condition 3.4, which simply states the asymptotic rate at which \( \varepsilon \) and \( d \) from the definition of \( h \) in (4) must increase as \( n, p \to \infty \) for our main result, Theorem 3.12, to be established. In fact, the previous three conditions were all for establishing nonasymptotic bounds of concentration.

**Condition 3.4.** For some fixed \( \rho \in (0, \frac{1}{2}) \), \( p^{\max\left\{ \frac{14}{7} - \frac{2}{1 - \rho^2} \right\}} = o(n) \). For the positive constant \( K_1 \) specified in (A2), as \( n \to \infty \) or \( n, p \to \infty \), \( \varepsilon \) satisfies

\[
\frac{\varepsilon}{9\Lambda_g} - K_1 \left( \frac{p\|Y\|^2}{\sqrt{n}} + p^2 \log(n) + \frac{n}{q} \cdot p^2 \sqrt{n} \right) \to_{P_x} \infty,
\]

\( d \) satisfies

\[
\frac{d \cdot n^2 p^2}{4\lambda_{\max}(X'X/n)} - K_1 \left( \frac{p\|Y\|^2}{\sqrt{n}} + p^2 \log(n) + \frac{n}{q} \cdot p^2 \sqrt{n} \right) \to_{P_x} \infty,
\]

and \( n = O_p(q) \), where \( q := \min_{1 \leq j \leq p} \{m_j\} \) with \( m_1, \ldots, m_p \) corresponding to the full model (i.e., all components active), and \( \varepsilon = \Lambda_g \cdot \tilde{\varepsilon} \) for some \( \tilde{\varepsilon} \) not depending on \( \Sigma \) or \( A_g \).

An important attribute of Condition 3.4 is the requirement that while the dimension of the VAR(1) model, \( p \), can be taken to infinity, it must be exceeded polynomially by the number of observed time instances, \( n \). This is in contrast to the model selection consistency result established for the high-dimensional linear regression setting in Williams & Hannig (2019), where \( p \) was allowed to grow subexponentially in \( n \). From our investigation, the primary reason for the difference here is that bounding the GFI Jacobian term proves to be much more difficult in the VAR model setting, but improving our bound remains an avenue of ongoing future work.

This section concludes with a definition of various quantities that will be referenced in the next section, and throughout the proofs.
Definition 3.5. \( N_1 \) is any positive constant such that \( n \geq N_1 \) implies

\[
1 - \frac{1 - c^{2n}}{n(1 - c^2)} \leq 1.
\]

\( N_2 \) is any positive constant such that \( n \geq N_2 \) implies

\[
1 + c^2 - 2\frac{c^2 - (c^2)^{n+1}}{n(1 - c^2)} \leq 1 + c^2.
\]

Additionally, \( N_3 \) is defined as in (S3).

\[
\begin{align*}
V_1 &:= 16\left(\sigma_{\text{max}}^0\right)^4 \left[ p^6 n^{1 - \frac{3p}{2}} \frac{\|\Gamma_n(0)\|_2^2}{\xi} + \left(\frac{3 + c^4}{(1 - c^2)^3 n}\right) \right] \\
&\quad + \frac{\delta^{-2} p^2}{(1 - c^2)\gamma n^{1 - 2\gamma}} + \left(\frac{3 + c^4}{(1 - c^2)^3 \xi}\right),
\end{align*}
\tag{9}
\]

with \( \xi = \frac{2\delta^2}{9\Lambda_g} \varepsilon \). The alternate \( \tilde{V}_1 \) denotes \( V_1 \) with \( \varepsilon \) replaced by \( c^2 \cdot \frac{9n^{1 + \frac{2}{3} - 3\varepsilon}}{2} \).

\[
\begin{align*}
V_2 &:= 4\delta^{-2} \left(\sigma_{\text{max}}^0\right)^4 \frac{(1 + c^2)}{(1 - c^2)^3} \frac{2\min\{|G|, p\}^2}{n} \\
V_3 &:= \frac{V_2}{4} + \delta^{-2} \left[ \frac{2p\left(\sigma_{\text{max}}^0\right)^2}{n(1 - c^2)} + \frac{p(p + 1)}{n}\right].
\end{align*}
\tag{10}
\tag{11}
\]

3.2. Results

Our strategy for establishing graph selection consistency in Theorem 3.12 is largely composed of the contents of Lemmas 3.6 and 3.8 and Theorems 3.9 and 3.10. Lemmas 3.6 and 3.8 describe, respectively, the GF concentration of the VAR(1) transition matrix around its least squares estimate and the concentration of the least squares estimate around an approximation to its expectation. The probability bounded in Lemma 3.6 is with respect to the joint GF distribution of \( A_g \) and \( \Sigma \). In contrast, Lemma 3.8 is a concentration inequality with respect to the data-generating mechanism (2), which derives its distribution from the errors \( U^{(t)} \sim N_p(0, I_p) \) for \( t \in \{1, \ldots, n\} \).

In what follows, we chose \( \varepsilon = \Lambda_g \cdot \bar{\varepsilon} \) for some \( \bar{\varepsilon} \) not depending on \( \Sigma \) or \( A_g \).

Lemma 3.6. For any \( G \) with \( |G| \leq np \),

\[
P \left( \|Z'_G\mathbf{W}^{-1}Z_G(\alpha_G - \hat{\alpha}_g)\|^2 \geq \varepsilon \right) \leq \frac{|G|\sqrt{2\Lambda_g}}{\sqrt{\pi\varepsilon}} e^{-\frac{\varepsilon}{2\Lambda_g}},
\]

where \( \hat{\alpha}_g := (Z'_G Z_G)^{-1} Z'_G Y \), and \( \Lambda_g := \|\mathbf{W}^{-1/2} Z_G\|^2_F \).

Recall that \( \Lambda_g \), which comes from the proof of this lemma, is a key component of our suggested default \( \varepsilon \) in (6) and of Condition 3.4. This results from the fact that \( \varepsilon \) must control for \( \Lambda_g \) in order to establish the well-behaved concentration of the GF distribution of \( \alpha_G \) that is...
exhibited by this lemma. The $\mathcal{W}^{-1/2}$ plays the role of appropriately scaling the design matrix $\mathcal{Z}_G$. Observe that for the full model $G = \{1, \ldots, p^2\}$,

$$\Lambda = \left\|\mathcal{W}^{-1/2}\mathcal{Z}\right\|_F^2 = \text{tr}(\mathcal{Z}'\mathcal{W}^{-1}\mathcal{Z}) = \text{tr}\left((\mathcal{A}'\mathcal{A}') \otimes \Sigma^{-1}\right) = \text{tr}(\mathcal{A}'\mathcal{A}') \cdot \text{tr}(\Sigma^{-1}),$$

which gives

$$E_x(\Lambda) = n \cdot \text{tr}(\Gamma_n(0)) \cdot \text{tr}(\Sigma^{-1}). \quad (12)$$

Thus, for a given graph $G$, $\Lambda_G$ is a combined measure of the covariance or dependence among the $p$ univariate time series in the VAR model, the contemporaneous error precision matrix, and the number of observed instances of the time series. This is what makes $\Lambda_G$ effective as a part of $\epsilon$ in the $h$-function for determining the $\epsilon$-admissibility of a given $\alpha_G$. Lemma 3.7 gives a probabilistic bound on $\Lambda_G$ as a function of $n$ and $p$, given the $h$-function constraint that $\min_{1 \leq j \leq p} \{ m^g_j \} \geq d$.

**Lemma 3.7.** For any $G$,

$$P\left(\Lambda_G \geq n^{1+\frac{d}{2}}p^3, \min_{1 \leq j \leq p} \{ m^g_j \} \geq d\right) \leq e^{-\left(\frac{d \cdot \frac{3}{2} \cdot \frac{9}{4} \cdot \frac{n}{\max(\mathcal{A})^2}}{\Lambda_G \cdot \frac{np}{\max(\mathcal{A})^2}}\right)\left\|\mathcal{Z}_G\right\|_F^2} \leq 2^{-\frac{|G|}{\epsilon_1}}.$$

where $\Lambda_G := \left\|\mathcal{W}^{-1/2}\mathcal{Z}_G\right\|_F^2$.

Next, consider the concentration of the least squares estimate.

**Lemma 3.8.** Assume Condition 3.1 holds. Then for all $n \geq \max\{N_1, N_2\}$, and for any $G$ with $|G| \leq np$,

$$P_x\left(\left\|\hat{\alpha}_g - \left(E_x(\mathcal{Z}_G^\prime \mathcal{Z}_G)^{-1}E_x(\mathcal{Z}_G^\prime Y)\right)\right\| \geq \frac{2\epsilon}{9n^{1+\frac{d}{2}}p^3\Lambda_G}\right) \leq V_1,$$

where $V_1$ is as in (9).

Materially, the three preceding lemmas are needed in the proofs of Theorems 3.9 and 3.10, presented next. These theorems are results of the behaviour of the EAS methodology coupled with the GF distribution (i.e., the Jacobian term); they are analogous to studying the behaviour of given priors for a (Bayesian) posterior distribution. Theorem 3.9 is a nonasymptotic concentration inequality that yields an upper bound on the rate at which the expected value (w.r.t. the joint GF distribution of $A_g$ and $\Sigma$) of the $h$-function times the Jacobian term diverges for non-$\epsilon$-admissible graphs.

**Theorem 3.9.** Take any $G$ with $G \not\subseteq G_o$ and $|G| \leq np$, and assume Conditions 3.1 and 3.3 hold. Then for all $n \geq \max\{N_1, N_2\}$,

$$E\left(h\left(\alpha_G, \{\sigma_j\}\right) \left| D_g^\prime \tilde{D}_g\right\| \right) \leq e^{-\frac{1}{2}(1-c)^{-2}(r^g_{\max}+(1+c)^2)\frac{\|\mathcal{Z}_G\|_F^2}{\sqrt{n}}} - \frac{|G|+p}{2} \times \frac{3|G|\sqrt{\Lambda_G} \cdot e^{-\frac{\epsilon}{9\Lambda_G}}}{\sqrt{\epsilon_1}} + e^{-\left(\frac{d \cdot \frac{3}{2} \cdot \frac{9}{4} \cdot \frac{n}{\max(\mathcal{A})^2}}{\Lambda_G \cdot \frac{np}{\max(\mathcal{A})^2}}\right)\left\|\mathcal{Z}_G\right\|_F^2} \times 2^{-\frac{|G|}{\epsilon_1} + 1}$$

with probability exceeding $1 - V_1$, where $V_1$ is as in (9), $r^g_{\max} := \max_{1 \leq j \leq p} |r^g_j|$. DOI: 10.1002/cjs.11726
Conversely, Theorem 3.10 is a nonasymptotic lower bound on the $h$-function times the Jacobian term for the oracle graph, $G_o$.

**Theorem 3.10.** Assume Conditions 3.1, 3.2, and 3.4 hold. Then for all $n \geq \max\{N_1, N_2, N_3\}$, with $N_3$ and the fixed $K_3 \in (0, 1)$ defined by (S3),

$$P_x \left( E \left( h \left( \alpha_{G_o}, \{\sigma_j\} \right) \mid \bar{D}_{s_0} \bar{D}_{s_0}' \right) \right) \geq (1 - K_3)e^{\frac{|G_o|^2}{4}} \geq 1 - V_1 - \tilde{V}_1 - 2V_2 - 2e^{-\frac{np}{4}} - V_3,$$

where $V_1$ and $\tilde{V}_1$, $V_2$, and $V_3$ are as in (9), (10), and (11), respectively.

Before stating the main result of this article, one final condition, Condition 3.11, is needed. In its absence, a weaker (yet still meaningful) statement of GF graphical consistency holds; we formulate this alternative statement as Corollary 3.13. The importance of Condition 3.11 is that it covers the gap left open in Theorem 3.9, since the theorem only bounds the GF probability of non-$\varepsilon$-admissible graphs (i.e., of those $G \notin G_o$).

**Condition 3.11.** For the positive constant $K_2$ specified in (A1),

$$\max_{G:G \subset G_o} \left\{ \frac{e^{K_2 \left( \frac{||Y||^2}{\sqrt{n}} + p^2 \log(n) \right)} \prod_{j=1}^{p} \left[ \frac{m_{G}^{G_o} - r_{G_o}^{G_o}}{m_{G}^{G} - r_{G}^{G}} \right]^{\frac{n - |r_{G_o}^{G_o}|}{2}} \right\} \rightarrow 0$$

as $n \to \infty$ or $n, p \to \infty$.

Recall from (5) that $m_{G}^{G}$ is the univariate RSS, corresponding to graph $G$, for the $j$th component of the VAR(1) model. Hence, this condition is a statement that the product of the ratio of RSS components for the true graph over that of any strict subgraph, taken to a power on the order of $n$, will vanish at a rate of $\exp \left\{ \frac{||Y||^2}{\sqrt{n}} \right\} = O_p \left( \exp \left\{ p^2 \sqrt{n} \right\} \right)$. This is not unreasonable to expect since for each $j \in \{1, \ldots, p\}$, $m_{G}^{G} = O_p(n)$, $m_{G_o}^{G_o} = O_p(n)$, $m_{G}^{G} \leq m_{G}^{G_o}$ for $G \in \{G: G \subset G_o\}$, and an explicit condition that the oracle model fits sufficiently better than all submodels is typical of model consistency results.

The main result of the present article, a statement of pairwise graphical selection consistency for the constructed EAS methodology, is now presented. This result demonstrates that the GF probability of the oracle graph will asymptotically dominate that of all other graphs.

**Theorem 3.12** (Pairwise selection consistency). Given Conditions 3.1–3.11, for any $G \subseteq \{1, \ldots, p^2\} \setminus G_o$,

$$\frac{r(G|Y)}{r(G_o|Y)} \rightarrow 0$$

as $n \to \infty$ or $n, p \to \infty$.

If Condition 3.11 is violated, Corollary 3.13 demonstrates that the GF mass function $r(G|Y)$ will concentrate asymptotically on the subset of graphs $\{G: G \subseteq G_o\}$. In practice, for sufficiently large $n$, this means that there will be a few graphs that the algorithm visits frequently, and the largest one (in cardinality) likely contains the greatest number of the oracle components.
Corollary 3.13 (Pairwise selection consistency). Relaxing Condition 3.11 in Theorem 3.12 gives, for any $G \subseteq \{1, \ldots, p^2\} \setminus \{G : G \subseteq G_o\}$,

$$
\frac{r(G|Y)}{r(G_o|Y)} \xrightarrow{P_x} 0
$$

as $n \to \infty$ or $n, p \to \infty$.

The additional corollary stated next demonstrates that the EAS methodology will concentrate all GF mass on the true model, asymptotically, for fixed $p$.

Corollary 3.14 (Strong selection consistency, fixed p). Given Conditions 3.1–3.11 and fixed $p$,

$$
\frac{r(G_o|Y)}{r(G_o|Y)} \xrightarrow{P_x} 1
$$

as $n \to \infty$.

Note the following short remark about the meaning of the difference between pairwise and strong model selection consistency. The statement of strong graph selection consistency is essentially a statement that the true model will be assigned large probability and all other models will be assigned small probabilities. Conversely, the implication of pairwise graph selection consistency is that the probability assigned to the true model will be large relative to each of the other model probabilities individually, but that all models (including the true model) may have small probabilities. Such a phenomenon is common for model selection paradigms in which the set of candidate models grows very fast with dimension (i.e., like $2^{p^2}$ in the case of a VAR(1) model).

4. SIMULATION RESULTS

While the theoretical pursuits of this article have been focused on the conditions and supporting lemmas/theorems needed for the EAS procedure to assign the highest probability to the oracle graph with probability converging to 1 as $n, p \to \infty$, we ultimately designed the EAS approach with more practical intuitions in mind. In applications, the true data-generating model, $G_o$, may itself contain redundant information (i.e., unnecessary active components), and through our $h$-function methodology, we are able to focus on recovering only the necessary active components. In doing so, at least for finite samples, the EAS approach re-defines what is meant by the true graph. The purpose of our asymptotic considerations was to illustrate the conditions needed for our re-defined notion of the true graph to correspond precisely to the oracle graph.

In this section, we demonstrate on synthetic data that when the theoretical conditions are satisfied, the EAS procedure performs as our asymptotic theory suggests, and is also able to perform as well as or better than existing methods in high-dimensional settings with respect to out-of-sample prediction error and estimation error. In fact, we find and present evidence to suggest that Conditions 3.1 and 3.2 are useful for high-dimensional settings. Moreover, Condition 3.1 is a simple and verifiable condition for actual observed data that inform of the sample size needed for competitive performance and is so well calibrated that we demonstrate deteriorating performance when it is not satisfied.

Furthermore, the EAS algorithm, with $\varepsilon = \Lambda_g$, does not require any tuning parameter to achieve out-of-sample predictive performance matching or exceeding that of competing methods such as LASSO or elastic net. The latter, more conventional methods, require cross-validation over a grid of tuning parameters, and the appropriateness of the grid depends on the scaling of the data (i.e., $\Sigma$). In contrast, via our $\Lambda_g$ component, the EAS algorithm is scale-invariant.
Similar invariance properties are also achieved for square-root LASSO estimators (e.g., Belloni, Chernozhukov & Wang, 2011; Xie & Xiao, 2018). For numerical results, the component \(d\) in the \(h\)-function is set at \(d = \min_{1 \leq j \leq p} \{m_j^{\text{en}}\}/10\), where \(G_{\text{en}}\) are the active components estimated by elastic net. And as discussed previously, the constraint \(|A_g|_2 \leq c\) in the \(h\)-function is replaced with \(|A_g|_2 < 1\) since \(c\) is not available on real data.

In the following two subsections, we present both low \((p = 4, n = 120)\) and high \((p = 10, n = 20\) and \(p = 30, n = 180)\) dimensional simulation studies on synthetic data generated according to model (2). For each of 100 random data-generating seeds, the transition matrix is randomly generated according to each of the five patterns described in Han, Lu & Liu (2015). In each instance of a transition matrix \(A^0\), the \(p\) diagonal components are active, and for patterns with additional randomly assigned active/inactive components, the probability of each component being generated as active is 0.01. Values of each diagonal component are assigned by sampling from the \(N(\pm 12, 1)\) distribution, while off-diagonal component values are assigned by sampling from the \(N(\pm 3, 1)\) distribution. As is common practice (e.g., Han, Lu & Liu, 2015), after a given \(A^0\) is randomly generated, it is rescaled so that \(|A^0|_2 = 0.5 = c\), and as in Han, Lu & Liu (2015) the contemporaneous error covariance matrix \(\Sigma^0 = I_p\).

In all simulation designs, the performance of the EAS algorithm is compared with that of LASSO and elastic net implementations, and to a recent “direct estimation of high-dimensional stationary VAR” estimation procedure proposed by Han, Lu & Liu (2015), which is formulated as a linear program (we denote this procedure by DELP for “direct estimation linear program”). The LASSO and elastic net routines are implemented using the Python module scikit-learn (Pedregosa et al., 2011), along with their built-in cross-validation procedures for time series data. For the DELP routine, the authors of Han, Lu & Liu (2015) were kind enough to provide their \(R\) code. However, we had to supplement their code by writing additional code to implement the cross-validation procedure proposed in Han, Lu & Liu (2015) for selecting their tuning parameter. Note that we generate synthetic data consistent with that described in Han, Lu & Liu (2015) so that the scaling of the data is appropriate for their default grid of tuning parameters for cross-validation.

For each synthetic data set in the various simulation study scenarios, the LASSO, elastic net, and DELP methods typically finish in under a minute, while the EAS algorithm may need a few minutes/hours. Completion of the EAS algorithm is more complicated because it entails pseudo-marginal MCMC for some specified number of importance samples and some specified number of MCMC steps. To be conservative, we set the number of importance samples and MCMC steps as large as possible while still finishing in approximately 1–2 days, parallelizing over the 100 seeds, 5 coefficient matrix patterns, and 3 simulation designs considered, with most of the time being spent on the designs with \(p = 30, n = 180\). Note that while the runtime required for the EAS algorithm is significantly longer than for optimization-based methods, the latter methods are typically relaxations of \(L_0\)-optimization problems, whereas the EAS method is an explicit solution and is MCMC-based. The statistical literature is saturated with strategies that relax \(L_0\)-optimization problems, and more investigations of explicit solutions could prove useful for sparking promising new ideas (see, e.g., Bertsimas, King & Mazumder, 2016). The code/workflow for reproducing all numerical results can be found at https://jonathanpw.github.io/research.

4.1. Definitions of Performance Metrics
A variety of metrics are considered for evaluating performance across procedures. For each random generator seed for each simulation design, \(2n\) instances of the time series are generated with \(X^{(0)} = 0_{px1}\). The first \(n\) are used for estimation, and the last \(n\) are set aside as an out-of-sample test set. As in Han, Lu & Liu (2015), on the out-of-sample test set, we compute the \(L_2\) prediction error, \(\frac{1}{n}\|\hat{Y} - \hat{A}X\|_2\), and the \(L_F\) prediction error, \(\frac{1}{n}\|\hat{Y} - \hat{A}X\|_F\), where \(\hat{A}\)
represents the estimated transition matrix on the first $n$ in-sample time instances. As in Basu & Michailidis (2015) and Ghosh, Khare & Michailidis (2019), we also calculate the estimation error, $\| \hat{A} - A^0 \|_F / \| A^0 \|_F$. For the EAS procedure, $\hat{A}$ is computed analogously to Bayesian model averaging, with least squares estimates used for every visited graph in the MCMC chain.

Additionally, we report $|G_{\text{MAP}}|$ as the number of nonzero (or active) components in the estimated graph for the frequentist LASSO, elastic net, and DELP procedures, and as the number of active components in the most frequently visited graph (i.e., maximum a posteriori probability or MAP) for the MCMC-based EAS algorithm. The false-positive rate (FPR) is computed as the number of the $p^2$ components in the estimated transition matrix incorrectly set active, as a proportion of the number of truly inactive components. Conversely, the false-negative rate (FNR) is computed as the number of the $p^2$ components in the estimated transition matrix incorrectly set inactive, as a proportion of the number of truly active components. For the EAS procedure, the FPR and FNR are computed based on the estimated $G_{\text{MAP}}$.

### 4.2. Low-Dimensional Setting

This first simulation design serves to demonstrate that the EAS procedure performs consistently with what the theory in Section 3 suggests for data with $p^2 < n$. For this simulation we present two additional performance metrics, $\hat{r}(G_0 | Y)$ and $\#\{G_{\text{MAP}} = G_o\}$. The former is the estimated GF probability of the oracle model, calculated as the number of times the MCMC algorithm visited $G_o$ divided by the number of steps of the chain. This metric is only available within the EAS framework because relative model probabilities are computed. The latter metric, $\#\{G_{\text{MAP}} = G_o\}$, is the proportion, over all 100 generated data sets, of instances in which the estimated $G_{\text{MAP}}$ corresponds precisely to $G_o$.

Observe from Table 2 that the EAS procedure performs very competitively with these existing methods; it shows better average performance metric values across the board, but all routines are within about one standard error of each other. Furthermore, the EAS algorithm selected a $G_{\text{MAP}}$ with 3–4 fewer active components than that of its competitors, on average, with $G_{\text{MAP}} = G_o$ for 81 of the 100 data sets. This is far better graph selection than the competing methods, which consistently over-select active components. Note that based on the proportion of data sets in which Condition 3.2 is satisfied, the oracle model is only identifiable for the EAS algorithm in 83% of the data sets. In other words, our theory would suggest that the EAS procedure should identify the true model in 83 of the 100 data sets considered, and in actuality, the EAS algorithm identified the true model in 81 of the 100 data sets.

### 4.3. High-Dimensional Setting

Tables 3–7 in this section display the results of two high-dimensional simulation designs in which $p^2 > n$, for all five transition matrix patterns. An important distinction to observe between the two designs, for all transition matrix patterns, is that Condition 3.1 is never satisfied for the $p = 10, n = 20$ case, while it is always satisfied for the $p = 30, n = 180$ case. This occurrence is by design to demonstrate the deteriorated performance of the EAS algorithm when this important, well-calibrated, and verifiable condition is not satisfied. In the $p = 30, n = 180$ case, the EAS algorithm performs just as well or better than the competing methods with respect to all metrics.

Notice also that the high-dimensional numerical results presented in this section do not list $\hat{r}(G_0 | Y)$ nor $\#\{G_{\text{MAP}} = G_o\}$ as performance metrics. For each of the estimation methods, the metric $\#\{G_{\text{MAP}} = G_o\}$ (and $\hat{r}(G_0 | Y)$ for EAS) produces zeros in almost all cases. For the EAS algorithm, this is due to the fact that Condition 3.2 is never satisfied for these high-dimensional simulation designs, and so the oracle model is not identifiable for the EAS procedure. However, we would not necessarily expect Condition 3.2 to be satisfied when $p^2 > n$ since our theory does not apply.

DOI: 10.1002/cjs.11726
TABLE 2: Random pattern transition matrix.

|        | oracle | eas  | delp | lasso | enet |
|--------|--------|------|------|-------|------|
| \( p = 4, n = 120 \) |        |      |      |       |      |
| L2     | 1.27   | 1.28 | 1.31 | 1.3   | 1.3  |
|        | (0.11) | (0.11) | (0.12) | (0.12) | (0.12) |
| LF     | 2.01   | 2.02 | 2.03 | 2.03  | 2.03 |
|        | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) |
| est err| 0.17   | 0.21 | 0.32 | 0.32  | 0.33 |
|        | (0.06) | (0.1) | (0.11) | (0.1) | (0.1) |
| \( |G_{\text{MAP}}| \) | 4.12  | 4.0  | 7.84 | 7.94  | 8.39 |
|        | (0.35) | (0.32) | (3.34) | (2.97) | (3.24) |
| FPR    | 0.01   | 0.32 | 0.33 | 0.36  | 0.36 |
|        | (0.02) | (0.28) | (0.25) | (0.27) |      |
| FNR    | 0.04   | 0.01 | 0.01 | 0.01  | 0.01 |
|        | (0.09) | (0.05) | (0.05) | (0.05) |      |
| \( \hat{r}(G_o|Y) \) | 0.7    |      |      |       |      |
|        | (0.32) |      |      |       |      |
| \#\{ G_{\text{MAP}} = G_o \} | 0.81   | 0.08 | 0.11 | 0.11  |      |

Note: See Section 4.1 for definitions of each performance metric, except for the last two that are described in Section 4.2. All metrics are quantities averaged over 100 generated data sets, and standard errors are in parentheses. The “oracle” column displays corresponding characteristics in the case that the oracle graph, \( G_o \), is known, and using the least squares estimate of \( A^0 \). Note that for Condition 3.1, \( 4(1 + c^2) = 5 \). Recall that a new set of active components \( G_o \) are generated for each data set, which gives the variability for \( |G_{\text{MAP}}| \) in the “oracle” column.

Moreover, recall that in finite samples, and particularly in high-dimensional settings with highly correlated data, the EAS framework was developed with the intuition that the oracle graph itself may not be \( \varepsilon \)-admissible. In these settings, the EAS methodology re-defines the notion of the “true” graph to be some nonredundant subgraph of the oracle graph, at least nonasymptotically. This is validated empirically in Tables 3–7 by observing that when Condition 3.1 is satisfied, the EAS algorithm almost always requires fewer active components to achieve on par or better performance than the competing methods, with respect to all metrics. Recall also that the EAS algorithm has no tuning parameter, while the competing methods use cross-validation to optimize out-of-sample prediction accuracy.

5. REAL DATA APPLICATION

As a final exposition of the EAS methodology developed for the VAR(1) model, this section presents the results of implementing the algorithm on real data. Monthly closing stock price data for \( p = 8 \) well-known companies from 1995 to 2018 are downloaded from Yahoo Finance via the R package BatchGetSymbols (Perlin, 2019). First differences of the time series are used for stationarity, and the data are split into two time periods, 1995–2006 (resulting in \( n = 144 \) observed time instances) and 2010–2018 (resulting in \( n = 108 \) observed time instances). Plots
### Table 3: Band pattern transition matrix.

|       | $p = 10, n = 20$ |       | $p = 30, n = 180$ |
|-------|-----------------|-------|-----------------|
|       | oracle | eas | delp | lasso | enet | oracle | eas | delp | lasso | enet |
| L2    | 3.04   | 3.29 | 4.13 | 2.94   | 2.9  | 1.92   | 2.02 | 2.03 | 2.02 | 2.02 |
|       | (0.68) | (0.65) | (5.9) | (0.49) | (0.45) | (0.09) | (0.09) | (0.1) | (0.1) | (0.1) |
| LF    | 3.46   | 3.54 | 3.55 | 3.4    | 3.39 | 5.53   | 5.64 | 5.67 | 5.65 | 5.65 |
|       | (0.22) | (0.24) | (0.71) | (0.18) | (0.18) | (0.06) | (0.06) | (0.07) | (0.07) | (0.07) |
| est err | 1.07 | 1.24 | 1.15 | 0.97   | 0.94 | 0.34   | 0.63 | 0.68 | 0.64 | 0.65 |
|       | (0.17) | (0.18) | (0.77) | (0.04) | (0.05) | (0.03) | (0.06) | (0.05) | (0.06) | (0.06) |
| $|G_{MAP}|$ | 28.0   | 11.64 | 8.89 | 3.43   | 19.63 | 88.0   | 22.32 | 43.43 | 49.78 | 60.85 |
|       | (0.0)  | (2.88) | (23.38) | (4.56) | (18.44) | (0.0)  | (2.92) | (25.39) | (11.55) | (29.62) |
| FPR   | 0.1    | 0.08 | 0.02 | 0.17   | 0.17 | 0.0    | 0.01 | 0.02 | 0.03 | 0.03 |
|       | (0.03) | (0.24) | (0.04) | (0.18) | (0.18) | (0.0)  | (0.03) | (0.01) | (0.03) | (0.03) |
| FNR   | 0.83   | 0.88 | 0.93 | 0.74   | 0.75 | 0.75   | 0.63 | 0.59 | 0.57 | 0.57 |
|       | (0.06) | (0.23) | (0.07) | (0.21) | (0.05) | (0.05) | (0.05) | (0.07) | (0.07) | (0.07) |

Under the null hypothesis, Condition 3.1 is 0.7112 (SE 0.2384) versus 5. Condition 3.2 satisfied for 0% of data sets.

Note: See Note for Table 2.

### Table 4: Cluster pattern transition matrix.

|       | $p = 10, n = 20$ |       | $p = 30, n = 180$ |
|-------|-----------------|-------|-----------------|
|       | oracle | eas | delp | lasso | enet | oracle | eas | delp | lasso | enet |
| L2    | 2.66   | 3.52 | 10.35 | 3.22 | 3.16 | 1.91 | 1.96 | 2.03 | 2.01 | 2.01 |
|       | (0.37) | (0.85) | (44.59) | (0.51) | (0.48) | (0.1) | (0.12) | (0.11) | (0.11) | (0.11) |
| LF    | 3.28   | 3.6 | 4.05 | 3.51 | 3.48 | 5.5 | 5.55 | 5.63 | 5.61 | 5.61 |
|       | (0.16) | (0.24) | (2.06) | (0.18) | (0.18) | (0.06) | (0.07) | (0.07) | (0.06) | (0.06) |
| est err | 0.48 | 1.08 | 1.42 | 0.95 | 0.92 | 0.17 | 0.34 | 0.5 | 0.46 | 0.46 |
|       | (0.12) | (0.16) | (1.49) | (0.05) | (0.06) | (0.03) | (0.09) | (0.05) | (0.05) | (0.05) |
| $|G_{MAP}|$ | 10.39 | 12.16 | 18.51 | 4.33 | 21.05 | 31.24 | 27.64 | 42.28 | 47.76 | 48.14 |
|       | (0.68) | (2.28) | (34.53) | (5.52) | (19.45) | (1.26) | (2.47) | (23.75) | (8.28) | (8.4) |
| FPR   | 0.09   | 0.17 | 0.03 | 0.18   | 0.18 | 0.0   | 0.01 | 0.02 | 0.02 | 0.02 |
|       | (0.03) | (0.35) | (0.04) | (0.19) | (0.19) | (0.0)  | (0.03) | (0.01) | (0.01) | (0.01) |
| FNR   | 0.63   | 0.68 | 0.8 | 0.55   | 0.29 | 0.14 | 0.04 | 0.04 | 0.04 | 0.04 |
|       | (0.14) | (0.33) | (0.19) | (0.29) | (0.29) | (0.08) | (0.04) | (0.04) | (0.04) | (0.04) |

Under the null hypothesis, Condition 3.1 is 0.6941 (SE 0.2265) versus 5. Condition 3.2 satisfied for 0% of data sets.

Note: See Note for Table 2.

DOI: 10.1002/cjs.11726
Table 5: Hub pattern transition matrix.

|       | oracle | eas  | delp | lasso | enet |
|-------|--------|------|------|-------|------|
| L2    | 3.06   | 3.3  | 6.52 | 3.02  | 2.98 |
|       | (0.64) | (0.66)| (29.28)| (0.61)| (0.58)|
| LF    | 3.44   | 3.54 | 3.68 | 3.41  | 3.4  |
|       | (0.21) | (0.2) | (1.53)| (0.2) | (0.2) |
| est err | 1.06  | 1.23 | 1.26 | 0.98  | 0.96 |
|       | (0.17) | (0.17)| (1.15)| (0.03)| (0.05)|
| |       |       |      |       |      |      |
| \(|G_{\text{MAP}}\) | 26.0   | 11.95| 11.9 | 2.89  | 19.74|
|       | (0.0)  | (2.48)| (29.42)| (4.84)| (20.78)|
| FPR   | 0.1    | 0.11 | 0.02 | 0.17  | 0.17 |
|       | (0.03) | (0.3) | (0.04)| (0.2) |      |
| FNR   | 0.82   | 0.86 | 0.94 | 0.73  | 0.73 |
|       | (0.07) | (0.29)| (0.09)| (0.25)|      |

Note: See Note for Table 2.

Table 6: Random pattern transition matrix.

|       | oracle | eas  | delp | lasso | enet |
|-------|--------|------|------|-------|------|
| L2    | 2.6    | 3.36 | 5.21 | 3.06  | 3.03 |
|       | (0.4)  | (0.8)| (9.99)| (0.49)| (0.5) |
| LF    | 3.24   | 3.55 | 3.71 | 3.45  | 3.43 |
|       | (0.17) | (0.26)| (0.95)| (0.21)| (0.22)|
| est err | 0.49  | 1.07 | 1.2  | 0.95  | 0.92 |
|       | (0.15) | (0.16)| (0.82)| (0.06)| (0.07)|
| \(|G_{\text{MAP}}\) | 10.9   | 12.49| 14.51| 3.69  | 19.28|
|       | (0.96) | (2.68)| (30.43)| (4.68)| (20.07)|
| FPR   | 0.1    | 0.13 | 0.02 | 0.17  | 0.17 |
|       | (0.03) | (0.31)| (0.03)| (0.19) |      |
| FNR   | 0.63   | 0.72 | 0.82 | 0.58  | 0.58 |
|       | (0.15) | (0.31)| (0.18)| (0.31) |      |

Note: See Note for Table 2.
TABLE 7: Scale–free pattern transition matrix.

|                  | \( p = 10, n = 20 \) |                  | \( p = 30, n = 180 \) |
|------------------|----------------------|------------------|-----------------------|
|                  | oracle eas delp lasso enet |                  | oracle eas delp lasso enet |
| L2               | 3.37 3.19 8.31 2.86 2.83 | 1.94 2.0 2.01 2.01 2.0 | (0.81) (0.75) (0.47) (0.47) (0.47) | (0.09) (0.1) (0.1) (0.1) (0.1) |
| LF               | 3.51 3.49 3.83 3.36 3.35 | 5.53 5.61 5.62 5.63 5.62 | (0.22) (0.22) (0.19) (0.18) (0.18) | (0.06) (0.06) (0.06) (0.07) (0.06) |
| est err          | 1.36 1.32 1.56 0.99 0.96 | 0.52 0.84 0.87 0.9 0.86 | (0.27) (0.19) (0.03) (0.04) (0.04) | (0.04) (0.04) (0.04) (0.05) (0.05) |
| \( |G_{\text{MAP}}| \) | 28.0 11.8 15.88 2.07 16.97 | 88.0 15.14 23.22 17.66 77.21 | (0.0) (2.85) (3.76) (18.88) (0.0) | (2.24) (4.07) (10.06) (35.83) |
| FPR              | 0.1 0.15 0.01 0.15 | 0.0 0.01 0.0 | (0.03) (0.34) (0.03) (0.18) | (0.0) (0.0) (0.0) (0.03) |
| FNR              | 0.84 0.82 0.96 0.78 | 0.86 0.79 0.83 0.63 | (0.07) (0.34) (0.07) (0.22) | (0.03) (0.04) (0.09) (0.11) |

r.h.s. Condition 3.1 is satisfied for 0% of data sets.

Note: See Note for Table 2.

of each time series for each time period are available with the code for reproducing our results at https://jonathanpw.github.io/research. In accordance with Lin & Michailidis (2017), due to a severe financial crisis, the time period 2007–2009 is omitted from the analysis. Condition 3.1 is approximately satisfied for the time period 2010–2018 (7.17 vs. 8 = 4(1 + 12)), but not for 1995–2006 (2.05 vs. 8 = 4(1 + 12)).

The results are displayed graphically in Figure 2. Nodes represent individual company stocks, and each edge label represents the marginal GF (or GF) inclusion probability of a particular component of \( A \), that is, the proportion of graphs, \( G \), (over all MCMC-sampled graphs) in which each component (i.e., edge) of \( A \) is active. Line widths are proportional to inclusion probabilities, and inclusion probabilities less than 0.05 are omitted.

Figure 2 suggests a variety of interesting findings. It is well known in the economics literature that the auto industry is strongly tied to economic cycles, but the pharmaceutical industry is not (Berman & Pfleeger, 1997). The high inclusion probability of 0.71 from Ford to Pfizer in the left panel of Figure 2 may reflect the early 2000 dot-com bubble burst, which hit the Ford stock price significantly (Haugh, Mourougane & Chatal, 2010); the Ford stock price may have acted as a proxy for the overall effect of economic activity on the Pfizer stock price. Further, Ford vehicles consisted of 25.5% of the US market share in 1995 (Knoema, 2020), but subsequently reduced to 14.29% (Carlier, 2021). Accordingly, Ford had less influence on the US economy during 2010–2018, which reflects the missing directed edge from Ford to Pfizer in the right panel of Figure 2 (Slickcharts, 2021). Moreover, the pharmaceutical industry is also largely driven by its own cycles due to planning and release of clinical trial results on specific dates each year, which likely explains the arrow from Pfizer to itself having an inclusion probability of 0.27.
FIGURE 2: Directed graph of marginal inclusion probabilities of components of the transition matrix, $A$, for monthly closing stock price of $p = 8$ companies, with $n = 144$ in the left panel and $n = 108$ in the right panel. First differences of the data are used. Each edge label represents the marginal GF (or GF) inclusion probability of a particular component of $A$, that is, the proportion of graphs, $G$, (over all MCMC-sampled graphs) in which each component (i.e., edge) of $A$ is active. Line widths are proportional to inclusion probabilities, and inclusion probabilities less than 0.05 are omitted.

Additionally, there are interesting interpretations for why the inclusion probabilities change between these two periods. For example, there is a link between Apple and Intel in the left panel of Figure 2 but not in the right panel, which may be attributed to the fact that the iPhone, first released in 2007, eventually eclipsed the Mac computer as Apple’s largest revenue source. More broadly, the high inclusion probability of 0.61 from Walmart to Apple reflects Apple’s corporate transition to a more general consumer product company in the 2010–2018 period, which is driven by the broader consumer market for electronics, as proxied by Walmart. The arrow does not point in the reverse direction, most likely because Walmart is a retailer in a larger consumer electronics market that includes more than just high-end products like those from Apple. Conversely, Microsoft is not so much a general consumer product company as compared with Apple, and does not have directed edges coming from Walmart.

Ultimately, though, the real data analysis conducted here is not a thorough investigation of these time series, but rather a proof of concept for how the EAS methodology can be useful on real data. A well-qualified study would require considerable additional analysis of the data, which is beyond the scope of the present article. Note that such a graphical representation of the results, with marginal inclusion probabilities for all components of $A$, is not possible via frequentist nor Bayesian point estimation based procedures. This is a major advantage of estimating relative model probabilities (i.e., $r(G|Y)$) versus simply coefficients. MCMC-based approaches are computationally more expensive, but they provide more information for uncertainty quantification.

6. CONCLUDING REMARKS

In summary, while BVAR models have been developed and explored empirically (primarily in the econometrics literature), there exist very few theoretical investigations of the repeated-sampling properties for BVAR models in the literature, and none for GF VAR models. To the best of our knowledge, our established pairwise and strong model selection consistency results are the first of their kind in the GF/Bayesian VAR literature. These types of results are sure to be followed by similar results in the high-dimensional BVAR literature, analogous to the
emergence of model selection with strong consistency results in the high-dimensional Bayesian linear regression literature such as Johnson & Rossell (2012), Narisetty & He (2014), and Williams & Hannig (2019).

All things considered, while it is required for our theory that \( n \) exceeds some polynomial of \( p \), consistent with our survey of the literature, it is claimed in Ghosh, Khare & Michailidis (2019) that general posterior consistency results are not available for “large \( p \), small \( n \)” settings. Furthermore, our graphical selection consistency results provide a theoretical guarantee for model selection, which is stronger than establishing estimation consistency of a point estimator of the VAR model parameters. As this is the first article to prove asymptotic theory for model selection consistency in a GF/Bayesian VAR model, we are forced to rely mostly on nonasymptotic second-moment concentration bounds. High-dimensional (\( p > n \)) consistency results require exponential tail bounds when establishing concentration of data-dependent quantities such as in Lemma 3.8. In future work, we hope to learn how to derive such exponential tail bounds for multivariate GF model selection theory.

Finally, recall that in finite samples, and particularly in high-dimensional settings with highly correlated data, the EAS framework was developed with the intuition that the oracle graph itself may not be \( \epsilon \)-admissible. In these settings, the EAS methodology re-defines the notion of the “true” graph to be some nonredundant subgraph of the oracle graph, at least nonasymptotically. Accordingly, with our EAS methodology, we hope to demonstrate the idea that to develop inherently scalable methodology, the key may be to re-think what one should hope to recover for useful statistical inference from a data-generating model.

REFERENCES

Ahelegbey, D. F., Billio, M., & Casarin, R. (2016). Sparse graphical vector autoregression: A Bayesian approach. Annals of Economics and Statistics/Annales d’Économie et de Statistique, 123/124, 333–361.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics, 37(2), 697–725.

Banbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector auto regressions. Journal of Applied Econometrics, 25(1), 71–92.

Basu, S., Li, X., & Michailidis, G. (2019). Low rank and structured modeling of high-dimensional vector autoregressions. IEEE Transactions on Signal Processing, 67(5), 1207–1222.

Basu, S. & Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics, 43(4), 1535–1567.

Belloni, A., Chernozhukov, V., & Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. Biometrika, 98(4), 791–806.

Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. The Annals of Statistics, 37, 905–938.

Berman, J. & Pfleeger, J. (1997). Which industries are sensitive to business cycles. Monthly Labor Review, 120, 19.

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. The Annals of Statistics, 44(2), 813–852.

Candes, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when \( p \) is much greater than \( n \). The Annals of Statistics, 35(6), 2313–2351.

Carlier, M. (2021). Ford motor company’s vehicle sales market share in the United States from 2015 to 2020.

Edlefsen, P. T., Liu, C., & Dempster, A. P. (2009). Estimating limits from Poisson counting data using Dempster–Shafer analysis. The Annals of Applied Statistics, 3, 764–790.

Fiecas, M., Leng, C., Liu, W., & Yu, Y. (2019). Spectral analysis of high-dimensional time series. Electronic Journal of Statistics, 13(2), 4079–4101.

Fraser, D. A. S. (2019). The p-value function and statistical inference. The American Statistician, 73(sup1), 135–147.
George, E. I., Sun, D., & Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, 142(1), 553–580.

Ghosh, S., Khare, K., & Michailidis, G. (2019). High dimensional posterior consistency in Bayesian vector autoregressive models. *Journal of the American Statistical Association*, 114(526), 735–748.

Ghosh, S., Khare, K., & Michailidis, G. (2021). Strong selection consistency of Bayesian vector autoregressive models based on a pseudo-likelihood approach. *The Annals of Statistics*, 49(3), 1267–1299.

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2), 436–451.

Giurcanu, M. C. (2017). Oracle m-estimation for time series models. *Journal of Time Series Analysis*, 38(3), 479–504.

Guo, S., Wang, Y., & Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika*, 103(4), 889–903.

Hall, E. C., Raskutti, G., & Willett, R. M. (2019). Learning high-dimensional generalized linear autoregressive models. *IEEE Transactions on Information Theory*, 65(4), 2401–2422.

Han, F., Lu, H., & Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research*, 16(1), 3115–3150.

Hannig, J., Iyer, H., Lai, R. C. S., & Lee, T. C. M. (2016). Generalized fiducial inference: A review and new results. *Journal of American Statistical Association*, 111(515), 1346–1361.

Haugh, D., Mourougane, A., & Chatal, O. (2010). The automobile industry in and beyond the crisis. In *OECD Economics Department Working Papers*, No. 745, OECD Publishing, Paris.

Jacob, P. E., Gong, R., Edlefsen, P. T., & Dempster, A. P. (2021). A Gibbs sampler for a class of random convex polytopes. *Journal of the American Statistical Association*, 116, 1181–1192.

Jameson, G. J. O. (2013). Inequalities for gamma function ratios. *American Mathematical Monthly*, 120, 936–940.

Johnson, V. E. & Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498), 649–660.

Knoema (2020). *Top vehicle manufacturers in the US market*, 1961–2016. https://knoema.com/infographics/floslle/top-vehicle-manufacturers-in-the-us-market-1961-2016.

Korobilis, D. (2013). VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, 28(2), 204–230.

Lin, J. & Michailidis, G. (2017). Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *The Journal of Machine Learning Research*, 18(1), 4188–4236.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions–Five years of experience. *Journal of Business & Economic Statistics*, 4(1), 25–38.

Loh, P.-L. & Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3), 1637.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*, Springer Science & Business Media, New York.

Martin, R. & Liu, C. (2015). *Inferential Models: Reasoning with Uncertainty*, Vol. 145, CRC Press, Boca Raton.

Medeiros, M. C. & Mendes, E. F. (2016). $\ell_1$-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1), 255–271.

Narisetty, N. N. & He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2), 789–817.

Negahban, S. & Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2), 1069–1097.

Nicholson, W. B., Wilms, I., Bien, J., & Matteson, D. S. (2020). High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, 21, 166–161.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Perlin, M. (2019). BatchGetSymbols: Downloads and organizes financial data for multiple tickers. https://CRAN.R-project.org/package=BatchGetSymbols
Raskutti, G., Yuan, M., & Chen, H. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3), 1554–1584.

Safikhani, A. & Shojaie, A. (2022). Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *Journal of the American Statistical Association*, 117(537), 251–264.

Schweder, T. & Hjort, N. L. (2016). *Confidence, Likelihood, Probability*, Vol. 41, Cambridge University Press, Cambridge.

Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*, CRC Press, Boca Raton.

Slickcharts (2021). S&P 500 Companies by Weight. https://www.slickcharts.com/sp500.

Taraldsen, G. & Lindqvist, B. H. (2013). Fiducial theory and optimal inference. *The Annals of Statistics*, 41, 323–341.

Veronese, P. & Melilli, E. (2015). Fiducial and confidence distributions for real exponential families. *Scandinavian Journal of Statistics*, 42, 471–484.

Williams, J. P. & Hannig, J. (2019). Nonpenalized variable selection in high-dimensional linear model settings via generalized fiducial inference. *The Annals of Statistics*, 47(3), 1723–1753.

Xie, F. & Xiao, Z. (2018). Square-root LASSO for high-dimensional sparse linear systems with weakly dependent errors. *Journal of Time Series Analysis*, 39(2), 212–238.

Xie, M. & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81, 3–39.

APPENDIX

This section provides statements of key supporting results, as well as proofs of our main theorem and its corollaries. See the Supplementary Material for proofs of all other results.

Five lemmas that are foundational to our proof techniques for establishing our theory for the EAS methodology are presented next. Nonasymptotic moment bounds on products of $\mathbf{A}'$ and $\mathbf{U}'$ (in the VAR model formulation (1)), with respect to $n$ and $p$, are the building blocks for any theoretical pursuit of understanding high-dimensional VAR models. These results are essentially a collection of second-moment bounds of the quantities and cross-quantities in $\Omega := \frac{1}{n} \begin{pmatrix} \mathbf{A}' \mathbf{X}' & \mathbf{X} \mathbf{U}'' \\ \mathbf{U}' \mathbf{X}' & \mathbf{U} \mathbf{U}'' \end{pmatrix}$.

**Lemma A.1.** Assume $\|A^0\|_2 \leq c$. Then for all $n \geq \max\{N_1, N_2\}$,

$$P_x \left( \left[ \lambda_{\text{min}} \left( \Omega - E_x(\Omega) \right) \right]^2 > \delta^2 \right) \leq V_3,$$

where $V_3$ is as in (11), $\Omega := \frac{1}{n} \begin{pmatrix} \mathbf{A}' & \mathbf{X} \mathbf{U}'' \\ \mathbf{U}' & \mathbf{U} \mathbf{U}'' \end{pmatrix}$, and $E_x(\Omega) = \left( \Gamma_n(0) I_p \right)$.

**Lemma A.2.** Assume $\|A^0\|_2 \leq c$. Then for all $n \geq N_1$,

$$\frac{1}{n^2} \text{tr} \left( E_x(\mathbf{A}' \mathbf{U}'' \mathbf{U} \mathbf{X}') \right) \leq \frac{p \sigma_{\text{max}}^2 \min \{ |G_o|, p \}^2}{n(1 - c^2)}.$$

**Lemma A.3.** Assume $\|A^0\|_2 \leq c$. Then for all $n \geq N_2$,

$$\text{tr} \left( \frac{1}{n^2} E_x \left( (\mathbf{A}' \mathbf{X}')^2 \right) - \Gamma_n^2(0) \right) \leq \frac{\delta^2}{4} V_2.$$
Lemma A.4. Assume \( \| A^0 \|_2 \leq c \). Then,

\[
\frac{1}{n^2} \text{tr} \left( E_x (X X' X U' A^0) \right) \leq \frac{2 (\sigma_{\max}^0)^3 c^2 \min \{|G_o|, p \}^2}{(1 - c^2)n}.
\]

Lemma A.5. Assume that Condition 3.1 holds. Then for all \( n \geq N_2 \),

\[
P_x \left( \lambda_{\min}(X X' / n) \geq \delta / 2 \right) \geq 1 - V_2,
\]

where \( V_2 \) is as in (10).

Proof of Theorem 3.12. Assume throughout this proof that \( n \geq \max\{N_1, N_2, N_3\} \) (see Definition 3.5). From (8),

\[
\frac{r(G|Y)}{r(G_0|Y)} = (2\pi)^{\frac{|G|-|G_0|}{2}} n^{\frac{|G_0|-|G|}{2}} E \left( h \left( \alpha_G, \{\sigma_j\} \right) |D_{G_0} D_G \right)^{\frac{1}{2}} / (h \left( \alpha_G, \{\sigma_j\} \right) |D_{G_0} D_G \right)^{\frac{1}{2}} E \left( h \left( \alpha_G, \{\sigma_j\} \right) | \tilde{D}_{G_0} \tilde{D}_G \right)^{\frac{1}{2}}
\]

\[
\times \prod_{j=1}^{p} \left| (X X')_{r^0_j} r^0_j \right|^{-\frac{n-|r^0_j|}{2}} \left( \frac{m_j}{2} \right)^{-\frac{n-|r^0_j|}{2}} \Gamma \left( \frac{n-|r^0_j|}{2} \right)
\]

From Jameson (2013),

\[
\Gamma \left( \frac{n-|r^0_j|}{2} \right) \leq \begin{cases} 
\left( \frac{n-|r^0_j|}{2} \right)^{-\frac{|r^0_j|}{2}} & \text{if } |r^0_j| - |r_j^0| \geq 1 \\
1 & \text{if } |r_j^0| - |r_j^0| = 0 \\
\left( \frac{n-|r^0_j|}{2} - 1 \right)^{-\frac{|r^0_j| - |r^0_j|}{2}} & \text{if } |r_j^0| - |r_j^0| \leq -1
\end{cases}
\]

and so for \( n - p \geq 4 \),

\[
\prod_{j=1}^{p} \frac{\Gamma \left( \frac{n-|r^0_j|}{2} \right)}{\Gamma \left( \frac{n-|r^0_j|}{2} \right)} \leq \prod_{j=1}^{p} \left\{ \begin{array}{ll}
\left( \frac{n}{2} \right)^{-\frac{|r^0_j|}{2}} & \text{if } |r^0_j| - |r_j^0| \geq 3 \\
\left( \frac{n}{2} \right)^{\frac{n-p}{2}} & \text{if } |r_j^0| - |r_j^0| = 2 \\
\left( \frac{n}{2} \right)^{-\frac{n-p}{2}} & \text{if } |r_j^0| - |r_j^0| = 1 \\
1 & \text{if } |r_j^0| - |r_j^0| = 0 \\
\left( \frac{n}{2} - 1 \right)^{-\frac{|r^0_j| - |r^0_j|}{2}} & \text{if } |r_j^0| - |r_j^0| \leq -1
\end{array} \right.
\]
\[ \leq \prod_{j=1}^{p} \left( \frac{n}{2} \right) \max \left\{ \frac{r_{\sigma_0}}{2}, 1 \right\} \]
\[ \leq \left( \frac{n}{2} \right)^{|G_0| + p}. \]

This bound, together with the simplification,
\[ \prod_{j=1}^{p} \left( \frac{m_{\sigma_j}^2}{2} \right)^{n - |r_{\sigma_j}| - \frac{1}{2}} = 2^{\frac{|G_0| - |G|}{2}} \prod_{j=1}^{p} \left( \frac{m_{\sigma_j}^2}{2} \right)^{a - |r_{\sigma_j}| - \frac{1}{2}} \]
gives
\[ \frac{r(G|Y)}{r(G_0|Y)} \leq \left( \frac{\pi}{n} \right) \left( \frac{n}{2} \right)^{\frac{|G_0| - |G|}{2} + p} \frac{E \left( h \left( \alpha_G, \{\sigma_j\} \right) | \tilde{D}_g \tilde{D}_g \right)}{E \left( h \left( \alpha_{G_0}, \{\sigma_j\} \right) | \tilde{D}_g \tilde{D}_g \right)} \]
\[ \times \prod_{j=1}^{p} \left[ \left( \frac{\sigma_{\tilde{D}_g}}{\sigma_{\tilde{D}_g}} \right)^{\frac{1}{2}} \left( \frac{m_{\sigma_j}^2}{2} \right)^{a - |r_{\sigma_j}| - \frac{1}{2}} \right]. \]

Further, by Lemmas S2.2 and S2.5,
\[ \frac{r(G|Y)}{r(G_0|Y)} \leq \left( \frac{\pi}{n} \right) \left( \frac{n}{2} \right)^{\frac{|G_0| - |G|}{2} + p} \frac{E \left( h \left( \alpha_G, \{\sigma_j\} \right) \right) \sigma_{\tilde{D}_g}^{\frac{1}{2} (1-c)^{-2} (r_{\sigma_j}^\delta + (1+c)^{1-\delta}) \frac{|Y|^2}{\sqrt{n}} - \frac{|G| + p}{2}}}{E \left( h \left( \alpha_{G_0}, \{\sigma_j\} \right) \right) \sigma_{\tilde{D}_g}^{\frac{1}{2}}} \]
\[ \times \prod_{j=1}^{p} \left[ \left( \frac{m_{\sigma_j}^2}{2} \right)^{a - |r_{\sigma_j}| - \frac{1}{2}} \right] \]

with probability exceeding \( 1 - 2V_2 \), where \( V_2 \) is as in (10). Then by Theorem 3.10, for the fixed \( K_3 \in (0, 1) \),
\[ \frac{r(G|Y)}{r(G_0|Y)} \leq \left( \frac{\pi}{n} \right) \left( \frac{n}{2} \right)^{\frac{|G_0| - |G|}{2} + p} \frac{E \left( h \left( \alpha_G, \{\sigma_j\} \right) \right) \sigma_{\tilde{D}_g}^{\frac{1}{2} (1-c)^{-2} (r_{\sigma_j}^\delta + (1+c)^{1-\delta}) \frac{|Y|^2}{\sqrt{n}} - \frac{|G| + p}{2}}}{(1 - K_3) \sigma_{\tilde{D}_g}^{\frac{1}{2}}} \]
\[ \times \prod_{j=1}^{p} \left[ \left( \frac{m_{\sigma_j}^2}{2} \right)^{a - |r_{\sigma_j}| - \frac{1}{2}} \right] \]

DOI: 10.1002/cjs.11726
with probability exceeding $1 - V_1 - \tilde{V}_1 - 4V_2 - 2e^{-\frac{np}{2}} - V_3$. Gathering terms, for some positive constant $K_2$ (not depending on $n$ nor $p$),

$$
\frac{r(G|Y)}{r(G_o|Y)} \leq E \left( h \left( \alpha_G, \{\sigma_j\} \right) \right) e^{K_2 \left( \frac{\ell(Y)^2}{\sqrt{n}} + p^2 \log(n) \right) \prod_{j=1}^{p} \left( \frac{m_j}{m_j^o} \right)^{n_j-o_j} \frac{1}{2}} \left( \frac{m_j^o}{m_j} \right)^{-n_j-o_j} \frac{1}{2},
$$

with probability exceeding $1 - V_1 - \tilde{V}_1 - 4V_2 - 2e^{-\frac{np}{2}} - V_3$. At this point, $E \left( h \left( \alpha_G, \{\sigma_j\} \right) \right)$ can be bounded as in the following two cases.

**Case 1:** $G \subset G_o$ with $|G| \in \{1, \ldots, |G_o| - 1\}$. In this case, Theorem 3.9 does not apply, so since $h \left( \alpha_G, \{\sigma_j\} \right) \leq 1$ uniformly,

$$
\frac{r(G|Y)}{r(G_o|Y)} \leq e^{K_2 \left( \frac{\ell(Y)^2}{\sqrt{n}} + p^2 \log(n) \right) \prod_{j=1}^{p} \left( \frac{m_j}{m_j^o} \right)^{n_j-o_j} \frac{1}{2}} \left( \frac{m_j^o}{m_j} \right)^{-n_j-o_j} \frac{1}{2},
$$

with probability exceeding $1 - V_1 - \tilde{V}_1 - 4V_2 - 2e^{-\frac{np}{2}} - V_3$. Then by Condition 3.11,

$$
\frac{r(G|Y)}{r(G_o|Y)} \frac{P_x}{r(G_o|Y)} \to 0 \text{ as } n \to \infty \text{ or } n, p \to \infty.
$$

**Case 2:** $G \not\subseteq G_o$ and $|G| \in \{1, \ldots, p^2\}$. By Lemma S2.6, and for some positive constant $K_1$ (not depending on $n$ nor $p$),

$$
\frac{r(G|Y)}{r(G_o|Y)} \leq E \left( h \left( \alpha_G, \{\sigma_j\} \right) \right) e^{K_1 \left( \frac{\ell(Y)^2}{\sqrt{n}} + p^2 \log(n) \right) \prod_{j=1}^{p} \left( \frac{m_j}{m_j^o} \right)^{n_j-o_j} \frac{1}{2}} \left( \frac{m_j^o}{m_j} \right)^{-n_j-o_j} \frac{1}{2},
$$

with probability exceeding $1 - V_1 - \tilde{V}_1 - 5V_2 - 3e^{-\frac{np}{2}} - V_3 - \frac{2 \left( \sigma_{max}^0 \right)^2}{\delta (1-c^2) \sqrt{n}}$. Therefore, by Theorem 3.9 and Condition 3.4,

$$
\frac{r(G|Y)}{r(G_o|Y)} \leq \left( e^{-\frac{c}{m_k}} + e^{-\left( \frac{d_n \omega_n}{4 \lambda x} \frac{p^2}{\sqrt{n}} - \frac{np}{2} \right) \prod_{j=1}^{G_o} \left( \frac{m_j^o}{m_j} \right)^{n_j-o_j} \frac{1}{2}} \right) \frac{K_1 \left( \frac{\ell(Y)^2}{\sqrt{n}} + p^2 \log(n) \right) \prod_{j=1}^{p} \left( \frac{m_j}{m_j^o} \right)^{n_j-o_j} \frac{1}{2}} \right) e^{K_1 \left( \frac{\ell(Y)^2}{\sqrt{n}} + p^2 \log(n) \right) \prod_{j=1}^{p} \left( \frac{m_j}{m_j^o} \right)^{n_j-o_j} \frac{1}{2}} \right)
$$

with probability exceeding $1 - 2V_1 - \tilde{V}_1 - 5V_2 - 3e^{-\frac{np}{2}} - V_3 - \frac{2 \left( \sigma_{max}^0 \right)^2}{\delta (1-c^2) \sqrt{n}}$. Thus, by Condition 3.4,

$$
\frac{r(G|Y)}{r(G_o|Y)} \frac{P_x}{r(G_o|Y)} \to 0 \text{ as } n \to \infty \text{ or } n, p \to \infty.
$$

**Proof of Corollary 3.13.** Omit case 1 in the proof of Theorem 3.12.
Proof of Corollary 3.14. Observe that

\[
    r(G_o|Y) = \frac{r(G_o|Y)}{\sum_{j=1}^{p^2} \sum_{G:|G|=j} r(G|Y)} = \frac{1}{1 + \sum_{j=1}^{p^2} \sum_{G\neq G_o:|G|=j} \frac{r(G|Y)}{r(G_o|Y)}}.
\]

Since \( p \) is fixed Theorem 3.12 gives,

\[
    \sum_{j=1}^{p^2} \sum_{G\neq G_o:|G|=j} \frac{r(G|Y)}{r(G_o|Y)} \xrightarrow{P} 0
\]

as \( n \to \infty \), which proves the desired result.

\[\square\]

Received 10 May 2021
Accepted 26 January 2022