Ridge Regularized Estimation of VAR Models for Inference and Sieve Approximation

Giovanni Ballarin[*]
University of Mannheim

April 6, 2021

Abstract: Developments in statistical learning have fueled the analysis of high-dimensional time series. However, even in low-dimensional contexts the issues arising from ill-conditioned regression problems are well-known. Because linear time series modeling is naturally prone to such issues, I propose to apply ridge regression to the estimation of dense VAR models. Theoretical non-asymptotic results concerning the addition of a ridge-type penalty to the least squares estimator are discussed, while standard asymptotic and inference techniques are proven to be valid under mild conditions on the regularizer. The proposed estimator is then applied to the problem of sieve approximation of VAR(∞) processes under moderately harsh sample sizes. Simulation evidence is used to discuss the small sample properties of the ridge estimator (RLS) when compared to least squares and local projection approaches: I use a Monte Carlo exercise to argue that RLS with a lag-adapted cross-validated regularizer achieve meaningfully better performance in recovering impulse response functions and asymptotic confidence intervals than other common approaches.

Keywords: ridge regression, vector autoregression, sieve approximation, impulse responses, time series.

[*] E-mail: `giovanni.ballarin@gess.uni-mannheim.de`.

1 Introduction

The enduring role of the vector autoregressive (VAR) model in macroeconomic research after its introduction by Sims (1980) can not be understated. The extensive textbook treatment of VARs (Lütkepohl, 2005, Kilian and Lütkepohl, 2017) shows that they are a keystone of applied time series research. Autoregressive models are especially popular as a tool to perform structural inference and to study the effects of dynamic shocks. However, the sensitive nature of the estimation problem has lead to studying the weak spots of linear autoregressive models. To name a few of interest for the purposes of this paper, one should consider the issues of model selection and post-selection inference; of large model estimation using data with small time samples; of models with low-rank or sparse structure.

Recent research has seen the introduction of a number of potential alternatives to VARs. Jordà (2005) proposed the local projection (LP) method for impulse response estimation. LPs have the advantages of not imposing a specific dynamical structure on the data, of being robust to nonlinearities and of solving the curse of dimensionality that comes with VAR models (Ramey, 2016)[1]. From a modeling perspective, factor augmented VARs and factor models (Bernanke et al., 2005, Stock and Watson, 2016, 2017) have gained significant traction when dealing with large economic times series data. Factor models have the advantage of reducing even considerable systems by synthesizing the core dynamics of many components down to a handful of underlying factors, but this comes at a cost of imposing a reduced-rank structure on shock propagation. A growing literature has shown interest in applying *regularization* methods to large VARs under assumptions of sparsity (Hsu et al., 2008, Song and Bickel, 2011, Krampe et al., 2020). Sparsity is suggestive in very large models, but has also been seen as a hard-to-defend assumption in reasonably large VARs (e.g. $10 \leq K \leq 100$ components) with data possibly coming from the same economy (Giannone et al., 2018).

This paper details the application of *ridge penalization* (Hoerl and Kennard, 1970, Tikhonov, 1943) as an alternative estimation strategy to ordinary least squares, with a focus on *inference*. Unlike (quasi) maximum likelihood methods or Bayesian methods, regularized least squares allow for a direct estimation strategy with the only additional complication of selecting appropriate regularization parameters. As such, no specific distributional assumptions on the process are needed nor parameter priors. The ridge approach is inherently more robust to settings where data availability is comparably poor. This situation is remarkably common in macroeconometrics, where VAR models might be estimated with as little as two observations per parameter if not less. Ridge regularization addresses the problem of ill-conditioned least squares by adding a positive diagonal matrix Λ to the sample scatter matrix before inversion is performed to solve the

[1] Local projections are also known to recover the same impulse responses as VARs in population (Plagborg-Møller and Wolf, 2021), and of being more robust in the face of high persistency (Montiel Olea and Plagborg-Møller, 2020), therefore they can be an attractive alternative.

normal equations. The cost incurred in adding the ridge penalty is in terms of bias: the larger the parameters in Λ are, the more coefficient estimates are shrunk towards zero with respect to their least squares counterparts. As a trade-off, a ridge estimator has lower finite-sample variance or, equivalently, is less sensitive to ill-conditioned samples. Under appropriate assumptions, I argue that these two effects together can be beneficial in performing VAR impulse response analysis: in simulations, ridge-estimated impulse responses compare favorably to least squares and local projection approaches, especially when the sample size is small. Therefore, the ridge estimation approach introduces a number of contributions that complement the aforementioned alternative frameworks.

I adapt the well-developed literature of ridge regression to the linear time series setting, and provide theoretical results on the shrinkage behavior of ridge and the implications for asymptotic inference. The idea is actually already established in Bayesian modeling, and is mentioned at least as far back as Hamilton (1994): in practice, using the Litterman-Minnesota prior effectively yields the equivalent of ridge regression on the posterior (Litterman, 1985, De Mol et al., 2008). However, the Bayesian perspective has a precise interpretation for the regularization parameter λ, stemming from the nature of the priors. In the frequentist context employed in this paper, I instead treat regularization independently and as such it must be set by the researcher, possibly in a data-driven manner.

First, I give a theoretical treatment of the shrinkage phenomenon when ridge is applied to time-dependent data: the bias-variance trade-off remains and the geometry of its effects depends crucially on the structure of the regularization matrix Λ. I focus on how uniform shrinkage (the most common kind) can be easily replaced by a flexible but treatable regularizer structure, which I dub *lag-adapted*, that naturally fits the structure of VAR models. While shrinkage to zero per se can be counterproductive – in that it systematically pushes estimated coefficients to be small, which could cause issues especially when the process is near-to-unit-root – this needs not always be true. In fact, lag-adapted regularizers are able to improve estimation performance by purposely inducing coefficient decay mostly at deeper lags, reflecting the nature of the underlying model.

Second, I develop the asymptotic theory of the RLS estimator (possibly with non-zero centering) and show that it is a consistent and asymtotically normal estimator under mild assumptions on the regularization hyperparameter matrix Λ: these results complement the knowledge of asymptotic normality for ridge in ordinary regression contexts (Fu and Knight, 2000). Importantly, strengthened assumptions are required to ensure that RLS is also *asymptotically pivotal*, a necessary condition for inference unless one were to results to de-biasing, an approach that I do not explore here. I also generalize the asymptotic inference results on *sieve* VARs (Lütkepohl and Poskitt, 1991) to the ridge estimator. Lag-adapted ridge regularization is a scheme which lends itself naturally to addressing the weaknesses caused by the construction of large autoregressive models, like in the sieve setting.

Finally, using a Monte Carlo simulation exercise, I show that sieve RLS impulse responses and confidence intervals can be more effective in recovering impulse response functions and achieving nominal coverage, even in small sample settings.

Related Literature. The use of ridge regression is in fact common in the forecasting literature: Inoue and Kilian (2008) use ridge regularization for forecasting U.S. consumer price inflation and argue that it compares favorably with bagging techniques; De Mol et al. (2008) obtain the ridge estimators in the Bayesian context for the purposes of forecasting; Ghosh et al. (2019) again study the Bayesian ridge, this time however in the high-dimensional context; Coulombe (2020), Babii et al. (2020) and Medeiros et al. (2021) compare LASSO, ridge and other machine learning techniques for forecasting with large economic datasets. The ridge penalty is considered within a more general mixed ℓ_1-ℓ_2 penalization setting in Smeekes and Wijler (2018), who discuss the performance and robustness of penalized estimates for forecasting purposes. Most recently, Li et al. (2021) call for a general exploration of shrinkage procedures in the context of structural impulse response estimation.

From a theoretical perspective, Coulombe (2020) shows that the estimation problem of VARs with time-varying parameters can be effectively recast as ridge regression. In order to estimate smoothed impulse responses, Barnichon and Brownlees (2019) propose to smooth local projection impulse response functions using B-splines and solving a ridge-type problem. Plagborg-Møller (2016) similarly adopts a local projection smoothing approach again relying on ridge penalization. Poignard (2018) is closely related to the present work: it develops the asymptotic theory of mixed ℓ_1-ℓ_2 penalized extremum estimators with dependent data under a double asymptotic regime, the focus of which is the study of an adaptive Sparse Group Lasso estimator. While here I do not consider the case of processes of growing dimension, the results of Poignard (2018) encompass the ridge case, although not for the purpose of VAR model estimation.

Outline Section 2 provides a formal treatment of the ridge estimator for VAR models, its shrinkage and its asymptotic theory. Section 3 discusses the application of the ridge estimator to the sieve approximation of VAR(∞) models. It presents the main simulation experiment comparing the ridge VAR estimator with least squares and local projections. Section 5 concludes. Section 6 contains the proofs of all asymptotic results. Additional discussion, simulations and the proofs of results of Section 2.1 can be found in the Supplementary Appendix.

2 Ridge Regularized VAR Estimation

Notation. Let $y_t = (y_{1t}, \ldots, y_{Kt})'$ be a K-dimensional vector autoregressive process with lag length $p \geq 1$,

$$y_t = \nu(t) + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + u_t$$

where $u_t = (u_{1t}, \dots, u_{Kt})'$ is additive noise such that $u_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_u)$. I will assume that $\nu(t) = 0$ in the remainder of the paper, so that y_t, with minor loss of generality, has no deterministic time trend component. One might equivalently think of y_t as a de-trended series. The quantity $(I_K - \sum_{i=1}^{p} A_i z^i)$ for $z \in \mathbb{C}$ is the characteristic polynomial and determines the stability of the process. The order h autocovariance matrix is $\Gamma(h) = \mathbb{E}[y_t y'_{t-h}]$. Given a sample of length T, define $Y = (y_1, \dots, y_T)$, $Y_t = (y'_t, y'_{t-1}, \dots, y'_{t-p+1})'$, $B = (A_1, \dots, A_p)$, $Z = (Y_0, \dots, Y_{T-1})$, $U = (u_1, \dots, u_T)$, and vectorized counterparts $\mathbf{y} = \text{vec}(Y)$, $\boldsymbol{\beta} = \text{vec}(B)$ and $\mathbf{u} = \text{vec}(U)$. Accordingly, $Y = BZ + U$ and $\mathbf{y} = (Z' \otimes I_K)\boldsymbol{\beta} + \mathbf{u}$, where $\Sigma_{\mathbf{u}} = I_K \otimes \Sigma_u$ and $\otimes$ is the Kronecker product (Lütkepohl, 2005).

Ridge regularization is a modification of the standard least squares objective by the addition of term dependent on the Euclidean norm of the coefficient vector. The *ridge regularized least squares* (RLS) estimator is therefore defined as

$$\hat{\boldsymbol{\beta}}^R(\Lambda) := \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - (Z' \otimes I_K)\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

where $\lambda > 0$ is the regularization coefficient or regularizer. When $\lambda\|\boldsymbol{\beta}\|^2$ is replaced with the quadratic form $\boldsymbol{\beta}'\Lambda\boldsymbol{\beta}$ for Λ positive definite matrix, the above is often termed Tikhonov regularization. Tikhonov regularization allows for more flexible regularization schemes as discussed below. However, to avoid confusion, I shall refer to it as "ridge", since Λ will always assumed to be diagonal. In the remainder, let $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_{K^2p}\}$, $\lambda_i > 0$ for all $i = 1, \dots, K^2p$. I will discuss the practical choice of $\{\lambda_i\}$ in detail below. For the partial ordering of regularization matrices, I shall write $\Lambda_1 \prec \Lambda_2$ if $\lambda_{1,i} < \lambda_{2,i}$ for all $i = 1, \dots, K^2p$; $\Lambda_1 \preceq \Lambda_2$ if $\lambda_{1,i} \leq \lambda_{2,i}$ for all i and $\exists\ j \in 1, \dots, K^2p$ such that $\lambda_{1,j} < \lambda_{2,j}$.

By solving the normal equations, the RLS estimator with regularization matrix Λ $(K^2p \times K^2p)$ is proven to be

$$\hat{\boldsymbol{\beta}}^R(\Lambda) = (ZZ' \otimes I_K + \Lambda)^{-1}(Z \otimes I_K)\mathbf{y} \tag{1}$$

When a centering vector $\boldsymbol{\beta}_0 \neq 0$ is additionally considered in the penalty term given by $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\Lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, the RLS estimator becomes

$$\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) = (ZZ' \otimes I_K + \Lambda)^{-1}((Z \otimes I_K)\mathbf{y} + \Lambda\boldsymbol{\beta}_0) \tag{2}$$

Finally, $\hat{B}^R_{de}(\Lambda, \boldsymbol{\beta}_0)$ is the *de-vectorized* coefficient estimator obtained from reshaping $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$. Supplementary Appendix A.1 gives the detailed derivations of these estimators. In Supplementary Appendix A.2 it is shown how, contrary to the least-squares estimator, the RLS estimator induces different structures to the regularization matrix Λ according to whether it is constructed in vectorized form or matrix form. The *matrix* RLS estimator is thus indicated with $\hat{B}^R(\Lambda_{Kp}, B_0)$, where $\Lambda_{Kp} = \text{diag}\{\lambda_1, \dots, \lambda_{Kp}\}$.

For a vector v, $\|v\|$ is intended as the Euclidean norm. For a matrix A, $\|A\|$ is intended as the spectral norm unless stated otherwise, and $\|A\|_F = \text{tr}\{A'A\}$ is the Frobenius norm.

The symbols $\xrightarrow{p}$ and $\xrightarrow{d}$ are used to indicate convergence in probability and convergence in distribution, respectively.

Assumptions. To study the estimator $\hat{\beta}^R(\Lambda)$ and its siblings I will make a number of assumptions on process y_t.

Assumption A. $\{u_t\}_{t=1}^T$ is a sequence of i.i.d. random variables with $\mathbb{E}[u_t] = 0$, $\mathbb{E}[u_t u_t'] = \Sigma_u$ being a nonsingular positive definite matrix, and $\mathbb{E}\,|u_{it}u_{jt}u_{mt}u_{nt}| \leq \kappa_4 < \infty \; i,j,m,n = 1,\dots,K$.

An i.i.d. noise assumption is standard and allows to prove the main asymptotic results with standard theoretical devices. Assuming u_t is white noise or assuming y_t respects strong mixing conditions (Davidson, 1994) would require more careful consideration in asymptotic arguments, see e.g. Boubacar Mainassara and Francq (2011). It would significantly complicate proofs, especially those related to sieve estimation. Further, it is well known from Andrews (1984, 1985) that strong mixing is not as radical a departure from independence as one would often perceive. Therefore, I consider the use of more general dependence assumptions beyond the scope of the following discussion.

The generalization of y_t to a process of *increasing* dimension K is clearly of high interest. Such case needs not encompass (ultra) high-dimensional settings directly, but is more pertinent to the uses of ridge where the number of regressors grows together with sample size. These are most common in i.i.d. regression settings but are becoming more prevalent in time series analysis, too.

Assumption B. There exists $\rho > c > 1$ such that $\det(I_K - \sum_{i=1}^p A_i z^i) \neq 0 \; \forall z \in \mathbb{C}$, $|z| \leq \rho$.

The process y_t is assumed to be strictly stable. A number of highly interesting situations do not satisfy this assumption, the most significant ones being unit roots, cointegrated VARs, and local-to-unity settings. The analysis of unit roots in VAR models is a key area of interest, because macroeconomic time series are well known to exhibit high persistence (Kilian and Lütkepohl, 2017). As I argue below, the ridge estimator naturally produces shrinkage. Hence, the well-known issues of small sample bias in VAR estimation (Hamilton, 1994) may be exacerbated. In particular, one may fear that, by adding a penalty, ridge estimates of highly persistent processes are even less reliable than those of LS/ML estimators. Incorrect identification of unit roots does not invalidate the use of LS or Maximum Likelihood estimators (Phillips, 1988, Park and Phillips, 1988, 1989, Sims et al., 1990), however inference could be significantly impacted as a result (Pesavento and Rossi, 2006, Mikusheva, 2007, 2012). Assumption B guarantees that y_t has no unit roots and, thus, ridge shrinkage is not biasing their estimation.

Assumption C. There exist $0 < \boldsymbol{m} \leq \boldsymbol{M} < \infty$ such that

$$\boldsymbol{m} \leq \omega_{\min}(\Gamma) \leq \omega_{\max}(\Gamma) \leq \boldsymbol{M}$$

where $\Gamma \equiv \Gamma(0)$ is the autocovariance matrix of y_t and $\omega_{\min}(\Gamma)$, $\omega_{\max}(\Gamma)$ are its smallest and largest eigenvalues respectively.

The assumption of a positive definite autocovariance matrix is standard in the literature regarding penalized estimation, and does not imposes significant additional constraints on the process y_t, cf. Assumption A. It is sufficient to ensure that for large enough T the plug-in sample autocovariance estimator is invertible even under vanishing Λ.

Assumption D. The regularization matrix Λ satisfies either:

(i) $\sqrt{T}^{-1}\Lambda \xrightarrow{p} 0$

(ii) $\sqrt{T}^{-1}\Lambda \xrightarrow{p} \Lambda_0$, where Λ_0 is positive definite.

(iii) $T^{-1}\Lambda \xrightarrow{p} 0$

(iv) $T^{-1}\Lambda \xrightarrow{p} \overline{\Lambda}_0$, where $\overline{\Lambda}_0$ is non-negative definite.

The options in Assumption D establish a number of rates at which the regularizer Λ can grow with the sample size. As I show below, such conditions are required for the penalty term not to overwhelm either Law of Large Number or Central Limit Theorem asymptotic rates, which would otherwise lead to singular (zero) estimates. The seminal paper Fu and Knight (2000) proves that Assumption D.(ii) is sufficient to produce a conventional, albeit generally complex, CLT results for the class of shrinkage estimators induced via norm penalty. This includes the ridge estimator, which is shown to have an asymptotically nonpivotal normal distribution, as its mean depends on the unknown model parameters. Under the more conservative assumption (i), this weakness is resolved. The main conclusion of this paper is that these results carry over to the time series context. On a conceptual level, it is hard to argue which setting would require an *increasing* rate of regularization, as intuition suggests that violations of Assumption D would likely be pathological. Supplementary Appendix A.3 shows that under the assumption that Γ is positive definite, the ordered eigenvalues $\hat{\omega}_i$ of a consistent estimator $\hat{\Gamma}_T$ converge in probability to $\omega_i(\Gamma) > 0$. Accordingly, ridge, intended as a *regularization* estimator, would demand $\Lambda \to 0$ as $T \to \infty$.

2.1 Ridge Shrinkage

A fundamental property of ridge regularized estimates is that they exhibit shrinkage with respect to least squares counterparts. The function

$$\mathcal{L}(\boldsymbol{\beta}; \lambda) = \|\boldsymbol{y} - (Z' \otimes I_K)\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

understood as a Lagrangian implies that the implicit constraint $\|\boldsymbol{\beta}\| = c$ is added to the least squares loss. Therefore, $\lambda > 0$ naturally leads to smaller estimates for $\hat{\boldsymbol{\beta}}^R$. When $\lambda\|\boldsymbol{\beta}\|$ is replaced by the penalty $\boldsymbol{\beta}\Lambda\boldsymbol{\beta}$, the relative shrinkage of different coefficients is more complex, so below I shall assume some additional structure on Λ.

In standard regression contexts, setting $\lambda > 0$ reduces estimator variance through the bias-variance trade-off (Hoerl and Kennard, 1970, van Wieringen, 2020), that is, the difference[2] $\mathrm{Var}[\hat{\boldsymbol{\beta}}^{LS}] - \mathrm{Var}[\hat{\boldsymbol{\beta}}^R(\lambda)]$ is a positive definite matrix. A first intuitive result is that at any fixed sample size T this property carries over to autoregressive ridge estimators.

Proposition 1. *Let Assumptions A-B hold. For a given T, if $\Lambda_1 \preceq \Lambda_2$ then*

$$\mathbb{P}(\|\hat{\boldsymbol{\beta}}^R(\Lambda_2)\| > \epsilon) < \mathbb{P}(\|\hat{\boldsymbol{\beta}}^R(\Lambda_1)\| > \epsilon) \qquad \forall \epsilon > 0$$

Proof. See Supplementary Appendix B.1.

Note that the above result is always applicable to uniform scaling matrices, that is $\Lambda = \lambda I_{K^2 p}$ for $\lambda > 0$. Proposition 1 is weaker than similar results in linear regression scenarios, because it yields no positive definite ranking of variance matrices, cf. Hoerl and Kennard (1970). This is primarily due to the complexity of recovering an explicit formula of the RLS estimators variance, an issue closely related to studying $\mathrm{Var}[\hat{\boldsymbol{\beta}}^{LS}]$. Computing the analytical moments of $\hat{\boldsymbol{\beta}}^{LS}$ in vector autoregressive setups remains mostly an open problem (Sawa, 1978, Nankervis and Savin, 1988, Bao, 2007), and I do not attempt to address it here.

From Proposition 1 one gathers that, for any given sample of length T, $\hat{\boldsymbol{\beta}}^R$ asymptotically concentrates its estimates in a neighborhood of the origin as $\lambda_i \to \infty$ across i,

$$\lim_{\Lambda \to \infty} \mathbb{P}(\|\hat{\boldsymbol{\beta}}^R(\Lambda)\| > \epsilon) = 0 \qquad \forall \epsilon > 0$$

and as such its variance is also asymptotically zero. For a ridge estimator centered at $\boldsymbol{\beta}_0$ the asymptotic neighborhood of concentration is located around $\boldsymbol{\beta}_0$. The proof of such a result is an immediate generalization of Proposition 1 so I omit it.

[2] By definition of $\hat{\boldsymbol{\beta}}^R$ it holds that $\hat{\boldsymbol{\beta}}^{LS} \equiv \hat{\boldsymbol{\beta}}^R(0)$.

Lag-Adapted Λ. I will now consider a special structure for Λ that is of interest when applying ridge specifically to a VAR model.[3] A *lag-adapted* regularization matrix is defined as

$$\Lambda^{\ell} = \mathrm{diag}\{\lambda_1, \dots, \lambda_p\} \otimes I_{K^2}$$

where each λ_i implies a different penalty for the coefficients of each coefficient matrix A_i, $i = 1, \dots, p$. The role of this specific structure for the regularizer will become clear below, and it will turn out to be especially important in the sieve application.

A question of interest is how the size of RLS estimates depends on the size of different regularizers. Choosing to study the family of lag-adapted matrices $\mathcal{F}^{\ell} = \{\mathrm{diag}\{\lambda_1, \dots, \lambda_p\} \otimes I_{K^2} \,|\, \lambda_i \in \mathbb{R}^+\}$ allows to obtain direct theoretical guarantees on the relative effects of larger ridge penalties.

Proposition 2. *For a subset $\mathcal{S} \subseteq \{1, \dots, p\}$ of cardinality $s = |\mathcal{S}|$, for $\Lambda^{\ell} \in \mathcal{F}^{\ell}$ define $\hat{\boldsymbol{\beta}}^R(\Lambda^{\ell})_{[\mathcal{S}]}$ as the vector of sK^2 coefficient estimates located at indexes $1 + K^2(j-1), \dots, K^2 j$ for $j \in \mathcal{S}$ (where j, $\mathcal{S}$ are ordered without loss of generality).*

(i) If $\Lambda^{\ell}_1 \preceq \Lambda^{\ell}_2$, then $\|\hat{\boldsymbol{\beta}}^R(\Lambda^{\ell}_2)_{[\mathcal{S}]}\| < \|\hat{\boldsymbol{\beta}}^R(\Lambda^{\ell}_1)_{[\mathcal{S}]}\|$ for any $\mathcal{S}$.

(ii) If $\lambda_1 < \lambda_2$, then $\|\hat{\boldsymbol{\beta}}^R(\lambda_2 \otimes I_{K^2 p})_{[\mathcal{U}]}\| < \|\hat{\boldsymbol{\beta}}^R(\lambda_1 \otimes I_{K^2 p})_{[\mathcal{U}]}\|$ for any $\mathcal{U} \subset \{1, \dots, K^2 p\}$.

(iii) Let $\hat{\boldsymbol{\beta}}^{LS}(\mathcal{S})$ be the least squares estimator of the autoregressive model with only the lags indexed by $\mathcal{S}$ included and $\mathcal{S}^c = \{1, \dots, p\} \setminus \mathcal{S}$. Then

$$\lim_{\substack{\Lambda^{\ell}_{[\mathcal{S}]} \to 0 \\ \Lambda^{\ell}_{[\mathcal{S}^c]} \to \infty}} \hat{\boldsymbol{\beta}}^R(\Lambda^{\ell}) = \hat{\boldsymbol{\beta}}^{LS}(\mathcal{S})$$

Proof. See Supplementary Appendix B.1.

By (1), it is clear that it is not possible to penalize each A_i independently.[4] However, Proposition 2 also shows that the limiting geometry of a lag-adapted ridge estimator is indeed identical to that of a specific least squares estimator. Intuitively, one may then think that by controlling the sizes of coefficients $\{\lambda_1, \dots, \lambda_p\}$ it is possible to achieve pseudo-selection of a

[3] The idea of lag-adapted regularization can be readily reinterpreted through the context of Bayesian methods as a sibling of the Minnesota prior (Litterman, 1985). Further, the concept of non-uniform shrinking over lags has already been put forward in the literature, see for example De Mol et al. (2008).

[4] The geometric intuition is that when setting $\Lambda^{\ell} = \mathrm{diag}\{1, 0, \dots, 0\} \otimes I_{K^2}$ the ridge estimator minimizes the Lagrangian with implicit constraint $\|\boldsymbol{\beta}_{1:K^2}\| = c$, where coefficients $\boldsymbol{\beta}_{1:K^2}$ map to the VAR matrix A_1. The direction of shrinkage is along the eigenvectors of $ZZ' \otimes I_K$, thus in general along all lags of y_t. It is however possible to achieve *component-wise* independent shrinkage by choosing an appropriate λ_k, $k = 1, \dots, K$ and applying the RLS estimator component-wise. While for least squares this approach is equivalent to that of the vectorized estimator, that is not generally true with the ridge estimator. It is outside the scope of this paper to discuss such a special case.

specific model. This behavior turns out to be fundamental in practice, because – at least in simulations – the data-driven selection of Λ^ℓ yields regularizers which, in some sense, "control" model complexity. These results are presented in more detail in Section 4.1.

Shrinkage Bias. Supplementary Appendix C contains additional simulation evidence aimed at better understanding the practical implications of the shrinkage bias described above. By considering both a two-dimensional VAR(1) model and a flexible VAR(1) specification with up to 100 series, I compare the bias behavior of the RLS estimator with isotropic regularizer $\Lambda = \lambda I_{K^2 p}$ between a number of models of different size and persistence.[5] In the small model the RLS estimator can in fact achieve a *smaller* bias than least-squares (although λ must be chosen by optimizing an unfeasible oracle loss). Unfortunately, with larger and more realistic models the same bias simulations show that this "advantage" disappears, even without very strong dependency. The conclusion is that in applications an isotropic ridge penalty is likely to produce estimates that are more biased than least squares. This clearly needs not be true for more general penalization matrices, yet in practice one should be attentive and weight how much bias one is willing to accept.

Ridge shrinkage effects are extremely relevant from both the theoretical and applied point of view. It is well known that a common issue in applications is the small sample bias of the least squares VAR estimator: the bias pushes estimates downward in univariate models (Marriott and Pope, 1954), but becomes significantly more complex to asses in multivariate settings (Nicholls and Pope, 1988, Pope, 1990, Engsted and Pedersen, 2014). The use of bias correction procedures (Kilian, 1998, Kilian and Lütkepohl, 2017) is often recommended to mitigate the problem, which can have a substantial impact, especially in terms of inference. The bias may have extreme effects on nonlinear transformations of parameters, like impulse responses, which most often are the true object of interest. As such, the use of the RLS estimator with non-trivial shrinkage ($0 \preceq \Lambda$) over the LS/ML estimators is hard to prescribe when the researcher knows or suspects that at least some of the components in a vector time series are strongly persistent.

2.2 Regularizer Selection

Given that ridge regularization is an established technique in the context of regression, the problem of empirically choosing the isotropic regularizer $\lambda I_{K^2 p}$ is well studied, with many competing approaches (Bauer and Lukas, 2011). Generalized cross-validation – often simply referred to as cross-validation (CV) – is one of the most commonly applied schemes and is straightforward to implement; it has been shown to be reliable even under dependence (Bauer and Lukas, 2011); and it can be easily tuned for robustness (Lukas, 2006, 2008, 2010). A common variation of CV

[5] In the 2-dimensional VAR(1) simulation I use the same DGP utilized in the well-known bias simulation study performed by Kilian (1998).

is given by *k-fold* cross validation, wherein the data is evenly split into k subsets called *folds* (Hastie et al., 2009). Often, the sample is also shuffled before constructing the set of k-folds, to avoid strictly contiguous sample subsets. Bergmeir et al. (2018) showed that for stationary AR(p) models, standard k-fold cross-validation lead to a valid estimation of prediction error so long as errors exhibit no correlation. An even more sophisticated variation is *non-dependent* CV (Bergmeir et al., 2018), but due to the extremely data-intensive nature of non-dependent CV however I shall not consider it here.[6] The naive alternative to CV is (quasi) out-of-sample (OOS) evaluation, which splits the data into an initial estimation subsample (periods $t = 0, \ldots, T_0 < T$) and a later testing subsample (periods $t = T_0 + p, \ldots, T$). Both CV and OOS evaluation have been applied in the recent literature on time series ridge applications. Relevant examples of k-fold CV applied to ridge estimators are Coulombe et al. (2020) and Coulombe (2020), although their use of ridge regression is not aimed at model estimation for the purpose of inference.

In this paper, I consider both cross-validation and out-of-sample validation as techniques to set λ in a data-driven manner. These methods are applied to the special case of lag-adapted regularizers, which give space for a more articulated regularization structure. Importantly, the numerical implementation of both out-of-sample validation and cross-validation will rely on solving *constrained* optimization problems, so that in simulations Λ has a finite (albeit large) upper bound. The details regarding this implementation can be found in Appendix 6.3.

2.3 Asymptotic Theory

It is now possible to state the main asymptotic result of the paper for the RLS estimator $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ with general regularization matrix Λ. I shall allow Λ and non-zero centering coefficient $\boldsymbol{\beta}_0$ to be, under appropriate assumptions, r.v.s dependent on sample size T. In particular, $\boldsymbol{\beta}_0$ may be a consistent estimator of $\boldsymbol{\beta}$.

Theorem 1. *Let $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ be the centered RLS estimator as in (2). Define $\hat{\Gamma}_T = T^{-1}ZZ'$, the residuals $\hat{U} = Y - \hat{B}^R_{de}Z$ and*

$$\hat{\Sigma}^R_T = T^{-1}\hat{U}\hat{U}'$$

Under Assumptions A-C, Assumption D.(ii), and assuming $\boldsymbol{\beta}_0 \xrightarrow{p} \underline{\boldsymbol{\beta}}_0$,

(a) $\hat{\Gamma}_T \xrightarrow{p} \Gamma$

(b) $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) \xrightarrow{p} \boldsymbol{\beta}$

(c) $\hat{\Sigma}^R_T \xrightarrow{p} \Sigma_u$

(d) $\sqrt{T}(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\Gamma^{-1}\Lambda_0(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}), \Gamma^{-1} \otimes \Sigma_u)$

[6] In fact, an issue of interest is how cross-validation and alternatives perform under small samples and relatively many VAR components, a setting for which non-dependent CV is clearly unsuitable.

Proof. See Section 6.1.

The above Theorem considers the most general case, and, as previously mentioned, gives the asymptotic distribution of $\hat{\boldsymbol{\beta}}^R$ under rather weak conditions for the regularizer Λ: the resulting normal limit is then clearly not pivotal in general, precluding inference. In fact, the Gaussian limit distribution depends not only on the unknown DGP coefficient $\boldsymbol{\beta}$ and autocovariance Γ, but on the asymptotic centering parameter $\underline{\boldsymbol{\beta}}_0$.

It is however possible, under strengthened assumptions for Λ or $\boldsymbol{\beta}_0$, that $\hat{\boldsymbol{\beta}}^R$ have a pivotal limiting distribution.

Theorem 2. *In the setting of Theorem 1, consider the following alternative assumptions:*

(1) Replace Assumption D.(ii) with Assumption D.(i).

(2) Replace Assumption D.(ii) with Assumption D.(iii), assume that $\boldsymbol{\beta_0} \xrightarrow{p} \boldsymbol{\beta}$ *and* $\sqrt{T}^{-1}(\boldsymbol{\beta_0} - \boldsymbol{\beta}) = O_p(1)$.

Then, results (a)-(c) hold and

(d′) $\sqrt{T}(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma_u)$

Proof. See Section 6.1.

The following corollary is immediate.

Corollary 1. *Let* $\hat{\beta}_0$ *be a consistent and asymptotically normal estimator of* $\boldsymbol{\beta}$*. Then, under assumptions (1) or (2) of Theorem 2 results (a)-(d′) hold.*

Asymptotic Shrinkage. Theorems 1 and 2 portray the ridge VAR estimator in a somewhat unfair light, because they result in asymptotic distributions that show no bias-variance tradeoff. In this sense, they can be disappointing: it would seem that ridge shrinkage yields no asymptotically smaller variance than least squares in the limit. This was already highlighted by Fu and Knight (2000). Of course, when these results are applied in finite samples shrinkage has an effect on $\Gamma^{-1} \otimes \Sigma_u$ because $\hat{\Sigma}_T^R$ is used to estimate the error term variance matrix.

To better understand why ridge can be valuable in practice, one should therefore consider the situation wherein a number of VAR coefficients are small (but not necessarily zero). Formally, for some $0 < n \leq p$ let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ where $\boldsymbol{\beta}_1 \in \mathbb{R}^{K^2(p-n)}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{K^2 n}$, and assume that $\boldsymbol{\beta}_2 \propto T^{-(1/2+\delta)}\, \boldsymbol{b}_2$ for $\delta > 0$. Notice that such ordered partitioning of $\boldsymbol{\beta}$ is without loss of generality. Also, the dimensions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are chosen to be multiples of K^2 to better conform to the lag-adapted setting I discuss below for the sieve framework; this is also without loss of generality, and simplifies intuition.

In this setting the subset of coefficients in $\boldsymbol{\beta}_2$ are clearly "small" with respect to sample size. This situation happens, for example, when one considers VAR(∞) process obtained by inverting VARMA(p,q) processes of finite AR and MA order: for $i \gg p$ coefficient matrices A_i show geometric decay and thus *in sample* they are "small" in the sense of $\boldsymbol{\beta}_2$. It is possible to penalize $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ differently when regularizing using the ridge penalty. Indeed, let $\Lambda = \text{diag}\{(L_1', L_2')'\} \otimes I_K$ where $L_1 \in \mathbb{R}_+^{K^2(p-n)}$ and $L_2 \in \mathbb{R}_+^{K^2 n}$. Assume that $\sqrt{T}^{-1} L_1 \xrightarrow{p} 0$ and $T^{-1} L_2 \xrightarrow{p} \overline{L}_2$ as $T \to \infty$. In particular,

$$\frac{\Lambda}{T} = \begin{bmatrix} T^{-1}\Lambda_1 & 0 \\ 0 & T^{-1}\Lambda_2 \end{bmatrix} \otimes I_K \xrightarrow{p} \begin{bmatrix} 0 & 0 \\ 0 & \overline{\Lambda}_2 \end{bmatrix} \otimes I_K = \overline{\Lambda}_{1:2} \otimes I_K, \qquad \overline{\Lambda}_2 \succ 0 \tag{3}$$

Because coefficients in $\boldsymbol{\beta}_2$ are small, it is possible to penalize them "strongly", that is, in a way that is not asymptotically negligible. This means one is exploiting some additional information on the importance of a subset of $\boldsymbol{\beta}$ to improve on the estimation efficiency of $\hat{\boldsymbol{\beta}}^R$.

It is thus possible to state a result explicitly showing the effect of shrinkage on the asymptotic distribution of the ridge estimator. For simplicity, I assume that $\hat{\boldsymbol{\beta}}^R$ features no centering, that is $\boldsymbol{\beta}_0 = 0$.

Theorem 3. *In the setting of Theorem 1, assume that, for $0 < n \leq p$,*

(1) $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ where $\boldsymbol{\beta}_1 \in \mathbb{R}^{K^2(p-n)}$ and $\boldsymbol{\beta}_2 \propto T^{-(1/2+\delta)}\, \boldsymbol{b}_2$ for $\delta > 0$, $\boldsymbol{b}_2 \in \mathbb{R}^{K^2 n}$.

(2) $\Lambda = \text{diag}\{(L_1', L_2')'\}$ where $L_1 \in \mathbb{R}_+^{K^2(p-n)}$ and $L_2 \in \mathbb{R}_+^{K^2 n}$.

(3) $\sqrt{T}^{-1} L_1 \xrightarrow{p} 0$ and $T^{-1} L_2 \xrightarrow{p} \overline{L}_2$ as $T \to \infty$..

(4) $\boldsymbol{\beta}_0 = 0$.

Define $\Gamma_{\overline{\Lambda}} = \Gamma + \overline{\Lambda}_{1:2}$ where $\overline{\Lambda}_{1:2} \succeq 0$ is defined as in (3). Then, results (a)-(c) hold and

(d″) $\sqrt{T}(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(0, \Gamma_{\overline{\Lambda}}^{-1}\, \Gamma\, \Gamma_{\overline{\Lambda}}^{-1} \otimes \Sigma_u\right)$

Proof. See Section 6.1.

It is easy to see that indeed the term $\Gamma_{\overline{\Lambda}}^{-1}\, \Gamma\, \Gamma_{\overline{\Lambda}}^{-1}$ in Theorem 3 is weakly smaller in the positive-definite sense than Γ^{-1}. Note that

$$\begin{aligned} \Gamma_{\overline{\Lambda}}^{-1}\, \Gamma\, \Gamma_{\overline{\Lambda}}^{-1} \preceq \Gamma^{-1} \quad & \Longleftrightarrow \quad (\Gamma + \overline{\Lambda}_{1:2})^{-1}\Gamma \preceq \Gamma^{-1}(\Gamma + \overline{\Lambda}_{1:2}) \\ & \Longleftrightarrow \quad I_{K^2p} - (\Gamma + \overline{\Lambda}_{1:2})^{-1}\overline{\Lambda}_{1:2} \preceq I_{K^2p} + \Gamma^{-1}\overline{\Lambda}_{1:2} \\ & \Longleftrightarrow \quad 0 \preceq ((\Gamma + \overline{\Lambda}_{1:2})^{-1} + \Gamma^{-1})\overline{\Lambda}_{1:2} \end{aligned}$$

The last inequality is true by definition of $\overline{\Lambda}_{1:2}$. This expression also makes clear that the shrinkage gains are "concentrated" at the components that have non-zero asymptotic shrinkage, i.e. those effected by $\overline{\Lambda}_{1:2}$.

Joint Inference. The theorems establishing the asymptotic normality of the RLS estimator also allow to prove joint limit results for both $\hat{\boldsymbol{\beta}}^R$ and the variance estimator $\hat{\Sigma}_T^R$, which are of fundamental interest for inference purposes requiring smooth transformations of the VAR coefficients, such as impulse responses (Lütkepohl, 1990).

Theorem 4. *Let $\hat{\boldsymbol{\sigma}}^R = vec(\hat{\Sigma}_T^R)$ and $\boldsymbol{\sigma} = vec(\Sigma_u)$. Under the assumptions of Theorem 1,*

$$\sqrt{T}\begin{bmatrix}\hat{\boldsymbol{\beta}}^R - \boldsymbol{\beta} \\ \hat{\boldsymbol{\sigma}}^R - \boldsymbol{\sigma}\end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix}\Gamma^{-1}\Lambda_0(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) \\ 0\end{bmatrix}, \begin{bmatrix}\Gamma^{-1} \otimes \Sigma_u & 0 \\ 0 & \Omega\end{bmatrix}\right)$$

Under assumption (1) or (2) of Theorem 2,

$$\sqrt{T}\begin{bmatrix}\hat{\boldsymbol{\beta}}^R - \boldsymbol{\beta} \\ \hat{\boldsymbol{\sigma}}^R - \boldsymbol{\sigma}\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix}\Gamma^{-1} \otimes \Sigma_u & 0 \\ 0 & \Omega\end{bmatrix}\right)$$

where $\Omega = \mathbb{E}\left[vec(u_t u_t')\, vec(u_t u_t')'\right] - \boldsymbol{\sigma}\boldsymbol{\sigma}'$.

Proof. See Section 6.1.

3 Sieve Approximation

It is a common assumption in autoregressive modeling that the data-generating process has finite lag order, and this order is either known or can be estimated by some criterion, often AIC or BIC. There is however an alternative. One might postulate the DGP to be a VAR(∞), so that any VAR(p) model fitted to sample realizations from said process can only yield a finite-order approximation. Because the true lag order is infinite, $p \to \infty$ with the sample size. This approach is commonly referred to as *sieve approximation*. The sieve view introduces significant changes to asymptotic analysis. Derivation of consistency and asymptotic normality results for the least squares estimator under the sieve framework goes back to Lewis and Reinsel (1985) and was further refined for inference in Lütkepohl (1990), Lütkepohl and Poskitt (1991).

A ridge-regularized estimator lends itself effectively to this setting. In the Bayesian context, it seems natural that a model with higher lag order should be tooled with priors penalizing deeper lags more than shallower lags, since often one considers VAR(∞) DGP with strong assumptions on the rates of decay of coefficients[7], as discussed by De Mol et al. (2008). The same idea can be applied to frequentist estimation by means of ridge regularization. In fact, the class of lag-adapted regularization matrices Λ^ℓ can yield exactly such a procedure, since they are parameterized by the a set $\{\lambda_1, \ldots, \lambda_p\}$ of coefficients, each λ_i penalizing matrix A_i, respectively. Indeed, it would simply amount to setting $0 \leq \lambda_1 < \ldots < \lambda_p$.

[7] The VAR(∞) representation of an invertible MA(1) process, where the autoregressive coefficient matrices have an exponential decay in p, is the easiest example of such decay.

3.1 Sieve Asymptotic Theory

Assume that y_t is a stable VAR(∞) process $y_t = \sum_{i=0}^{\infty} A_i y_{t-i} + u_t$ with Wold representation

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \qquad \sum_{i=0}^{\infty} i^{1/2} \|\Phi_i\|_F < \infty$$

where the transfer function $\Phi(z) = \sum_{i=0}^{\infty} \Phi_i z^i$ has no roots inside the complex circle $\|z\| \leq \rho$ for $\rho > 1$ (Lütkepohl and Poskitt, 1991). A finite lag order VAR(p) model, $y_t = \sum_{i=1}^{p} A_i y_{t-i} + u_t$, is then fit to the infinite order data generating process, and estimators $\hat{A}_i^{LS}$ and $\hat{A}_i^{R}(\Lambda)$ are obtained by least squares and RLS, respectively.

Sieve LS and sieve RLS estimators require different conditions to remain consistent and asymptotically normal than those previously discussed. Lewis and Reinsel (1985) derive appropriate rates for p to vary with T such that usual asymptotic properties of least squares hold. I consider their original setting, and I adapt their argument to show that additional convergence rates must be assumed for the lag-adapted regularizer Λ^{ℓ}. I focus on the lag-adapted RLS estimator because it has a number of benefits compared to more general formulations. Its structure is, as argued above, well suited to the sieve approach. It is also a good compromise between degrees-of-freedom and parsimony in choosing Λ, especially in applications. Thirdly, it simplifies analytical formulas significantly, allowing a precise theoretical treatment.

The lag-adapted structure enables the direct use of the matrix ridge estimator[8] $\hat{B}^R$ in place of the general (vectorized) form $\hat{\boldsymbol{\beta}}^R$, therefore below I use $\hat{B}^R$ to streamline notation.

Theorem 5. *Let $\hat{B}_p^R(\Lambda^{\ell})$ be the matrix ridge regularized estimator with lag-adapted regularizer of lag order p. Under Assumptions A-C, and*

(i) p is such that $p^2/T \to 0$ as $T \to \infty$

(ii) p is such that $p^{1/2} \sum_{j=p+1}^{\infty} \|A_j\|_F \to 0$ as $p, T \to \infty$

(iii) Λ^{ℓ} is such that $(T-p)^{-1} \|\Lambda^{\ell}\| \to 0$ as $p, T \to \infty$

Then

$$\|\hat{B}_p^R(\Lambda^{\ell}) - B_p\| \xrightarrow{p} 0$$

Proof. See Section 6.2.

Theorem 6. *Let $\hat{B}_p^R(\Lambda^{\ell})$ be the matrix ridge regularized estimator with lag-adapted regularizer of lag order p. Under Assumptions A-C, and*

(i) p is such that $p^3/T \to 0$ as $T \to \infty$

(ii) p is such that $T^{1/2} \sum_{j=p+1}^{\infty} \|A_j\|_F \to 0$ as $p, T \to \infty$

[8] This equivalence follows from the discussion in Supplementary Appendix A.2.

(iii) Λ^ℓ *is such that* $(T-p)^{-1/2}\|\Lambda^\ell\| \to 0$ *as* $p, T \to \infty$

Then

$$(T-p)^{1/2}(\hat{\boldsymbol{\beta}}_p^R(\Lambda^\ell) - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, V_p^{-1} \otimes \Sigma_u)$$

where $\hat{\boldsymbol{\beta}}_p^R(\Lambda^\ell) = vec(\hat{B}_p^R(\Lambda^\ell))$ *and* V_p *is the upper-left* $(Kp \times Kp)$ *block of infinite-dimensional matrix* $[\Gamma(m-n)]_{m,n=1,2,\dots} = \Gamma_\infty$.

Proof. See Section 6.2.

Sieve Inference. A most interesting application for the sieve RLS estimator is its use for inference purposes. It is important to note that the results of Theorem 4 are not sufficient in this context, since it assumes p to be finite and known. Fortunately, the theory of inference for sieve estimation is well studied. Chapter 15, Lütkepohl (2005) gives a thorough overview, and the theory within can be readily adapted to the case of ridge regression given Theorems 5 and 6. Lütkepohl and Poskitt (1991) provides the core theoretical derivation of the limiting distribution of IRFs by relying on the results of Lewis and Reinsel (1985). Given the proof of Theorem 6, it is therefore sufficient to generalize Theorem 1 of Lütkepohl and Poskitt (1991) as to encompass the use of $\hat{B}_p^R(\Lambda^\ell)$ in place of the usual finite-order least squares estimator. In fact, it suffices to show that Lemma 2, Lütkepohl and Poskitt (1991), p. 493 holds if least squares sieve residuals are replaced with sieve RLS residuals under assumptions (i)-(iii) of Theorem 6. I give a proof of such result in Lemma 2, Appendix 6.2.

The MA coefficient matrices obtained from sieve AR estimates are then asymptotically normally distributed (cf. Proposition 15.4, Lütkepohl (2005)):

Proposition 3. *Under the assumptions of Theorem 6, let* $\hat{\Phi}_i^R$ *be the MA coefficient estimates obtained recursively as*

$$\hat{\Phi}_i^R(p) = \sum_{j=1}^{i} \hat{\Phi}_{j-i}^R(p)\hat{A}_j^R(\Lambda^\ell)$$

where $\hat{A}_j^R(\Lambda^\ell)$ *is the lag* j *submatrix of coefficients estimates in* $\hat{B}_p^R(\Lambda^\ell)$ *if* $i \leq p$ *and* $\hat{A}_j^R(\Lambda^\ell) := 0$ *if* $j > p$*. Then,*

$$\sqrt{T} vec(\hat{\Phi}_i^R(p) - \Phi_i) \xrightarrow{p} \mathcal{N}\left(0, \Sigma_u^{-1} \otimes \sum_{j=0}^{i-1} \Phi_j \Sigma_u \Phi_j'\right)$$

Proof. See Section 6.2.

While the above proposition discusses explicitly only non-orthogonalized MA coefficient estimates, the pertinent results of Lütkepohl and Poskitt (1991) apply here too, since the sieve estimator of Σ_u is consistent, as discussed in Appendix 6.2. It is therefore possible to directly apply textbook results for inference on structural impulse responses to sieve ridge estimates, with the only additional cost being that assumption (iii) of Theorem 5 and Theorem 6 must hold.

4 Impulse Response Inference

4.1 Monte Carlo Simulation

To study the quality of ridge-regularized estimator I perform a simulation exercise that focuses in inference on dynamic multipliers. In order to have a fair comparison between methods, I use a VAR(∞) data-generating process: in order to make such specification feasible to simulate numerical, a VARMA$(1, 1)$ process is used. I rely on the specification of Kilian and Kim (2011) since it has already been extensively used to gauge the performance of both finite-order and sieve methods in the literature.

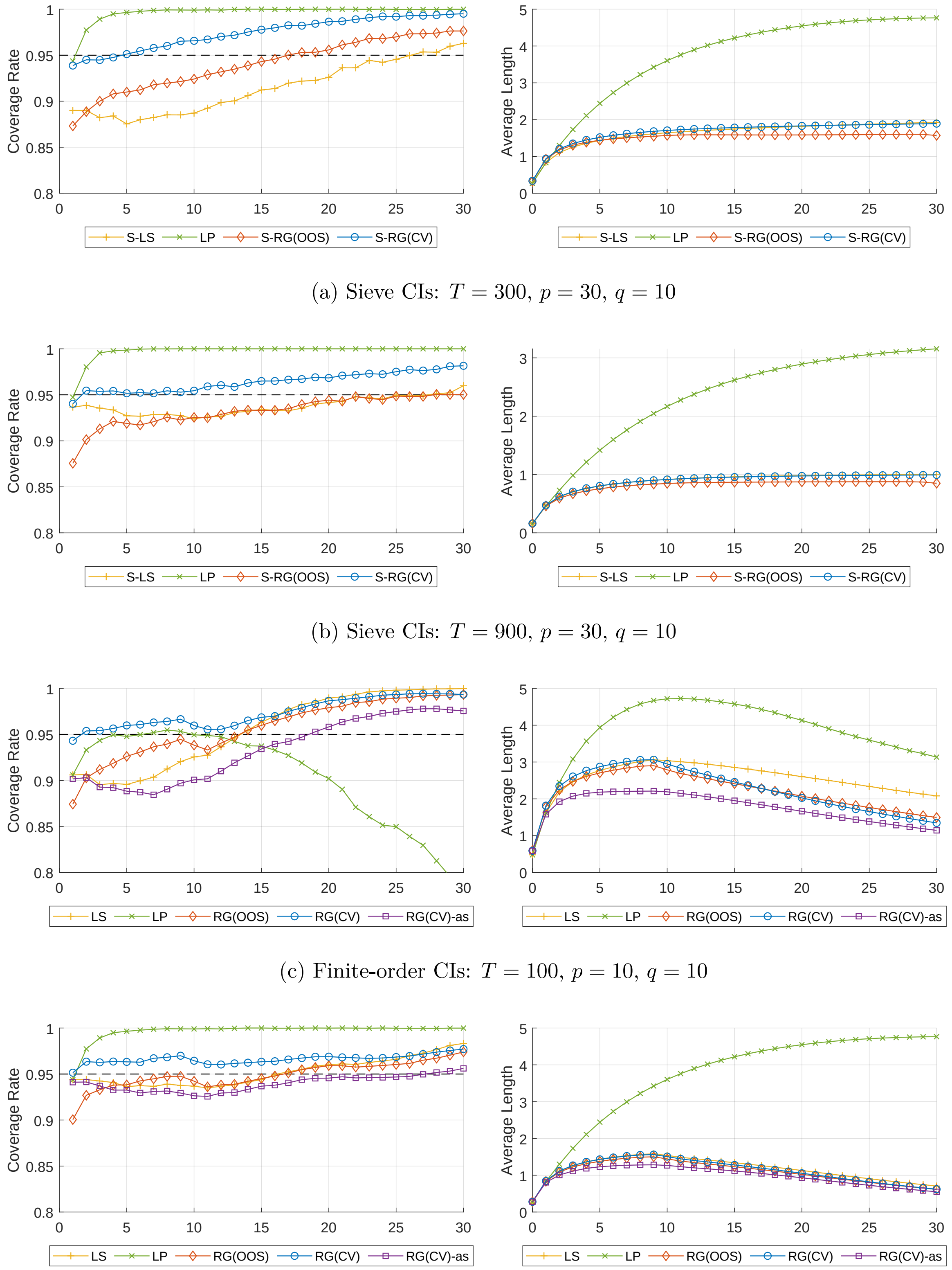
This Monte Carlo exercise covers inference horizons up to $H = 30$ periods ahead, and encompasses the following estimators and asymptotic inference formulations. Finite-order (LS) and sieve (S-LS) least-squares VAR(p) estimators, where $p = 10$ in the finite-order setting while $p = 30$ in the sieve setting. Local projections (LP) estimator of order $q = 10$, with Newey-West covariance estimates. Finite-order ridge VAR(p) estimator, with (RLS-as) and without (RLS) asymptotic shrinkage correction, where $p = 10$. Sieve ridge VAR(p) estimator (S-RLS), where $p = 30$. All ridge estimators are constructed with lag-adapted regularization. I use both 10-fold cross-validation (CV) and out-of-sample (OOS) validation (on the last 20% of the sample) to select Λ^ℓ.

For the sake of comparability, as H is kept fixed with respect to p two sample size regimes are considered. For sieve inference, given a set maximal inference horizon H, it is necessary to estimate a VAR(H) model in order to construct correct Delta method CIs (cf. Lütkepohl (1988), proof of Theorem 1). On the other hand, finite-order VAR modeling imposes no such relationship between p and H. Because the number of parameters to be estimated in the sieve simulations is 300% that of the finite-order simulations, in sieve simulations $T = \{300, 900\}$ and $p = H = 30$; in finite-order simulations, $T = \{100, 300\}$. The ratios T/p remain the same across methods. Local projections are considered independent of these concerns since there is no direct "sieve" LP counterpart.

Finally, to avoid the introduction of an additional tuning parameter to the ridge procedure, when the asymptotic variance matrix with shrinkage $\hat{\Gamma}_{T,\overline{\Lambda}}^{-1} \hat{\Gamma}_T \hat{\Gamma}_{T,\overline{\Lambda}}^{-1}$ is estimated, I set $n = 0$ and $\overline{\Lambda}_{1:2} = \hat{\Lambda}$, where $\hat{\Lambda}$ is the regularized obtained by either OOS or CV. More details on the numerical implementation of the estimators can be found in Appendix 6.3.

Confidence Intervals Coverage. In Figure 1 impulse responses and Delta method confidence intervals (CIs) based on the various estimation strategies detailed above are compared using $B = 1000$ replications. In all simulations the nominal CI level is set to 95%. Coverage

Figure 1: Average coverage rates and lengths of 95% CIs.



(a) Sieve CIs: $T = 300$, $p = 30$, $q = 10$

(b) Sieve CIs: $T = 900$, $p = 30$, $q = 10$

(c) Finite-order CIs: $T = 100$, $p = 10$, $q = 10$

(d) Finite-order CIs: $T = 300$, $p = 10$, $q = 10$

Note: Coverage (left panel) and average length (right panel), $B = 1000$ Monte Carlo replications.

and lengths in Figure 1 are averaged across all individual impulse responses.[9] I refrain from producing bootstrapped CIs as they would require additional theoretical justification in the case of ridge, although bootstrap techniques are potentially of high interest.[10] To produce structural dynamic responses, I use a recursive identification strategy: each estimation method also recovers a lower-triangular matrix P via the Cholesky decomposition of $\hat{\Sigma}_u$.[11]

The takeaway of this simple simulation exercise is that impulse response confidence intervals based on RLS estimates using cross-validation yield the most robust estimation overall. In the sieve setting, Figures 1a and 1b suggest that the improvements of coverage come at virtually no costs in length: the better performance of sieve RLS (S-RLS) CIs is driven by the smoothing of IRFs induced by regularization. Comparing S-RLS(OOS) and S-LS intervals further highlights the point, because coverage is comparable although ridge produces slightly shorter intervals. These results translate very well to the finite-order estimation setting, Figures 1c and 1d. To note is that the application of Theorem 3 makes for arguably *too much* regularization, see the subpar quality of the RG(CV)-as confidence intervals. This is possibly due to the fact that n was chosen exogenously and such that the full penalization matrix effects the asymptotic variance. At this moment, it is unclear whether the tuning of n from data would result in gain worth the cost of additional complexity, and this is a question that could be tackled in future research.

Of note is that, while impulse responses produced by penalized estimation are smoother than LS or LP IRFs, residuals are naturally larger: the former is an implication of the ridge regularization shrinkage properties; the latter is due to the definition of the ridge estimator as a sub-optimal solution to the non-penalized least squares problem. This translates to regularization-based CIs being potentially larger than their least-squares counterparts, at least at shorter horizons. In the above simulation setting, such behavior is extremely mild and hard to read from Figure 1. But the effects of residuals inflation may be much more noticeable in smaller samples, where regularization plays a bigger role. Whether this behavior degrades or improves the statistical properties of ridge CIs is also likely to depend on the specific process that is being analyzed, and on the precise concerns of the researcher.

4.2 Model Selection Revised

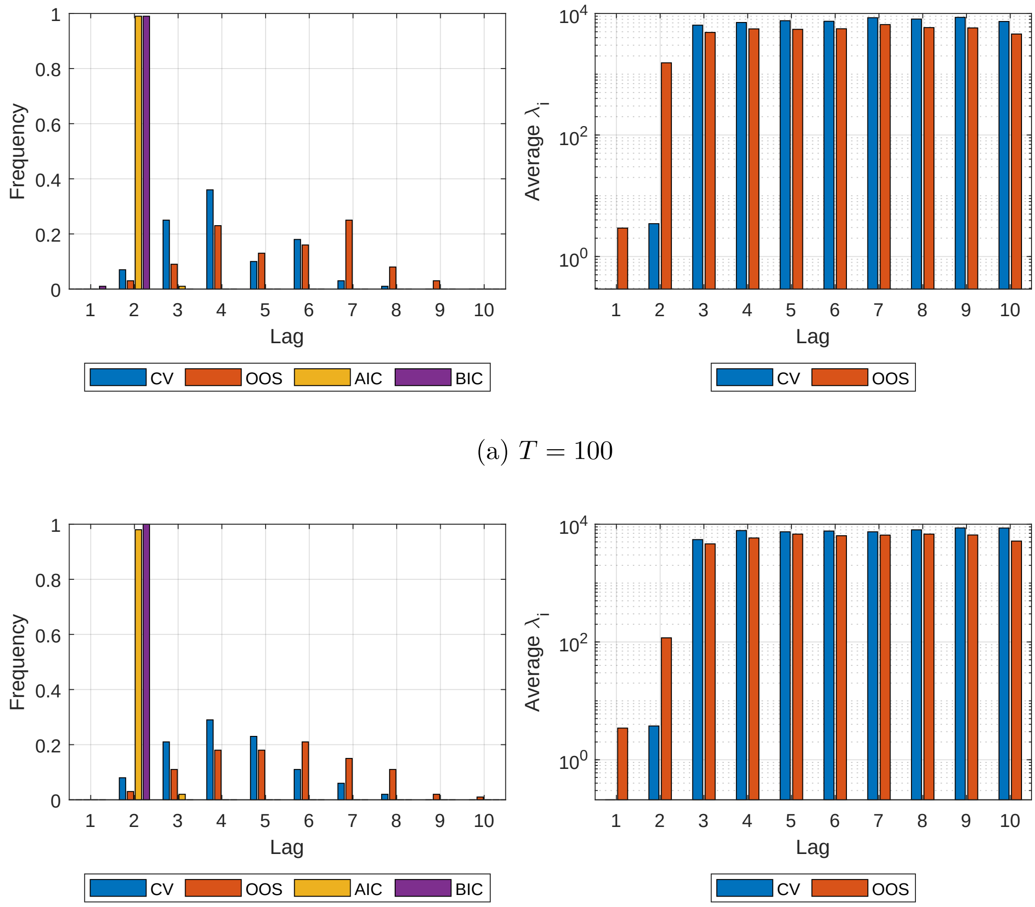
One might believe that ridge, by partially doing away with model selection, is doing double duty. Consider the following: in practice, when estimating $\hat{B}_p^R(\Lambda^\ell)$, one could simply set p to

[9] For a graphical illustration of the estimated IRFs and related confidence intervals on a single simulated sample, please see Figure 3 in Supplementary Appendix D.

[10] Theoretical results for sieve inference with VAR models has already been developed in Inoue and Kilian (2002), for example.

[11] This approach at identifying structural relations between time series components is by no means unique. The recursive approach is also the one built-in into the DGP by design, since Kilian and Kim (2011) define $\Sigma_u = PP'$ for a given lower-triangular P. The same specification is also used in Inoue and Kilian (2002).

Figure 2: Comparison of OOS and CV methods for choosing lag-adapted Λ^ℓ.



(a) $T = 100$

(b) $T = 300$

Note: Frequency (left) of selection of a given number of VAR lags. For OOS and CV methods, the frequency is counted as fraction of times in which λ_i, $1 \leq i \leq 10$, is such that $\lambda_i < 9900$ (less than 90% of the upper bound); for AIC and BIC, frequency is counted as the fraction of times a model of lag length p is selected with $p_{\max} = 10$. Average value (right) of λ_i for OOS and CV methods. $B = 1000$ replications.

be conservatively much larger than any value selected by information criteria like AIC or BIC. Does this mean that the question of model selection has been sidestepped? Unfortunately, or perhaps reassuringly, this is not really the case. While it is true that the data-driven choice of Λ^ℓ via either OOS or CV in principle gives arbitrary flexibility, there is a limit to the amount of information that can be extracted from any given sample. This limitation does not depend on the finite-order VAR or sieve VAR framework either, as both out-of-sample and cross-validation techniques are agnostic about the *interpretation* of the estimated model.

To better explain, Figure 2 shows the behavior of both OOS and CV methods in selecting $\{\lambda_i\}_{i=1}^p$ for VAR(10) models versus the lag selection choices of both AIC and BIC criteria.

Information criteria are estimated with $p_{\max} = 10$. The data is drawn from the sample data-generating process of the Monte Carlo impulse response simulation. Since OOS/CV do not chose a value for lag depth p, I consider the frequency of times a number p of coefficients $\{\lambda_i\}$ are selected such that $\lambda_i < 9900$.[12] While OOS/CV cannot properly produce model selection, they indeed penalize coefficients at lags > 3 virtually as much as possible under the optimization constraints.[13] Figure 2 also suggests that OOS tends to systematically penalize less, while CV is sharper and produces more "compact" models. While many coefficient matrices, therefore, feature small coefficients due to shrinkage, the highly non-linear nature of the mappings from VAR coefficients to IRFs (and CIs) means that a priori one can not immediately argue that strongly penalized matrices are unimportant.

Indeed, OOS and CV are free to *not* penalize coefficients at deeper lags, so that the lag order "trimming" that AIC and BIC perform can not be easily compared one-to-one to the flexible specification of lag-adapted RLS. The results of Figure 2 should be taken as limited evidence of the general properties of both out-of-sample and cross-validation methods. Finally, the simplicity of a VARMA(1,1) data-generating process needs to be taken into account, too.

5 Conclusion

In this paper, I have formally introduced ridge regularization to the estimation problem of vector autoregressive models. While the ridge has been applied to VARs many times before, previous work has mostly concentrated on either its Bayesian derivation as a consequence of specific priors, or on its efficacy in forecasting tasks, often in comparison to a number of other regularization methods. I have instead worked on first showing that the properties of the ridge known in the i.i.d. cross-sectional regression setting carry over to the linear time series setting, albeit with complications. Secondly, I have derived the asymptotic properties of the RLS estimator, which are useful for inference. Finally, I developed the application of ridge to the VAR sieve approximation framework, showing that its use can be beneficial when one is performing impulse response function analysis.

At first glance, Theorems 1, 2 and 4 can be disconcerting, in that they show that the RLS estimator yields no *asymptotic* gains when compared to textbook LS. In fact, all limit quantities are the same. This really can be seen in two ways. If one wants to preserve standard asymptotic properties, e.g. consistency, then there are technical constraints that must be imposed on the rate of convergence of the regularizer Λ. Assumption D considered here does not allow for

[12] This threshold is chosen because it is equivalent to $\lambda_i < 0.9\,\lambda_{\text{upperbound}}$, where $\lambda_{\text{upperbound}} = 10\,000$ is the upper bound for λ_i in OOS/CV optimizations, see Appendix 6.3.

[13] I have discussed above that the shrinkage behavior of the non-degenerate RLS estimator does not allow for penalizing different lags *independently* of others, cf. Proposition 1. However, lag-adapted regularizers arguably have the best structure to target a higher penalty at a specific coefficient matrix.

asymptotic improvements, i.e. there is no bias-variance tradoff in the limit. On the other hand, Theorem 3, under appropriately modified assumptions, shows that shrinkage plays a role in the asymptotic distribution of ridge estimators: unfortunately, the applicability of Theorem 3 is undermined by the fact that a partitioning parameter $0 \leq n \leq p$ has to be additionally tuned. This makes it hard to say whether asymptotic shrinkage leads to more accurate in-sample inference than "standard" asymptotics.

The simulation results of Section 4.1 indicate that RLS estimates improve significantly over standard LS or local projection methods, especially when sample sizes are smaller. Accordingly, the suggestion from this paper is to consider the RLS estimator really as a *small sample* method, which trades estimation bias for impulse response smoothness. Its use can be beneficial in inference when one is facing a shortage of data, which make non-regularized estimates weak in numerical estimation. For suitably large samples it might be that the bias-variance tradeoff of ridge, coupled with its additional numerical complexity in tuning the regularization parameters, is less beneficial or wholly unnecessary.

More generally, consider the researcher dealing with medium-to-large dimensional VAR models, with possibly a small sample dataset or under the necessity to consider a large number of autoregressive lags, or both. To them, the modification of the least squares estimator to the ridge-penalized RLS estimator can be very useful. This usefulness clearly requires one to know the potential limitations and pitfalls of this regularization technique. Yet, there are a number of questions that would be interesting to address which I have not touched upon directly in this paper.

Future Research. I have put a significant focus on the lag-adapted structure with out-of-sample and cross-validation methods. An important puzzle is whether there are meaningfully better ways to choose Λ, perhaps complementing or even doing away with the lag-adapted structure. The comprehensive review of Bauer and Lukas (2011) showcases a wide range of alternative techniques, but it is fundamental to strike a balance between the flexibility of a loose structure on Λ and numerical feasibility. In a VAR setting, the number of parameters grows quadratically in model size K and linearly in lag size p, so that large models can require many thousands of parameters. This can create significant computational challenges when implementing or evaluating ridge estimators. Further, it is desirable that any data-driven method for choosing Λ be, in theory, robust to the correlation structure of time series data.

The option of applying the bootstrap has been left open as I have relied strictly on asymptotic (Delta method) techniques. The main advantage of bootstrapped inference would be that one could more directly address the finite sample estimation bias and variance due to regularization. A problem in implementing a bootstrap scheme - apart from deriving the appropriate theory - should be that if Λ needs to be selected repeatedly, then the procedure might be computationally taxing. However, it seems likely that one could take some "shortcuts" which are asymptotically negligible but make computation more efficient, like in Kilian (1998).

Finally, given the increasing availability of very large datasets ("big data"), the high-dimensional setting seems a natural next step in the study of any regularized VAR estimator. Inference in such environments requires completely different theoretical devices, see for example the de-biased estimators of Krampe et al. (2020). In particular, one wonders how the dense estimates of ridge compare theoretically and practically with the sparse estimates provided by high-dimensional techniques like de-sparsified LASSO, especially when dealing with very large macroeconometric datasets.

6 Appendix

6.1 Ridge Asymptotic Theory

Proof. Theorem 1

(a) Assumptions A-B imply directly that $\hat{\Gamma}_T$ is a consistent estimator for Γ: in particular, y_t is a stationary, stable and ergodic VAR process.

(b) Rewriting $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ yields

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) &= (ZZ' \otimes I_K + \Lambda)^{-1}[(Z \otimes I_K)((Z' \otimes I_K)\boldsymbol{\beta} + \mathbf{u}) + \Lambda\boldsymbol{\beta}_0] && (4)\\
&= (ZZ' \otimes I_K + \Lambda)^{-1}[(ZZ' \otimes I_K)\boldsymbol{\beta} + (Z \otimes I_K)\mathbf{u} + \Lambda\boldsymbol{\beta}_0] && (5)\\
&= (ZZ' \otimes I_K + \Lambda)^{-1}[(ZZ' \otimes I_K + \Lambda)\boldsymbol{\beta} + (Z \otimes I_K)\mathbf{u} + \Lambda(\boldsymbol{\beta}_0 - \boldsymbol{\beta})] && (6)\\
&= \boldsymbol{\beta} + (ZZ' \otimes I_K + \Lambda)^{-1}\Lambda(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + (ZZ' \otimes I_K + \Lambda)^{-1}(Z \otimes I_K)\mathbf{u} && (7)
\end{aligned}$$

I study the last two terms of the last inequality separately. The first term is $o_p(1)$ under Assumption D.(ii),

$$\left(\left(\frac{ZZ'}{T}\right) \otimes I_K + \frac{\Lambda}{T}\right)^{-1} \frac{\Lambda}{T}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) = \left(\left(\frac{ZZ'}{T}\right) \otimes I_K + o_p(1)\right)^{-1} o_p(1)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \xrightarrow{p} 0 \quad (8)$$

since $(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = (\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}_0) + (\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) = (\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}_0) + o_p(1)$. Considering the matrix sequence

$$\boldsymbol{\zeta}_T = [T^{-1}(ZZ'), T^{-1}\Lambda] \quad (9)$$

which under Assumptions B and D.(ii) converges in probability to $[\Gamma, 0]$, by the continuous mapping theorem (Davidson, 1994) the second term gives

$$\left(\left(\frac{ZZ'}{T}\right) \otimes I_K + o_p(1)\right)^{-1} \left(\frac{1}{T}(Z \otimes I_K)\mathbf{u}\right) \xrightarrow{p} \Gamma^{-1}\,\mathbb{E}[(Z \otimes I_K)\mathbf{u}] = 0 \quad (10)$$

under Assumption A.

(c) The residuals $\hat{U}$ can be written as

$$\hat{U} = Y - \hat{B}^R_{de}Z = BZ + U - \hat{B}^R_{de}Z = U + (B - \hat{B}^R_{de})Z \quad (11)$$

Thus

$$\frac{\hat{U}\hat{U}'}{T} = \frac{UU'}{T} + (B - \hat{B}^R_{de})\left(\frac{ZZ'}{T}\right)(B - \hat{B}^R_{de})' + (B - \hat{B}^R_{de})\left(\frac{ZU'}{T}\right) + \left(\frac{UZ'}{T}\right)(B - \hat{B}^R_{de})' \quad (12)$$

From (a) we have that $\mathrm{vec}(B) - \mathrm{vec}(\hat{B}^R_{de}) = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R = o_p(1)$, so $(B - \hat{B}^R_{de}) \xrightarrow{p} 0$, while $T^{-1}(ZZ') \xrightarrow{p} \mathbb{E}[ZZ']$ and $T^{-1}(ZU') \xrightarrow{p} \mathbb{E}[ZU'] = 0$: the terms involving these quantities then vanish asymptotically. Lastly, the first term of the sum gives

$$\frac{UU'}{T} = \frac{1}{T}\sum_{t=1}^{T} u_t u_t' \xrightarrow{p} \mathbb{E}[u_t u_t'] = \Sigma_u \quad (13)$$

for $T \to \infty$ under Assumptions A and B.

(d) With the same expansion used in (b),

$$\sqrt{T}(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}) = Q_T^{-1} \frac{\Lambda}{\sqrt{T}} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + Q_T^{-1} \left(\frac{1}{\sqrt{T}} (Z \otimes I_K) \mathbf{u} \right) \tag{14}$$

where $Q_T = (T^{-1}(ZZ') + T^{-1}\Lambda) \xrightarrow{p} \Gamma^{-1}$. Following the arguments above, the first term in the sum converges in probability,

$$Q_T^{-1} \frac{\Lambda}{\sqrt{T}} (\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} + o_p(1)) \xrightarrow{p} \Gamma^{-1} \Lambda_0 (\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}_0) \tag{15}$$

The second term has normal limiting distribution,

$$Q_T^{-1} \left(\frac{1}{\sqrt{T}} (Z \otimes I_K) \mathbf{u} \right) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma_u) \tag{16}$$

see Lütkepohl (2005), Proposition 3.1. By Slutky's theorem claim (d) follows.

□

Proof. Theorem 2

(1) Since Assumption (i) implies Assumption (ii), results (a)-(c) are unchanged, while (d) now involves the limit

$$Q_T^{-1} \frac{\Lambda}{\sqrt{T}} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) = Q_T^{-1} \cdot o_p(1) \cdot (\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} + o_p(1)) \xrightarrow{p} 0 \tag{17}$$

yielding (d′).

(2) Assuming $\boldsymbol{\beta}_0 \xrightarrow{p} \underline{\boldsymbol{\beta}}_0$ simplifies the terms in the proof of Theorem 1 since now $(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = o_p(1)$. The weaker convergence rate imposed by Assumption D.(iii) does not influence results (a)-(c), but matters for the limiting distribution. However, since $\sqrt{T}\boldsymbol{\beta}_0$ is asymptotically normal,

$$Q_T^{-1} \frac{\Lambda}{\sqrt{T}} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) = Q_T^{-1} \cdot \frac{\Lambda}{T} \cdot \sqrt{T}(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}_0) = Q_T^{-1} \cdot o_p(1) \cdot O_p(1) \xrightarrow{p} 0 \tag{18}$$

and again (d′) follows.

□

Proof. Theorem 3

The stated results reduce to studying the behavior of two components used in the proof of Theorem 1 and Theorem 2, under the additional simplification of $\boldsymbol{\beta}_0 = 0$.

(a) Identical to result (a) in Theorem 1.

(b) Consistency follows by noting that

$$\frac{\Lambda}{T} \boldsymbol{\beta} \xrightarrow{p} \begin{bmatrix} 0 & 0 \\ 0 & \overline{\Lambda}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \overline{\Lambda}_2 \sqrt{T}^{-1} \boldsymbol{b}_2 \end{bmatrix} \xrightarrow{p} 0 \tag{19}$$

(c) Follows from (c), Theorem 1 and (b) above.

(d″) The bias term in the expression of $\sqrt{T}(\hat{\boldsymbol{\beta}}^R - \boldsymbol{\beta})$ is driven by

$$\frac{\Lambda}{\sqrt{T}}\boldsymbol{\beta} \propto \begin{bmatrix} \sqrt{T}^{-1}\Lambda_1\,\boldsymbol{\beta}_1 \\ \sqrt{T}^{-1}\Lambda_2\,T^{-(1/2+\delta)}\boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} \sqrt{T}^{-1}\Lambda_1\,\boldsymbol{\beta}_1 \\ T^{-1}\Lambda_2\,(T^{-\delta}\boldsymbol{b}_2) \end{bmatrix} \xrightarrow{p} 0 \tag{20}$$

meaning there is no asymptotic bias. On the other hand, notice that

$$\left(\left(\frac{ZZ'}{T}\right)\otimes I_K + \frac{\Lambda}{T}\right)^{-1} \xrightarrow{p} (\Gamma + \overline{\Lambda}_{1:2})^{-1}\otimes I_K \tag{21}$$

Setting $(\Gamma + \overline{\Lambda}_{1:2}) = \Gamma_{\overline{\Lambda}}$ yields the claim since there are no further simplifications in the asymptotic variance formula, cf. proof of (d), Theorem 1.

□

Proof. THEOREM 4

I make a straightforward adaptation of the proof found in Hamilton (1994), Proposition 11.2. Define $\hat{\Sigma}_T^* = T^{-1}(UU')$, which is expanded to

$$\begin{aligned}\hat{\Sigma}_T^* &= \frac{1}{T}(Y - BZ)(Y-BZ)' \qquad (22)\\ &= \frac{1}{T}\Big(Y - \hat{B}^R_{de}Z + (\hat{B}^R_{de} - B)Z\Big)\Big(Y - \hat{B}^R_{de}Z + (\hat{B}^R_{de} - B)Z\Big)' \qquad (23)\\ &= \hat{\Sigma}_T^R + (\hat{B}^R_{de} - B)\left(\frac{ZZ'}{T}\right)(\hat{B}^R_{de} - B)' + \\ &\quad + \frac{1}{T}\Big((Y - \hat{B}^R_{de}Z)Z'(\hat{B}^R_{de} - B)' + (\hat{B}^R_{de} - B)Z(Y - \hat{B}^R_{de}Z)'\Big) \qquad (24)\end{aligned}$$

Contrary to the least squares estimator, cross-terms do not cancel out since for $0 \preceq \Lambda$ the residuals $(Y - \hat{B}^R_{de}Z)$ are not in the orthogonal space of Z. From the consistency results of Theorem 1,

$$(\hat{B}^R_{de} - B)\left(\frac{ZZ'}{T}\right)(\hat{B}^R_{de} - B)' = o_p(1)\left(\frac{ZZ'}{T}\right)o_p(1) \xrightarrow{p} 0 \tag{25}$$

and

$$\sqrt{T}(\hat{B}^R_{de} - B)\left(\frac{ZZ'}{T}\right)(\hat{B}^R_{de} - B)' = O_p(1)\left(\frac{ZZ'}{T}\right)o_p(1) \xrightarrow{p} 0 \tag{26}$$

Further,

$$\sqrt{T}\left[\frac{1}{T}(Y - \hat{B}^R_{de}Z)Z'(\hat{B}^R_{de} - B)'\right] = \left(\frac{\hat{U}Z'}{T}\right)\sqrt{T}(\hat{B}^R_{de} - B)' \xrightarrow{p} 0 \tag{27}$$

since again $\sqrt{T}\hat{B}^R_{de}$ is asymptotically normal, and $T^{-1}(\hat{U}Z') = T^{-1}(UZ') + (B - \hat{B}^R_{de})\cdot T^{-1}(ZZ') = T^{-1}(UZ') + o_p(1) \xrightarrow{p} \mathbb{E}[UZ'] = 0$. The same holds for the remaining transpose term, too.

It is thus proven that $\sqrt{T}(\hat{\Sigma}_T^* - \hat{\Sigma}_T^R) \xrightarrow{p} 0$, meaning $\sqrt{T}(\hat{\Sigma}_T^* - \Sigma_u) \xrightarrow{p} \sqrt{T}(\hat{\Sigma}_T^R - \Sigma_u)$ so that the two terms may be exchanged in computing the joint asymptotic distribution. Theorem 1 accordingly yields

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_0^R - \boldsymbol{\beta} \\ \mathrm{vec}(\hat{\Sigma}_T^R) - \mathrm{vec}(\Sigma_u) \end{bmatrix} \xrightarrow{p} \begin{bmatrix} Q_T^{-1}\frac{\Lambda}{\sqrt{T}}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + Q_T^{-1}\frac{1}{\sqrt{T}}\boldsymbol{\xi} \\ \frac{1}{\sqrt{T}}\boldsymbol{\eta} \end{bmatrix} \tag{28}$$

where $\boldsymbol{\xi} = (Z \otimes I_K)\mathbf{u}$ and $\boldsymbol{\eta} = \operatorname{vec}(UU' - \Sigma_u)$. As in Hamilton (1994), Proof of Proposition 11.2, $(\boldsymbol{\xi}', \boldsymbol{\eta})'$ is a martingale difference sequence, thus the claim

$$\sqrt{T}\begin{bmatrix} \hat{\boldsymbol{\beta}}_0^R - \boldsymbol{\beta} \\ \operatorname{vec}(\hat{\Sigma}_T^R) - \operatorname{vec}(\Sigma_u) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} \Gamma^{-1}\Lambda_0(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) \\ 0 \end{bmatrix}, \begin{bmatrix} \Gamma^{-1} \otimes \Sigma_u & 0 \\ 0 & \Omega \end{bmatrix}\right) \tag{29}$$

as $T \to \infty$ follows. When the strengthened assumptions (1) or (2) of Theorem 2 are used instead, the non-zero limiting mean vanishes

$$Q_T^{-1}\left(\frac{\Lambda}{T}\right)\sqrt{T}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \xrightarrow{p} 0 \tag{30}$$

proving that the joint asymptotic distribution is mean-zero Gaussian.

Finally, to compute the explicit expression of the asymptotic variance Ω, one must take care and note that u_t is not assumed to be normally distributed, contrary to the remainder of the proof in Hamilton (1994), pp. 342-343. A correct expression for i.i.d. non-Gaussian u_t can be found in Remark 2.1, Brüggemann et al. (2016), yielding

$$\Omega = \operatorname{Var}[\operatorname{vec}(u_t u_t')] = \mathbb{E}[\operatorname{vec}(u_t u_t')\operatorname{vec}(u_t u_t')'] - \boldsymbol{\sigma}\boldsymbol{\sigma}' \tag{31}$$

where $\boldsymbol{\sigma} = \operatorname{vec}(\Sigma_u)$. □

6.2 Sieve Asymptotic Theory

Proof. Theorem 5

To begin with, by rewriting

$$\begin{aligned} \hat{B}_p^R(\Lambda^\ell) &= \left(\frac{YZ'}{T-p}\right)\left(\left(\frac{ZZ'}{T-p}\right) + (T-p)^{-1}\Lambda^\ell\right)^{-1} && (32) \\ &= \hat{\Gamma}_p^1(\hat{\Gamma}_p + (T-p)^{-1}\Lambda^\ell)^{-1} && (33) \\ &= \hat{\Gamma}_p^1\hat{L}_p^{-1} && (34) \end{aligned}$$

I can consider the expansion

$$\hat{B}_p^R(\Lambda^\ell) - B_p = \hat{\Gamma}_p^1\hat{L}_p^{-1} - B_p\hat{L}_p\hat{L}_p^{-1} = \left[(T-p)^{-1}(U_pZ') - B_p(T-p)^{-1}\Lambda^\ell\right]\hat{L}_p^{-1} \tag{35}$$

where $U_p = Y - B_pZ$. It then follows that (Wiener and Masani, 1958)

$$\|\hat{B}_p^R(\Lambda^\ell) - B_p\|_F \leq \|\hat{L}_p^{-1}\|\|U_{1T}\|_F + \|\hat{L}_p^{-1}\|\|U_{2T}\|_F \tag{36}$$

where

$$U_{1T} = (T-p)^{-1}((U_p - U)Z'), \qquad U_{2T} = (T-p)^{-1}(UZ') \tag{37}$$

I now prove that $\|\hat{L}_p^{-1}\|$ is bounded above in probability uniformly over p. Using the upper bound $\|\hat{L}_p^{-1}\| \leq \|\Gamma_p^{-1}\| + \|\hat{L}_p^{-1} - \Gamma_p^{-1}\|$ and noting that $\|\Gamma_p^{-1}\|$ is bounded above by some constant F uniformly over $p \geq 1$ (Lewis and Reinsel, 1985), I focus on the second term. Notice that

$$\mathbb{E}\left[\|\hat{\Gamma}_p - \Gamma_p\|^2\right] \leq \mathbb{E}\left[\|\hat{\Gamma}_p - \Gamma_p\|_F^2\right] \leq C_1\frac{(pK)^2}{(T-p)} \to 0 \tag{38}$$

by Assumption (i) and Hannan (1970), Chapter 4. Then

$$\begin{aligned}
\|\hat{L}_p^{-1} - \Gamma_p^{-1}\| &\leq \|\hat{L}_p^{-1}(\hat{L}_p - \Gamma_p)\Gamma_p^{-1}\| && (39)\\
&\leq \|\hat{L}_p^{-1}\|\|(\hat{\Gamma}_p - \Gamma_p) - (T-p)^{-1}\Lambda^{\ell}\|\|\Gamma_p^{-1}\| && (40)\\
&\leq \left[F + \|\hat{L}_p^{-1} - \Gamma_p^{-1}\|\right]\left[\|\hat{\Gamma}_p - \Gamma_p\| + \|(T-p)^{-1}\Lambda^{\ell}\|\right] F && (41)
\end{aligned}$$

It follows from the above inequalities that

$$0 \leq \Upsilon_{p,T}^{\Lambda} = \frac{\|\hat{L}_p^{-1} - \Gamma_p^{-1}\|}{(F + \|\hat{L}_p^{-1} - \Gamma_p^{-1}\|)F} \leq \|\hat{\Gamma}_p - \Gamma_p\| + \|(T-p)^{-1}\Lambda^{\ell}\| \xrightarrow{p} 0 \tag{42}$$

under Assumption (iii). Hence $\|\hat{L}_p^{-1} - \Gamma_p^{-1}\| = F^2\Upsilon_{p,T}^{\Lambda}/(1 - F\Upsilon_{p,T}^{\Lambda})$ also converges in probability to zero as $T \to \infty$.

The remainder of the proof is as in Lewis and Reinsel (1985), Theorem 1. □

Proof. Theorem 6

The proof is simple, in that it amounts to adjusting Theorem 2 of Lewis and Reinsel (1985) to work with estimator $\hat{B}_p^R(\Lambda^{\ell})$. Then, Lewis and Reinsel (1985), Theorem 3 and 4 follow under the stated assumptions – since they do not directly depend on the specific form of the VAR coefficient estimator – completing the proof by virtue of the Cramér-Wold device.

Let $\{l(p)\}$ be a sequence of $(Kp \times 1)$ vectors such that for $p = 1, 2, \ldots$

$$0 < M_1 \leq \|l(p)\|^2 \leq M_2 < \infty \tag{43}$$

I prove that, for $(\hat{\boldsymbol{\beta}}_p^R(\Lambda^{\ell}) - \boldsymbol{\beta}_p) = (\text{vec}\{\hat{B}_p^R(\Lambda^{\ell})\} - \text{vec}\{B_p\})$,

$$(T-p)^{1/2} l(p)'(\hat{\boldsymbol{\beta}}_p^R(\Lambda^{\ell}) - \boldsymbol{\beta}_p) - (T-p)^{1/2} l(p)'\text{vec}\left\{(T-p)^{-1}(UZ')\Gamma_p^{-1}\right\} \xrightarrow{p} 0 \tag{44}$$

as $T \to \infty$. One may rewrite the above terms:

$$\begin{aligned}
&(T-p)^{1/2} l(p)'\left[\text{vec}\left\{\left(\frac{U_p Z'}{(T-p)} - B_p(T-p)^{-1}\Lambda^{\ell}\right)\hat{L}_p^{-1}\right\} - \text{vec}\left\{\frac{UZ'}{(T-p)}\Gamma_p^{-1}\right\}\right] && (45)\\
&= (T-p)^{1/2} l(p)'\left[\text{vec}\left\{\frac{(U_p Z)\hat{L}_p^{-1} - (UZ')\Gamma_p^{-1}}{(T-p)}\right\} - \text{vec}\left\{B_p(T-p)^{-1}\Lambda^{\ell}\hat{L}_p^{-1}\right\}\right] && (46)
\end{aligned}$$

Since

$$(U_p Z)\hat{L}_p^{-1} - (UZ')\Gamma_p^{-1} = (U_p Z' - UZ')(\hat{L}_p^{-1} - \Gamma_p^{-1}) + (UZ')(\hat{L}_p^{-1} - \Gamma_p^{-1}) + (U_p Z' - UZ')\Gamma_p^{-1} \tag{47}$$

then equation (44) can be separated into the following terms

$$\begin{aligned}
l(p)'[((\hat{L}_p^{-1} - \Gamma_p^{-1}) \otimes I_K)\text{vec}&\left\{(T-p)^{1/2}\, U_{1T}\right\} + \\
&+ ((\hat{L}_p^{-1} - \Gamma_p^{-1}) \otimes I_K)\text{vec}\left\{(T-p)^{1/2}\, U_{2T}\right\} + \\
&+ (\Gamma_p^{-1} \otimes I_K)\text{vec}\left\{(T-p)^{1/2}\, U_{1T}\right\} +
\end{aligned}$$

$$
\begin{aligned}
&- \operatorname{vec}\{B_p'\}'(\hat{L}_p^{-1} \otimes I_K)\operatorname{vec}\{(T-p)^{-1/2}\Lambda^\ell\}] \\
&= w_{1T} + w_{2T} + w_{3T} - w_{4T} \qquad (48)
\end{aligned}
$$

where U_{1t}, U_{2T} and $\hat{L}_p^{-1}$ are defined as for Theorem 5; terms w_{1T}, w_{2T}, w_{3T} and w_{4T} are defined from the respective terms in the sum at the LHS.

Like in the proof of Theorem 5, it holds that

$$
|w_{1t}| \leq \|l(p)\| \cdot p^{1/2} \, \|\hat{L}_p^{-1} - \Gamma_p^{-1}\| \cdot \|p^{-1/2}(T-p)^{1/2} U_{1T}\|_F \qquad (49)
$$

where $\|l(p)\| \leq M_2^{1/2}$ and it can be shown that $p^{1/2}\|\hat{L}_p^{-1} - \Gamma_p^{-1}\| \xrightarrow{p} 0$ under Assumption (i). Also,

$$
\mathbb{E}[\|p^{-1/2}(T-p)^{1/2} U_{1T}\|] \leq C_1 (T-p)^{1/2} \sum_{j=p+1}^{\infty} \|A_j\|_F \to 0 \qquad (50)
$$

as $T \to \infty$ (Lewis and Reinsel, 1985) by Assumption (ii), and thus $w_{1T} \xrightarrow{p} 0$. Using similar arguments it is easy to prove that $w_{2T} \xrightarrow{p} 0$, and additionally $w_{3T} \xrightarrow{p} 0$ under Assumptions (i)-(ii).

Finally, notice the last term may be written as

$$
\begin{aligned}
w_{4T} &= \operatorname{vec}\{B_p'\}' \left[((\hat{L}_p^{-1} - \Gamma_p^{-1}) \otimes I_K) + (\Gamma_p^{-1} \otimes I_K) \right] \operatorname{vec}\{(T-p)^{-1/2}\Lambda^\ell\} \qquad (51) \\
&= w_{4T}^A + w_{4T}^B \qquad (52)
\end{aligned}
$$

for which $|w_{4T}| \leq \left|w_{4T}^A\right| + \left|w_{4T}^B\right|$, where

$$
\left|w_{4T}^A\right| \leq \|\boldsymbol{\beta}_p\| \cdot \|\hat{L}_p^{-1} - \Gamma_p^{-1}\| \cdot (T-p)^{-1/2} \|\Lambda^\ell\| \xrightarrow{p} 0 \qquad (53)
$$

trivially, given previous results and Assumption (iii), and further

$$
\left|w_{4T}^B\right| \leq \|\boldsymbol{\beta}_p\| \cdot \|\Gamma_p^{-1}\| \cdot (T-p)^{-1/2} \|\Lambda^\ell\| \leq C_2 \cdot (T-p)^{-1/2} \|\Lambda^\ell\| \to 0 \qquad (54)
$$

since $\|\boldsymbol{\beta}_p\| \leq \sum_{j=1}^{\infty} \|A_j\|_F < \infty$, $\|\Gamma_p^{-1}\|$ is uniformly bounded above over $p \geq 1$, and Assumption (iii) again holds. Thus $|w_{iT}| \xrightarrow{p} 0$ for $i = 1, \ldots, 4$ as claimed. □

Asymptotic Theory. Below I adapt the proof of Lemma 2, Lütkepohl and Poskitt (1991), p. 493 to the case when the estimator of a VAR(p) model with $p < \infty$ under a VAR(∞) is given by the RLS (lag-adapted) estimator $\hat{B}_p^R(\Lambda^\ell)$.

Define

$$
\hat{\Sigma}_p^R = \frac{1}{T} \sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{p} \hat{B}_{p,j}^R(\Lambda^\ell) y_{t-j} \right) \left(y_t - \sum_{j=1}^{p} \hat{B}_{p,j}^R(\Lambda^\ell) y_{t-j} \right)' \qquad (55)
$$

where clearly $\hat{B}_{p,j}^R(\Lambda^\ell)$ for $j = 1, \ldots, p$ are the lag coefficient blocks of $\hat{B}_p^R(\Lambda^\ell)$.

Lemma 1. *Under the conditions of Theorem 6,*

$$
\sqrt{T} vech\{\hat{\Sigma}_p^R - \Sigma\} \xrightarrow{d} \mathcal{N}(0, \Omega_\Sigma)
$$

Proof. First, by defining $y_{t,p+1} = (y'_t, y'_{t-1}, \dots, y'_{t-p})'$ and $\hat{A}^R_0 = I_K$, $\hat{A}^R_j = \hat{B}^R_{p,j}(\Lambda^\ell)$ for $j = 1, \dots, p$, where the explicit dependency on Λ^ℓ has been dropped for convenience of notation, and $\hat{A}^R(p) = (I_K, -\hat{A}^R_1, \dots, -\hat{A}^R_p)$, one can write

$$\hat{\Sigma}^R_p = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{A}^R(p)y_{t,p+1}\right)\left(\hat{A}^R(p)y_{t,p+1}\right)' = \hat{A}^R(p)\left(\frac{1}{T}\sum_{t=1}^{T} y_{t,p+1}y'_{t,p+1}\right)\hat{A}^R(p)' \tag{56}$$

I then note that

$$\hat{\Sigma}^R_p - \Sigma = (\hat{\Sigma}^R_p - \hat{\Sigma}_p) + (\hat{\Sigma}_p - \Sigma) \tag{57}$$

where $\hat{\Sigma}_p$ is the least squares estimator of the same VAR(p) model. Write $\hat{\Sigma}_p - \Sigma = (\hat{\Sigma}_p - \hat{\Sigma}) + (\hat{\Sigma} - \Sigma)$ for non-lag-truncated estimator $\hat{\Sigma}$ of Σ. Since $\sqrt{T}\|\hat{\Sigma}_p - \hat{\Sigma}\| = o_p(1)$ (Hannan and Kavalieris, 1986), then $\sqrt{T}\text{vech}\{\hat{\Sigma}_p - \Sigma\} \xrightarrow{d} N(0, \Omega_\Sigma)$ as proven in Lütkepohl and Poskitt (1991), p. 494. Thus to generalize the lemma to the ridge estimator it suffices to show that $\sqrt{T}\|\hat{\Sigma}^R_p - \hat{\Sigma}_p\| = o_p(1)$.

Write $\widetilde{\Gamma}_p = T^{-1}\sum_{t=1}^{T} y_{t,p+1}y'_{t,p+1}$, and let $\hat{A}(p)$ and $\hat{\Sigma}_p$ be defined for least squares similarly to $\hat{A}^R(p)$ and $\hat{\Sigma}^R_p$, respectively,

$$\begin{aligned}\hat{\Sigma}^R_p - \hat{\Sigma}_p = {}& \hat{A}^R(p)\widetilde{\Gamma}_p(\hat{A}^R(p) - \hat{A}(p)) + (\hat{A}^R(p) - \hat{A}(p))\widetilde{\Gamma}_p\hat{A}^R(p) + \\ & + (\hat{A}^R(p) - \hat{A}(p))\widetilde{\Gamma}_p(\hat{A}^R(p) - \hat{A}(p))\end{aligned} \tag{58}$$

Then, with the same expressions of $\hat{B}^R_{p,j}(\Lambda^\ell)$ and $\hat{B}_{p,j}$ as in the proof of Theorem 5,

$$\begin{aligned}\sqrt{T}\|\hat{A}^R(p) - \hat{A}(p)\| &= \sqrt{T}\|\hat{\Gamma}^1_p\left((\hat{\Gamma}_p + (T-p)^{-1}\Lambda^\ell)^{-1} - \hat{\Gamma}^{-1}_p\right)\| && (59)\\ &\leq \|\hat{\Gamma}^1_p\| \cdot \|((\hat{\Gamma}_p \cdot T^{-1/2}(T-p)(\Lambda^\ell)^{-1} + T^{-1/2}I)^{-1}\| && (60)\\ &= o_p(1) && (61)\end{aligned}$$

since $\|\hat{\Gamma}^1_p\|$ is bounded above in probability (Lewis and Reinsel (1985), p. 397) and $\sqrt{T}(T-p)^{-1}\|\Lambda^\ell\| \xrightarrow{p} \infty$ by assumptions (i)-(iii), Theorem 6. By taking the term-by-term norm upper bound of $\|\hat{\Sigma}^R_p - \hat{\Sigma}_p\|$ from the above expansion the claim is proven. □

6.3 Sieve Inference Simulation

Data-generating Process. The VARMA(1,1) model is taken from Kilian and Kim (2011),

$$y_t = A_1 y_{t-1} + \epsilon_t + M_1\epsilon_{t-1}$$

where

$$A_1 = \begin{bmatrix} 0.5417 & -0.1971 & -0.9395 \\ 0.04 & 0.9677 & 0.0323 \\ -0.0015 & 0.0829 & 0.8080 \end{bmatrix}, \qquad M_1 = \begin{bmatrix} -0.1428 & -1.5133 & -0.7053 \\ -0.0202 & 0.0309 & 0.1561 \\ 0.0227 & 0.1178 & -0.0153 \end{bmatrix}$$

and $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, PP')$ where

$$P = \begin{bmatrix} 9.2325 & 0 & 0 \\ -1.4343 & 3.6070 & 0 \\ -0.7756 & 1.2296 & 2.7555 \end{bmatrix}$$

Regularizer Selection. For a VAR(p) and a lag-adapted Λ^{ℓ}, a collection $\{\lambda_1, \ldots, \lambda_p\}$ must be chosen. To implement OOS validation and CV for the ridge estimators, I rely on MATLAB optimization routines, in both cases using the optimization function `patternsearch` from the MATLAB Optimization Toolbox. The domain of optimization is chosen to be the hypercube $[0, 10^4]^p$.

There are two reasons behind the choice of a constrained maximization problem for setting Λ. First, an unconstrained optimization solution does not guarantee that the RLS estimator can be numerically computed: the issue is that for λ_i sufficiently large, the matrix $(ZZ' \otimes I_K + \Lambda^{\ell})$ can not be numerically inverted without running into precision issues. Secondly, executing a (global) optimizer over an unbounded optimization domain could potentially stall by diverging away at least over some dimensions, leading to unfeasible Monte Carlo simulations times.

Due to practical computational constraints, I keep the MATLAB limit of a maximum 3000 functional evaluation for optimization convergence. This is because, in general, there is no guarantee that the high dimensional OOS/CV objective function is concave. Thus a time limit must be enforced to avoid some MC simulations stalling. In applications, it is suggested to always employ advanced optimization routines, e.g. genetic or pattern-based optimizers like `patternsearch`, if possible. When one only requires to estimate the VAR model once, then the selection of Λ is a one-time cost. The gains of better optimization solutions therefore are often superior to the higher computational costs one incurs in when using more sophisticated routines. However, in selecting Λ one must also be careful in checking that a given cross- or out-of-sample validated solution does not produce numerical issues when computing RLS estimates using the closed-form formula, which entails a matrix inversion operation.

Parameter selection with many lags. The minimization problem of OOS/CV losses suffers from the curse of dimensionality. This is somehow effectively mitigated when using a lag-adapted Λ^{ℓ} regularizer since the loss depends only on p (non-negative) parameters, rather then $K^2 p$. But the issue remains whenever p is chosen large, as it happens in the case of sieve VAR modeling. In the simulation of Section 4.1, $p = H = 30$ is already a large enough value to make OSS/CV methods hard to optimize.

I use a shortcut to make computation more efficient. Such simplification stems from the following observation. Since by design the data-generating process is VARMA(1,1), the associated VAR(∞) coefficients will be geometrically decaying already after the first lag. Therefore after the first few lags penalization can be quite harsh whilst still having little bias. This is precisely reflected in Figure 2, too. The shortcut then is to use OOS/CV to estimate $\{\lambda_1, \ldots, \lambda_r\}$ where $r < p$, then extrapolate and use $\{\lambda_1, \ldots, \lambda_{r-1}, \lambda_r, \ldots, \lambda_r\}$, where λ_r is repeated $p - r$ times, to penalize a VAR(p) ridge estimator. This strategy is not generally appropriate, because it could be that even at relatively deep lags some coefficients are large, thus an excessive amount of bias could be introduced. Therefore in practice caution must be taken in applying such shortcut, and it is in fact preferable to avoid it.

In the simulations of Section 4.1, for sieve RLS estimation of VAR(30) models OOS/CV methods exploit the shortcut above with $r = 10$.

References

Andrews, D. W. K. (1984). Non-Strong Mixing Autoregressive Processes. *Journal of Applied Probability*, 21(4):930–934.

Andrews, D. W. K. (1985). A Nearly Independent, but Non-Strong Mixing, Triangular Array. *Journal of Applied Probability*, 22(3):729–731.

Babii, A., Ghysels, E., and Striaukas, J. (2020). Machine Learning Time Series Regressions with an Application to Nowcasting. *arXiv:2005.14057 [econ, stat]*.

Bao, Y. (2007). The Approximate Moments of the Least Squares Estimator for the Stationary Autoregressive Model under a General Error Distribution. *Econometric Theory*, 23(05):1013.

Barnichon, R. and Brownlees, C. (2019). Impulse Response Estimation by Smooth Local Projections. *The Review of Economics and Statistics*, 101(3):522–530.

Bauer, F. and Lukas, M. A. (2011). Comparing parameter choice methods for regularization of ill-posed problems. *Mathematics and Computers in Simulation*, 81(9):1795–1841.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *Q. J. Econ.*, 120(1):387–422.

Bhatia, R. (1997). *Matrix Analysis*. Springer, New York, NY, New York, NY, USA.

Boshnakov, G. N. and Iqelan, B. M. (2009). Generation Of Time Series Models With Given Spectral Properties. *Journal of Time Series Analysis*, 30(3):349–368.

Boubacar Mainassara, Y. and Francq, C. (2011). Estimating structural VARMA models with uncorrelated but non-independent error terms. *Journal of Multivariate Analysis*, 102(3):496–505.

Brüggemann, R., Jentsch, C., and Trenkler, C. (2016). Inference in VARs with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, 191(1):69–85.

Coulombe, P. G. (2020). Time-Varying Parameters as Ridge Regressions. *arXiv:2009.00401 [econ, stat]*.

Coulombe, P. G., Leroux, M., Stevanovic, D., and Surprenant, S. (2020). How is Machine Learning Useful for Macroeconomic Forecasting? *arXiv:2008.12477 [econ, stat]*.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.

De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328.

Engsted, T. and Pedersen, T. Q. (2014). Bias-Correction in Vector Autoregressive Models: A Simulation Study. *Econometrics*, 2(1):45–71.

Fu, W. and Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Ghosh, S., Khare, K., and Michailidis, G. (2019). High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models. *Journal of the American Statistical Association*, 114(526):735–748.

Giannone, D., Lenza, M., and Primiceri, G. E. (2018). Economic Predictions with Big Data: The Illusion of Sparsity. SSRN Scholarly Paper ID 3166281, Social Science Research Network, Rochester, NY.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hannan, E. J. (1970). *Multiple Time Series*. Wiley, New York.

Hannan, E. J. and Kavalieris, L. (1986). Regression, Autoregression Models. *Journal of Time Series Analysis*, 7(1):27–49.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82.

Hsu, N.-J., Hung, H.-L., and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657.

Inoue, A. and Kilian, L. (2002). Bootstrapping Smooth Functions of Slope Parameters and Innovation Variances in VAR(∞) Models*. *International Economic Review*, 43(2):309–331.

Inoue, A. and Kilian, L. (2008). How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation. *Journal of the American Statistical Association*, 103(482):511–522.

Jordà, Ò. (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, 95(1):161–182.

Kilian, L. (1998). Small-sample Confidence Intervals for Impulse Response Functions. *The Review of Economics and Statistics*, 80(2):218–230.

Kilian, L. and Kim, Y. J. (2011). How Reliable Are Local Projection Estimators of Impulse Responses? *Review of Economics and Statistics*, 93(4):1460–1466.

Kilian, L. and Lütkepohl, H. (2017). *Structural vector autoregressive analysis*. Cambridge University Press.

Krampe, J., Paparoditis, E., and Trenkler, C. (2020). Impulse Response Analysis for Sparse High-Dimensional Time Series. *arXiv:2007.15535 [stat]*.

Lewis, R. and Reinsel, G. C. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis*, 16(3):393–411.

Li, D., Plagborg-Møller, M., and Wolf, C. K. (2021). Local projections vs. vars: Lessons from thousands of dgps. *Working Paper.* Preliminary and incomplete - comments welcome!Matlab code (GitHub).

Li, J. and Liao, Z. (2020). Uniform nonparametric inference for time series. *Journal of Econometrics*, page 14.

Litterman, R. B. (1985). Forecasting with Bayesian vector autoregressions five years of experience. Working Papers 274, Federal Reserve Bank of Minneapolis.

Lukas, M. A. (2006). Robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems*, 22(5):1883.

Lukas, M. A. (2008). Strong robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems*, 24(3):034006.

Lukas, M. A. (2010). Robust gcv choice of the regularization parameter for correlated data. *The Journal of integral equations and applications*, pages 519–547.

Lütkepohl, H. (1988). Asymptotic Distribution of the Moving Average Coefficients of an Estimated Vector Autoregressive Process. *Econometric Theory*, 4(1):77–85.

Lütkepohl, H. (1990). Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models. *The Review of Economics and Statistics*, 72(1):116–125.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis.* Springer.

Lütkepohl, H. and Poskitt, D. S. (1991). Estimating Orthogonal Impulse Responses via Vector Autoregressive Models. *Econometric Theory*, 7(4):487–496.

Marriott, F. H. C. and Pope, J. A. (1954). Bias in the Estimation of Autocorrelations. *Biometrika*, 41(3/4):390–402.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., and Zilberman, E. (2021). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Mikusheva, A. (2007). Uniform Inference in Autoregressive Models. *Econometrica*, 75(5):1411–1452.

Mikusheva, A. (2012). One-Dimensional Inference in Autoregressive Models With the Potential Presence of a Unit Root. *Econometrica*, 80(1):173–212.

Montiel Olea, J. L. and Plagborg-Møller, M. (2020). Local projection inference is simpler and more robust than you think. *Working Paper.*

Nankervis, J. C. and Savin, N. E. (1988). The exact moments of the least-squares estimator for the autoregressive model corrections and extensions. *Journal of Econometrics*, 37(3):381–388.

Nicholls, D. F. and Pope, A. L. (1988). Bias in the Estimation of Multivariate Autoregressions. *Australian Journal of Statistics*, 30A(1):296–309.

Olea, J. L. M. and Plagborg-Møller, M. (2020). Local Projection Inference is Simpler and More Robust Than You Think. *arXiv:2007.13888 [econ].*

Park, J. Y. and Phillips, P. C. (1988). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory*, 4(3):468–497.

Park, J. Y. and Phillips, P. C. (1989). Statistical inference in regressions with integrated processes: Part 2. *Econometric Theory*, pages 95–131.

Pesavento, E. and Rossi, B. (2006). Small-sample confidence intervals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics*, 21(8):1135–1155.

Phillips, P. C. B. (1988). Regression Theory for Near-Integrated Time Series. *Econometrica*, 56(5):1021–1043.

Plagborg-Møller, M. (2016). *Essays in Macroeconometrics.* PhD thesis, Harvard University.

Plagborg-Møller, M. and Wolf, C. K. (2021). Local projections and vars estimate the same impulse responses. *Econometrica.* Forthcoming.

Poignard, B. (2018). Asymptotic theory of the adaptive Sparse Group Lasso. *Annals of the Institute of Statistical Mathematics*, 72(1):297–328.

Pope, A. L. (1990). Biases of Estimators in Multivariate Non-Gaussian Autoregressions. *Journal of Time Series Analysis*, 11(3):249–258.

Ramey, V. (2016). Macroeconomic Shocks and Their Propagation. In *Handbook of Macroeconomics*, volume 2, pages 71–162. Elsevier.

Sawa, T. (1978). The exact moments of the least squares estimator for the autoregressive model. *Journal of Econometrics*, 8(2):159–172.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1):1–48.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica: Journal of the Econometric Society*, pages 113–144.

Smeekes, S. and Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting*, 34(3):408–430.

Song, S. and Bickel, P. J. (2011). Large Vector Auto Regressions. *arXiv:1106.3915 [q-fin, stat].*

Stock, J. H. and Watson, M. W. (2016). Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In Taylor, J. B. and Uhlig, H., editors, *Handbook of Macroeconomics*, volume 2, pages 415–525. Elsevier.

Stock, J. H. and Watson, M. W. (2017). Twenty Years of Time Series Econometrics in Ten Pictures. *Journal of Economic Perspectives*, 31(2):59–86.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198.

van Wieringen, W. N. (2020). Lecture notes on ridge regression. *arXiv:1509.09169 [stat]*.

Whitney, H. (1972). *Complex Analytic Varieties*, volume 131. Addison-Wesley Reading.

Wiener, N. and Masani, P. (1958). The prediction theory of multivariate stochastic processes, II: The linear predictor. *Acta Mathematica*, 99:93–137.

Supplementary Appendix

A Preliminaries

A.1 LS and RLS Estimators.

Lütkepohl (2005), Chapter 3, shows that the multivariate least squares and GLS estimator of parameter vector $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = ((Z'Z)^{-1}Z \otimes I_K)\mathbf{y}$$

as the minimizer of $S(\boldsymbol{\beta}) = \mathrm{tr}[(Y - BZ)'\Sigma_u(Y - BZ)]$. The multivariate ridge-regularized least squares (RLS) – or, simply, ridge – estimator considered in this paper is defined to be the minimizer of the regularized problem,

$$\begin{aligned} S^R(\boldsymbol{\beta}; \Lambda) &= \mathrm{tr}[(Y - BZ)'(Y - BZ)] + \mathrm{tr}[B'\Lambda B] \\ &= \mathbf{y}'\mathbf{y} + \boldsymbol{\beta}'(ZZ' \otimes I_K)\boldsymbol{\beta} - 2\boldsymbol{\beta}'(Z \otimes I_K)\mathbf{y} + \boldsymbol{\beta}'\Lambda\boldsymbol{\beta} \end{aligned}$$

The first partial derivative,

$$\frac{\partial S^R(\boldsymbol{\beta}; \Lambda)}{\partial \boldsymbol{\beta}} = 2(ZZ' \otimes I_K)\boldsymbol{\beta} - 2(Z \otimes I_K)\mathbf{y} + 2\Lambda\boldsymbol{\beta}$$

gives the normal equations $(ZZ' \otimes I_K + \Lambda)\boldsymbol{\beta} = (Z \otimes I_K)\mathbf{y}$. The Hessian $\partial^2 S^R(\boldsymbol{\beta})/\partial^2\boldsymbol{\beta} = 2(ZZ' \otimes I_K + \Lambda)\boldsymbol{\beta}$ is positive definite when $\Lambda > 0$, thus indeed the minimum is achieved by

$$\hat{\boldsymbol{\beta}}^R(\Lambda) = (ZZ' \otimes I_K + \Lambda)^{-1}(Z \otimes I_K)\mathbf{y}$$

Identical derivations prove that re-centering the ridge penalty at $\boldsymbol{\beta}_0 \in \mathbb{R}^{K^2 p}$ produces the estimator

$$\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) = (ZZ' \otimes I_K + \Lambda)^{-1}((Z \otimes I_K)\mathbf{y} + \Lambda\boldsymbol{\beta}_0)$$

A.2 Structure of the Regularization Matrix

The vectorized RLS estimator $\hat{\boldsymbol{\beta}}^R(\Lambda)$ has maximal flexibility in terms of the regularization structure that matrix $\Lambda = \mathrm{diag}\{\lambda_{1,1}, \ldots, \lambda_{K,p}\}$ $(K^2 p \times K^2 p)$ imposes. Since $\boldsymbol{\beta}$ contains all the coefficients of $(A_1, \ldots, A_p)$ it is indeed possible to individually penalize each lag of each series differently. In fact, by relaxing the assumption that Λ be a diagonal matrix, even more general penalization structures are possible, although I do not consider them in this paper.

An interesting special case arises if the RLS estimator is instead written in its matrix form[14],

$$\hat{B}^R(\Lambda_{Kp}) = YZ'(ZZ' + \Lambda_{Kp})^{-1}$$

[14] For details in the least squares case, see again Lütkepohl (2005), Chapter 3. The derivations for the ridge estimator are identical.

where here it is of note that $\Lambda_{Kp} > 0$ has size $(Kp \times Kp)$. The regularization structure imposed is different in general than that in $\hat{\boldsymbol{\beta}}^R(\Lambda)$: Λ_{Kp} induces *column-wise* ridge regularization, which penalizes coefficient estimates uniformly over each of the Kp columns of B. The associated vectorized estimator then simplifies:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^R(\Lambda_{Kp}) &= ((ZZ' + \Lambda_{Kp}) \otimes I_K)^{-1}(Z \otimes I_K)\mathbf{y} \\ &= ((ZZ' + \Lambda_{Kp})^{-1} Z \otimes I_K)\mathbf{y}\end{aligned}$$

On the other hand, the *de-vectorized* RLS estimator is given by

$$\hat{B}^R_{de}(\Lambda_{K^2p}) = \text{reshape}(\boldsymbol{\beta}^R(\Lambda), K, Kp)$$

that is, $\hat{B}^R_{de}$ is simply a restructuring of the vectorized estimator into a matrix with identical dimensions to B. Importantly then, $\hat{B}^R_{de}(\Lambda_{K^2p})$ is equivalent to $\hat{B}^R(\Lambda_{Kp})$ if $\Lambda_{K^2p} = \Lambda_{Kp} \otimes I_K$. Because $\boldsymbol{\beta}^R(\Lambda_{K^2p})$ and $\hat{B}^R_{de}(\Lambda_{K^2p})$ allow for the most generality, I will consider them the "standard" RLS estimators, so the dimension subscript to Λ will be dropped unless explicitly required.

A.3 Autocovariance and Asymptotic Conditioning

The conditioning of the autocovariance matrix $\Gamma = \mathbb{E}(y_t y_t')$ is an important measure for the role that the regularization in the RLS estimator should be playing. This in turn depends on the eigenvalues of Γ_T with respect to those of Γ. Hoerl and Kennard (1970) showed in the linear regression setting that, when the sample covariance matrix deviates significantly from the identity matrix, its small eigenvalues excessively inflate the variance of least squares estimates, even though the regression problem itself is well-posed. This fragility is inherently a byproduct of finite sampling, and partially due to numerical procedures. Nowadays, numerical precision is virtually not a concern anymore, as robust linear algebra procedures are implicitly implemented in most scientific languages and toolboxes. Yet estimation issues tied to small or unfavorable data samples remain extremely relevant from both theoretical and practical viewpoints.

In the spirit of ridge as a regularization procedure, the following Lemma establishes convergence in probability of the ordered eigenvalues of the sample autocovariance matrix.

Lemma A.1. *If* $\hat{\Gamma}_T = T^{-1} \sum_{t=1}^{T-1} y_t y_t' \xrightarrow{p} \Gamma$ *where* $\Gamma \in \mathbb{R}^{K \times K}$ *is positive definite, then*

$$\omega_j\left(\hat{\Gamma}_T\right) \xrightarrow{p} \omega_j(\Gamma)$$

where $\omega_j(A)$ *is the* j *largest eigenvalue of* A*.*

Proof. First, recall that for all matrices $A \in \mathbb{R}^{K \times K}$, the determinant $\det(A)$ is clearly a continuous mapping[15]. Furthermore, for any polynomial $g(z) = z^n + a_1 z^{n-1} + \ldots + a_n$, $a_i \in \mathbb{C}$ factored

[15] This follows from $\det(A_{i,j}) = \sum_{\varsigma} \text{sgn}(\varsigma) \prod_{i=1}^{K} A_{\varsigma(i),i}$ for ς permutation over $\{1, \ldots, K\}$

as $g(z) = (z - w_1)\cdots(z - w_n)$, $w_i \in \mathbb{C}$, where the ordering of roots w_i is arbitrary, it holds that for any $\epsilon > 0$ there exists $\delta > 0$ such that for every polynomial $h(z) = z^n + b_1 z^{n-1} + \ldots + b_n$ with $|a_i - b_i| < \delta$ decomposed as $g(z) = (z - \overline{w}_1)\cdots(z - \overline{w}_n)$, $|w_i - \overline{w}_i| < \epsilon$, $i = 1, \ldots, n$, see Whitney (1972), Appendix V.4. This in particular implies that the roots of the characteristic polynomial of matrix A are continuous functions of its coefficients.

Let $\varrho_{\hat{\Gamma}_T}(z) = z^K + a_1 z^{K-1} + \ldots + a_K = (z - \hat{\omega}_1)\cdots(z - \hat{\omega}_K)$ and $\varrho_{\Gamma}(z) = z^K + b_1 z^{K-1} + \ldots + b_K = (z - \omega_1)\cdots(z - \omega_K)$ be the (real) characteristic polynomials of $\hat{\Gamma}_T$ and Γ respectively. Because of the continuity arguments above, for every $\epsilon > 0$ there exist δ_1, $\delta_2 > 0$ such that

$$\mathbb{P}(|\hat{\omega}_i - \omega_i| > \epsilon) \leq \mathbb{P}(|a_i - b_i| > \delta_1) \tag{62}$$

$$\leq \mathbb{P}(\|\hat{\Gamma}_T - \Gamma\| > \delta_2) \tag{63}$$

for $i \in \{1, \ldots, K\}$. Since by assumption $\hat{\Gamma}_T \xrightarrow{p} \Gamma$, the RHS of the above converges to zero as $T \to \infty$, thus $\hat{\omega}_i \xrightarrow{p} \omega_i$. □

B Proofs

B.1 Ridge Shrinkage

Proof. Proposition 1. Since K is fixed, the submultiplicative and compatibility properties of the spectral norm, together with the notion that the Gram matrix ZZ' has nonnegative eigenvalues, imply

$$\|\hat{\boldsymbol{\beta}}^R(\Lambda)\| \leq \|(ZZ' \otimes I_K + \Lambda)\| \cdot \|(Z \otimes I_K)\boldsymbol{y}\| \tag{64}$$

$$\leq \operatorname{tr}\{(ZZ' \otimes I_K + \Lambda)^{-1}\} \cdot \|(Z \otimes I_K)\boldsymbol{y}\| \tag{65}$$

$$< \overline{\operatorname{tr}}(\Lambda) \cdot C_{\boldsymbol{y}} \tag{66}$$

where $\overline{\operatorname{tr}}(\Lambda) = \left(\sum_{i=1}^{K^2 p} \lambda_i^{-1}\right)$. The last inequality follows from $\lambda_i > 0$ and the application to $(ZZ' \otimes I_K + \Lambda)$ of Weyl's inequalities, see Theorem III.2.1 in Bhatia (1997). In the above $C_{\boldsymbol{y}}$ is a random variable that does not depend on Λ.

By Chebychev's inequality and under Assumptions A and B, for $\epsilon > 0$

$$\mathbb{P}(\|\hat{\boldsymbol{\beta}}^R(\Lambda)\| > \epsilon) \leq \mathbb{P}\left(\overline{\operatorname{tr}}(\Lambda)\, C_{\boldsymbol{y}} > \epsilon\right) \tag{67}$$

$$\leq \frac{1}{\epsilon}\, \mathbb{E}\left[\overline{\operatorname{tr}}(\Lambda)\, C_{\boldsymbol{y}}\right] \tag{68}$$

$$\leq \frac{1}{\epsilon}\, \overline{\operatorname{tr}}(\Lambda)\, \underbrace{\mathbb{E}\left[C_{\boldsymbol{y}}\right]}_{C_1} \tag{69}$$

where $C_1 < \infty$ is a constant. A proof of this is simple to obtain given that $C_{\boldsymbol{y}}$ is a function of fourth moments of y_t, and can be found in Lemma 2 below. Since the sequence $\Lambda_1 \preceq \Lambda_2$ implies $\overline{\operatorname{tr}}(\Lambda_2) < \overline{\operatorname{tr}}(\Lambda_1)$, applying the last inequality to $\hat{\boldsymbol{\beta}}^R(\Lambda_1)$ and $\hat{\boldsymbol{\beta}}^R(\Lambda_2)$ yields

$$\mathbb{P}(\|\hat{\boldsymbol{\beta}}^R(\Lambda_2)\| > \epsilon) - \mathbb{P}(\|\hat{\boldsymbol{\beta}}^R(\Lambda_1)\| > \epsilon) \leq \frac{1}{\epsilon}\left(\overline{\operatorname{tr}}(\Lambda_2) - \overline{\operatorname{tr}}(\Lambda_1)\right) C_1 < 0 \tag{70}$$

□

Lemma 2. *Under assumptions A-B, it holds that*

$$\mathbb{E}[\|(Z \otimes I_K)\boldsymbol{y}\|] < \infty$$

Proof. First one rewrites the random variable above as

$$(Z \otimes I_K)\boldsymbol{y} = \text{vec}\{YZ'\} = \text{vec}\left\{\sum_{t=0}^{T-1} y_{t+1}Y_t\right\} \tag{71}$$

where $Y_t = (y_t', y_{t-1}', \ldots, y_{t-p+1}')$. Therefore to bound $\|(Z \otimes I_K)\boldsymbol{y}\|$ is to bound $\|YZ'\|_F$, which can be done transparently as follows. The Frobenius norm can be expanded to

$$\begin{aligned}
\|YZ'\|_F^2 &= \left\|\sum_{t=0}^{T-1} y_{t+1}(y_t', y_{t-1}', \ldots, y_{t-p+1}')\right\|_F^2 && (72)\\
&\leq \sum_{t=0}^{T-1} \|y_{t+1}(y_t', y_{t-1}', \ldots, y_{t-p+1}')\|_F^2 && (73)\\
&= \sum_{t=0}^{T-1}\sum_{i=0}^{p-1} \|y_{t+1}\, y_{t-i}'\|_F^2 && (74)\\
&= \sum_{t=0}^{T-1}\sum_{i=0}^{p-1}\sum_{n=1}^{K}\sum_{m=1}^{K} (y_{t+1,n}\, y_{t-i,m})^2 && (75)
\end{aligned}$$

and under Assumptions A-B the fourth moment $\mathbb{E}[(y_{t+1,n}\, y_{t-i,m})^2]$ is finite (cf. Hamilton (1994), Proposition 7.10). Jensen's inequality then yields

$$\mathbb{E}[\|(Z \otimes I_K)\boldsymbol{y}\|] \leq \left(\mathbb{E}\left[\|YZ'\|_F^2\right]\right)^{\frac{1}{2}} \leq \left(\sum_{t=0}^{T-1}\sum_{i=0}^{p-1}\sum_{n=1}^{K}\sum_{m=1}^{K} \mathbb{E}\left[(y_{t+1,n}\, y_{t-i,m})^2\right]\right)^{\frac{1}{2}} < \infty \tag{76}$$

□

Proof. PROPOSITION 2. Notice that any lag-adapted regularization matrix can be written as $\Lambda^\ell = \Lambda_p \otimes I_{K^2} = (\Lambda_p \otimes I_K) \otimes I_K$, so that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^R(\Lambda^\ell) &= ((ZZ' + \Lambda_p \otimes I_K) \otimes I_K)^{-1}(Z \otimes I_K)\mathbf{y} && (77)\\
&= ((ZZ' + \Lambda_p \otimes I_K)^{-1} \otimes I_K)(Z \otimes I_K)\mathbf{y} && (78)
\end{aligned}$$

Then, by the properties of the Kronecker product,

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}^R(\Lambda)_s\| &= \|((ZZ' + \Lambda_p \otimes I_K)^{-1} \otimes I_K)(Z \otimes I_K)\mathbf{y}\| && (79)\\
&\leq \|((ZZ' + \Lambda_p \otimes I_K)^{-1} \otimes I_K)\| \cdot \|(Z \otimes I_K)\mathbf{y}\| && (80)\\
&\leq \text{tr}\{(ZZ' + \Lambda_p \otimes I_K)^{-1}\} \cdot C_0 && (81)\\
&\leq \sum_{i=1}^{Kp} \lambda_i^{-1} \cdot C_0 && (82)
\end{aligned}$$

where $C_0 > 0$ is a constant not dependent on Λ_p, and the last step follows again from Weyl's inequalities.

(i) The assumption $\Lambda_1^\ell \preceq \Lambda_2^\ell$ implies by definition that $\lambda_{1,i} < \lambda_{2,i}$ for at least some $i \in \{1, \dots, K^2p\}$. Then

$$\|\hat{\boldsymbol{\beta}}^R(\Lambda_2^\ell)\| - \|\hat{\boldsymbol{\beta}}^R(\Lambda_1^\ell)\| \leq \sum_{i=1}^{Kp} \left[\lambda_{2,j}^{-1} - \lambda_{1,j}^{-1}\right] \cdot C_0 < 0 \tag{83}$$

By constructing an appropriate selection matrix[16] $R(s)$ it is immediate to show that $\|\hat{\boldsymbol{\beta}}^R(\Lambda)_{[\mathcal{S}]}\| = \|R(s)\hat{\boldsymbol{\beta}}^R(\Lambda)\|$. Since $\|R(s)\| = 1$ it follows again

$$\begin{aligned} \|\hat{\boldsymbol{\beta}}^R(\Lambda)_s\| &\leq \|R(s)\| \cdot \|((ZZ' + \Lambda_p \otimes I_K)^{-1} \otimes I_K)\| \|(Z \otimes I_K)\mathbf{y}\| && (84) \\ &\leq \sum_{i=1}^{K^p} \lambda_i^{-1} \cdot C_0 && (85) \end{aligned}$$

therefore one may use the same inequality applied above when $\mathcal{S} = \{1, \dots, p\}$.

(ii) The result concerning the isotropic regularizer $\Lambda^\ell = \lambda \otimes I_{K^2p}$ is trivial given (i).

(iii) Without loss of generality due to the ordering of lags in Z, one may write the Gram matrix ZZ' in a block fashion,

$$ZZ' + \Lambda_p = \begin{bmatrix} (ZZ')_{[\mathcal{S}]} + \Lambda_{[\mathcal{S}]} & D \\ D' & (ZZ')_{[-\mathcal{S}]} + \Lambda_{[-\mathcal{S}]} \end{bmatrix} \tag{86}$$

where $(ZZ')_{[-\mathcal{S}]}$ is the sub-matrix containing all the components *not* indexed by subset $\mathcal{S}$, and the subscript has been dropped from Λ_p for ease of notation.
Define $A_{[\mathcal{S}]} = (ZZ')_{[\mathcal{S}]} + \Lambda_{[\mathcal{S}]}$, $B_{[-\mathcal{S}]} = (ZZ')_{[-\mathcal{S}]} + \Lambda_{[-\mathcal{S}]}$ and $\varDelta = (B_{[-\mathcal{S}]} - D'A_{[\mathcal{S}]}^{-1}D)$. The matrix block-inversion formula yields

$$(ZZ' + \Lambda_p)^{-1} = \begin{bmatrix} A_{[\mathcal{S}]}^{-1} + A_{[\mathcal{S}]}^{-1} D \varDelta^{-1} D' A_{[\mathcal{S}]}^{-1} & A_{[\mathcal{S}]}^{-1} D \varDelta^{-1} \\ -\varDelta^{-1} D' A_{[\mathcal{S}]}^{-1} & \varDelta^{-1} \end{bmatrix} \tag{87}$$

If $\Lambda_{[\mathcal{S}]} \to 0$ and $\Lambda_{[-\mathcal{S}]} \to \infty$, then $A_{[\mathcal{S}]} \to (ZZ')_{[\mathcal{S}]}$, $B_{[-\mathcal{S}]} \to \infty$. Therefore $\varDelta^{-1} \to 0$, since for $\Lambda_{[-\mathcal{S}]}$ sufficiently large $\|B_{[-\mathcal{S}]}^{-1} D' A_{[\mathcal{S}]}^{-1} D\| < 1$ and thus the Sherman-Morrison-Woodbury formula implies

$$\|(B_{[-\mathcal{S}]} - D'A_{[\mathcal{S}]}^{-1}D)^{-1}\| \leq \frac{\|B_{[\mathcal{S}]}^{-1}\|}{1 - \|B_{[-\mathcal{S}]}^{-1} D' A_{[\mathcal{S}]}^{-1} D\|} \to 0 \tag{88}$$

The above results finally give

$$((ZZ' + \Lambda_p)^{-1} Z \otimes I_K)\boldsymbol{y} \to \begin{bmatrix} (ZZ')_{[\mathcal{S}]} & 0 \\ 0 & 0 \end{bmatrix} (Z \otimes I_K)\boldsymbol{y} = \hat{\boldsymbol{\beta}}^{LS}(\mathcal{S}) \tag{89}$$

□

[16] The *selection* matrix for vector v $(K^2p \times 1)$ and index $s \in \{1, \dots, p\}$ is the $(K^2p \times K^2p)$ matrix $R(s) = \mathrm{diag}\{e_s\} \otimes I_{K^2}$, where e_s is the s canonical basis vector.

C Bias Simulation

While both plain and ridge regularized least squares estimators are biased when applied to linear time series models, a priori there is no saying whether one dominates the other in terms of finite-sample performance with respect to bias. A complication comes from the fact the RLS is not a single estimator, but indeed, for any given sample, one has to consider the family[17]

$$\mathcal{R}(Y) = \{\hat{\boldsymbol{\beta}}^R(\Lambda) \mid \Lambda = \operatorname{diag}\{\lambda_1, \dots, \lambda_{K^2p}\},\ \lambda_i \in \mathbb{R}^+\}$$

For a fixed sample then, it might be the case that

$$\inf_{\hat{\beta}^R \in \mathcal{R}(Y)} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R\| < \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|$$

If $\mathcal{R}(Y)$ is instead considered as a family of random variables, then this leads to the more general question of whether the mechanism of shrinkage gives the RLS estimator uniform advantage over the LS estimator

$$\inf_{\hat{\beta}^R \in \mathcal{R}(Y)} \mathbb{E}[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R\|] < \mathbb{E}[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|]$$

Such a result would imply that for any sample size T, it is always possible to construct an RLS estimator with lower bias than its least squares counterpart. The complexity of proving such a claim is apparent by noting that – in geometric terms – this amounts to showing that the infimum bias norm over the set of random variable $\mathcal{R}(Y)$ is never achieved by the LS estimator.

In order to give a more complete picture of the effects of ridge regularization and shrinkage on bias in small samples, I construct a simple Monte Carlo experiment following the one introduced in Kilian (1998). The original setting was also aimed at comparing the effects of bias in vector autoregression estimation for the purposes of inference, but it focused on comparing bootstrap schemes with respect to relative coverage. In the present case, I use the same setup design of Kilian (1998), but only focus on the immediate bias of the RLS estimator versus least squares. The data-generating process is a bivariate VAR(1) process $y_t = By_{t-1} + u_t$ with variable persistence controlled by entry B_{11} in the coefficient matrix B and i.i.d. Gaussian error term u_t with non-diagonal variance matrix. Because the main objective is to understand whether expanding the singleton of the LS estimator to the $\mathcal{R}(Y)$ family gives room for potential improvement (at least theoretically) I pit the textbook least squares VAR estimator (Lütkepohl, 2005) against the oracle-optimal RLS estimator $\hat{\boldsymbol{\beta}}^R(\lambda_{\text{opt}})$, where λ_{opt} is the minimizer of

$$\mathcal{B}(\lambda) = \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\lambda)\|$$

This is an oracle optimizer in the sense that in any applied setting the true parameter vector $\boldsymbol{\beta}$ is clearly unavailable.[18]

[17] Here for simplicity I focus on the ridge least squares without centering, but the family $\mathcal{R}$ easily generalizes to that given by $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$, although size and complexity grow significantly.

[18] This is a somehow unfair comparison with respect to least squares. It is pretty obvious that if one had access

Table 1: Average estimation bias comparison in small sample.

Sample Size	B_{11}	Bias($\hat{\boldsymbol{\beta}}$)	Bias($\hat{\boldsymbol{\beta}}^R(\lambda_{\text{opt}})$)	Avg λ_{opt}	Avg $\|\hat{\boldsymbol{\beta}}^R\|/\|\hat{\boldsymbol{\beta}}\|$	$q(\lambda_{\text{opt}} = 0)$
$T = 50$	-0.9	0.188	0.173	3.88	0.97	0
	-0.5	0.229	0.209	4.75	0.94	0
	0	0.252	0.217	8.92	0.9	0
	0.5	0.249	0.214	5.98	0.93	0
	0.9	0.244	0.21	4.61	0.96	0
	0.97	0.239	0.213	3.66	0.97	0
	1	0.245	0.223	3.17	0.97	0
$T = 100$	-0.9	0.123	0.113	5.14	0.98	0
	-0.5	0.157	0.146	5.78	0.96	0
	0	0.172	0.148	11.92	0.93	0
	0.5	0.17	0.145	8.68	0.95	0
	0.9	0.153	0.134	5.63	0.98	0
	0.97	0.155	0.136	5.3	0.98	0
	1	0.156	0.14	4.56	0.98	0

$B = 1000$ replications.

The results of this simulation are shown in Table 1. The first two columns present the experimental sample size and process specification; the second and third columns list the average bias for the LS and (oracle-optimal) RLS estimators respectively; the fourth column gives the average optimal λ; the fifth column lists the average norm shrinkage of the RLS estimates as a fraction of the least squares estimates norm; and the last column lists the fraction of simulations for which λ_{opt} was set to zero[19]. Two outcomes are of interest: first, it appears that the ridge family $\mathcal{R}(Y)$ contains estimators that improve on the expected bias with respect to textbook least squares. The gains are larger whenever y_t has lower persistence, suggesting that regularization can not be as effective in mitigating bias of near-unit-root processes. In all cases of B_{11}, the optimal ridge parameters is meaningfully different from zero, with λ_{opt} increasing monotonically as $|B_{11}|$ is reduced. This is an intuitive pattern, as one would expect that as the coefficients in matrix B become smaller, there is more leeway to shrink while reducing bias. Secondly, the fraction of times the optimal λ is set exactly zero[20], $q(\lambda_{\text{opt}} = 0)$, is always null, meaning that in this experimental setting the bias-optimal member of $\mathcal{R}(Y)$ is never the LS estimator.

to an oracle criterion, setting λ optimally would become a trivial task, at least when the concern is to minimize expected bias. As I discuss below, this is most surely not the case in practice, and in the regression setting there is a large literature on data-driven criteria to set the (uniform-scale) ridge parameter, see Bauer and Lukas (2011) for a general review.

[19] Further details on the implementation of the Monte Carlo simulation are given in Appendix C.1.

[20] Due to numerical precision and optimization, I consider λ_{opt} to be set to zero whenever $\lambda_{\text{opt}} < 10^{-6}$.

Table 2: Average estimation bias comparison in medium-large models.

Model Size	T	Bias($\hat{\boldsymbol{\beta}}$)	Bias($\hat{\boldsymbol{\beta}}^R(\lambda_{\text{opt}})$)	Avg λ_{opt}	Avg $\Vert\hat{\boldsymbol{\beta}}^R\Vert/\Vert\hat{\boldsymbol{\beta}}\Vert$	$q(\lambda_{\text{opt}}=0)$
$K = 15$	100	1.23	1.222	0.56	0.9937	0.05
	200	0.826	0.822	0.6	0.9969	0
	500	0.507	0.506	0.8	0.9984	0
$K = 50$	100	5.46	5.422	0.41	0.992	0
	200	3.091	3.085	0.38	0.9975	0
	500	1.768	1.767	0.38	0.9992	0

$B = 1000$ replications.

Unfortunately, it turns out that without adding more degrees of freedom to the regularizer the above (theoretical) "bias advantage" of the RLS estimator vanishes very quickly as soon as one increases the dimensionality of the VAR process under consideration. This conclusion comes from another simple but relevant simulation experiment that employs DGPs with $K \gg 2$. I make use of the procedure proposed in Boshnakov and Iqelan (2009) to generate VARIMA models of arbitrary dimension and spectral properties[21]. Their method for creating synthetic DGPs is especially useful, since producing large vector autoregressive models under the constraint of stability is especially complex when done manually. I again consider VAR(1) processes. I further simplify the variance matrix of the error term to be the identity. For this experiment, given the discussion regarding the applicability of RLS to highly persistent processes, I restrict the set of absolute eigenvalues of coefficient matrix A_1 to be $\{0.1, 0.3, 0.5, 0.7\}$. This is without much loss of generality, as one gathers from Table 1 that RLS should simply show a more prominent gain versus least squares. I consider a DGP of dimension $K = 15$, to approximate a medium-sized macroeconomic time series collection, and one with $K = 50$, a large (but not truly high-dimensional) system, which is representative of nowadays common estimation tasks. Finally, to allow comparability, sample sizes T are kept common between models, noting that least squares remains a feasible estimation technique in all cases. As evidenced by Table 2, even a relatively mild increase in the number of components has a significant impact on the gains that one might reap for uniform shrinkage in the RLS estimates. While again RLS does improve on LS estimates, and only in the case $K = 15$ and $T = 100$ the least squares estimate is selected 5% of the time, these improvements are nearly null. Given such slim margins, one should expect that in practice, when K is of realistic size and λ is chosen in a data-driven manner, the ridge estimator will be more biased than least squares.

[21] I directly rely on the `R` library '`mcompanion`' developed by the authors and available on `CRAN`.

C.1 Implementation

I follow Kilian (1998) in choosing a stationary bivariate VAR(1) process $y_t = (y_{1t}, y_{2t})$ given by

$$y_t = \begin{bmatrix} B_{11} & 0 \\ 0.5 & 0.5 \end{bmatrix} y_{t-1} + u_t, \qquad u_t \overset{\text{iid}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \right)$$

where $B_{11} \in \{-0.9, -0.5, 0, 0.5, 0.9, 0.97, 1\}$. Since $-1 < B_{11} \leq 1$ the process is stationary unless $B_{11} = 1$. For $B_{11} = 1$, the process is cointegrated, but this has no strong impact on the behavior of the estimator themselves because of the continuity of finite-sample distributions. Setting $B_{11} = 0$ implies a white noise process for the first component of y_t, which invalidates inference on impulse responses of y_{1t} due to a degenerate asymptotic covariance (Lütkepohl, 1990); this does not, however, adversely influence the parameter estimates themselves.

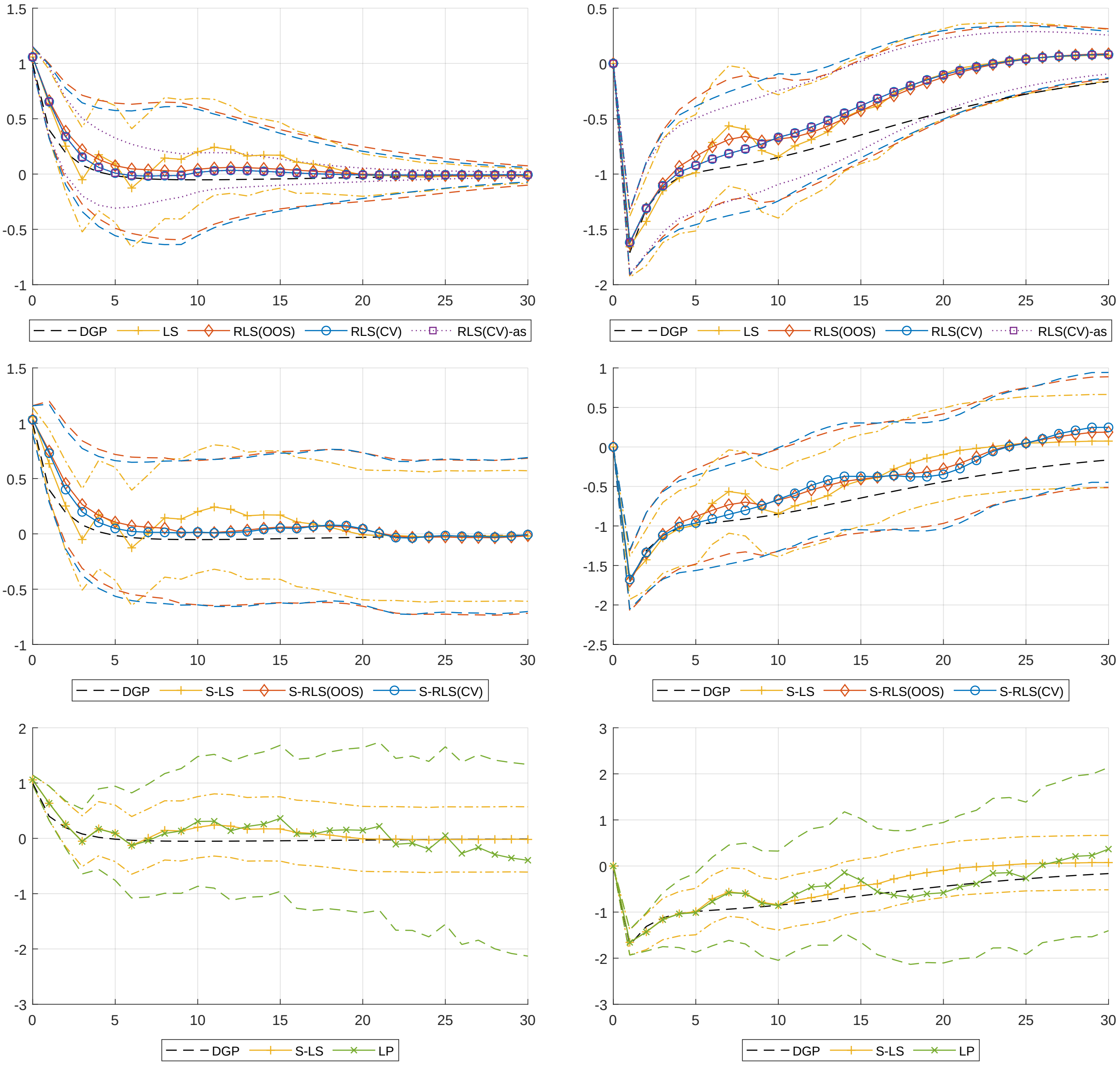
As remarked by Kilian (1998), under the given bivariate VAR process, sample sizes of $T = 50$ and $T = 100$ are realistic once contextualized in the most common application scenarios due to the geometric growth of the number of parameters with component size K. However, it is important to highlight that when regularization is employed, the number of components K influences estimation by more than just effecting the rate of increase in degrees-of-freedom (as a function of sample size). For all simulations I use $B = 1000$ Monte Carlo replications to draw samples from the specified DGP. For the RLS estimator, I consider the matrix estimator $\hat{B}^R$ with the uniform-scale ridge regularization matrix λI_{Kp}. To find the optimal regularizer λ_{opt} at every replication step, I minimize the oracle objective

$$\mathcal{B}(\lambda) = \|B - \hat{B}^R(\lambda I_{Kp})\|_F$$

where $\|\cdot\|_F$ is the Frobenius norm. I employ the MATLAB function `fmincon` using the standard '`interior-point`' optimization routine constrained over the interval $[0, 1000]$ and initial starting value 0.01. In all simulations the optimizer always converges within the constraint.

D Impulse Responses Illustration

Figure 3: Impulse responses and CIs of different estimators for a simulated sample.



Note: IRFs shown are Φ_{11} (left side panels) and Φ_{12} (right side panels). From top to bottom row: finite-order VAR(10) model, LS (yellow pluses), RLS(OOS) (red diamonds), RLS(CV) (blue dots) and RLS(CV)-as with asymptotic variance shrinkage (purple squares) estimates; sieve VAR(30) model, S-LS (yellow pluses), S-RLS(CV) (blue dots) and S-RLS(OOS) (red diamonds) estimates; sieve VAR(30) S-LS (yellow pluses) and LP (green crosses) estimates with $q = 10$ lags, Newey-West standard errors. Sample size $T = 300$.