JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **46**: 235–257 (2025)
Published online 18 February 2024 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12737

SPECIAL ISSUE ARTICLE

RIDGE REGULARIZED ESTIMATION OF VAR MODELS FOR INFERENCE

GIOVANNI BALLARIN
Department of Economics, University of Mannheim, Mannheim, Germany

Ridge regression is a popular method for dense least squares regularization. In this article, ridge regression is studied in the context of VAR model estimation and inference. The implications of anisotropic penalization are discussed, and a comparison is made with Bayesian ridge-type estimators. The asymptotic distribution and the properties of cross-validation techniques are analyzed. Finally, the estimation of impulse response functions is evaluated with Monte Carlo simulations and ridge regression is compared with a number of similar and competing methods.

Received 17 July 2023; Accepted 30 January 2024

Keywords: Impulse responses; inference; ridge regularization; vector autoregression.

JEL. C32; C51; C52.

MOS subject classification: 60G10; 60G25; 62F12; 62J07; 62M10; 62M20; 62P20.

1. INTRODUCTION

While the idea of using ridge regression for vector autoregressive model estimation dates back to Hamilton (1994), there seems to be no complete analysis of its properties and asymptotic theory in the literature. This article fills this gap by analyzing the geometric and distributional properties of ridge in a VAR estimation framework, discussing its comparison to well-known Bayesian approaches and deriving the validity of cross-validation as a selection procedure for the ridge penalty.

First, I show that the shrinkage induced by the ridge estimator, while intuitive in the setting of an isotropic penalty, produces complex effects when estimating a VAR model with a more flexible penalization scheme. This implies that the benefits of the bias-variance trade-off (Hastie, 2020) may be hard to gauge a priori. I provide a tractable example where ridge can yield estimates that have higher autoregressive dependence than the least squares solution. To better understand how different ridge penalization strategies can be designed, I also make a comparison with Bayesian VAR estimators commonly used in macroeconometric practice.

Second, I generalize the analysis of Fu and Knight (2000) and prove the consistency and asymptotic normality of the ridge estimator, a result that seems to be missing in the literature. For standard inference, the ridge penalty should either be negligible in the limit or its centering converge in probability to the true parameter vector. In both these cases, there is no asymptotic bias and no reduction in variance. Alternatively, in settings where a researcher is willing to assume that a subset of the VAR parameters features small coefficients, one can achieve an asymptotic reduction of variance by correctly tuning the ridge penalty matrix. I further derive the properties of cross-validation, which is a popular approach in practical applications to tune penalized estimators (Hastie *et al.*, 2009; Bergmeir *et al.*, 2018). More specifically, I show that cross-validation is able to select penalties that are asymptotically valid

*Correspondence to: Giovanni Ballarin, Department of Economics, University of Mannheim, L7, 3-5, Mannheim, 68131, Germany.
Email: giovanni.ballarin@gess.uni-mannheim.de

for inference. In passing, I also prove that, in an autoregressive setup, the time dependence of regressors has an exponentially small effect on in-sample prediction error evaluation.

Lastly, I use Monte Carlo simulations to study the performance of the different ridge approaches discussed, focusing on impulse response inference. I consider two exercises: one is based on a three-variable VARMA(1,1) data generating process from Kilian and Kim (2011); the other is a VAR(5) model estimated in levels from a set of seven macroeconomic series, following Giannone *et al.* (2015). The finding is that ridge can lead to improvements over unregularized methods in impulse response confidence interval length, while Bayesian estimators show the best overall performance due to the underlying flexibility of their priors.

1.1. Related Literature

This article does not discuss the high-dimensional setting, where the number of regressors grows together with the sample size. Some important work has been done in this direction already. Dobriban and Wager (2018) derive an explicit expression for the predictive risk of ridge regression assuming a high-dimensional random effects model. Other works in this vein are Liu and Dobriban (2020); Patil *et al.* (2021) and Hastie *et al.* (2022), which are mostly focused on penalty selection by cross-validation, as well as structural features of ridge. Generally speaking, the complexity of analyzing ridge regression in high dimensions is a challenge to precisely understanding its practical implications. As I show below, in the context of finite-dimensional VARs, asymptotic inference demands that the ridge penalty becomes asymptotically negligible at appropriate rates. Thus, a challenge is understanding in what way high-dimensional time series problems would benefit from the use of ridge. This question is beyond the scope of this article.

In the time series forecasting literature, ridge regression is commonly used for prediction. I provide a partial list of contributions in this direction. 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) use a Bayesian VAR with posterior mean equivalent to a ridge regression in forecasting; Ghosh *et al.* (2019) again study the Bayesian ridge, this time, however, in the high-dimensional context; Goulet Coulombe *et al.* (2022), Fuleky (2020), Babii *et al.* (2021), and Medeiros *et al.* (2021) compare LASSO, ridge and other machine learning techniques for forecasting with large economic datasets. Fuleky (2020) gives a textbook treatment of penalized time series estimation, including ridge, but does not discuss inference. The ridge penalty is considered within a more general mixed ℓ_1-ℓ_2 penalization setting in Smeekes and Wijler (2018), who study the performance and robustness of penalized estimates for constructing forecasts.

Regarding inference, Li *et al.* (2024) provided a general exploration of shrinkage procedures in the context of structural impulse response estimation. Very recently, Cavaliere *et al.* (2023) suggested a methodology for inference on ridge-type estimators that relies on bootstrapping. Finally, shrinkage of autoregressive models to constrained submodels was discussed by Hansen (2016b) in a more general setting.

Finally, various estimation problems can either be cast as or augmented with ridge-type regressions. Goulet Coulombe (2023) shows that the estimation of VARs with time-varying parameters can be written as ridge regression. Plagborg-Møller (2016) and Barnichon and Brownlees (2019) both use ridge to derive smoothed local projection impulse response functions.

1.2. Outline

Section 2 provides a discussion of the ridge penalty and the ridge VAR estimator. In Section 3 I deal with the properties of ridge-induced shrinkage in the autoregressive coefficients. I discuss the connections between frequentist and Bayesian ridge for VAR models within Section 4. Section 5 develops the asymptotic theory and inference result in the case where there is no asymptotic shrinkage. This includes studying the property of cross-validation under dependence. Section 6 provides inference and CV results in a setting where some shrinkage of a subset of parameters is possible. Section 7 presents Monte Carlo simulations focused on impulse response estimation. Section 8 concludes. Finally, the Data S1 Supplementary Appendix contains more detailed derivations, as well as all proofs, additional tables and further information on simulations.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12737

1.3. Notation

Define $\mathbb{R}_+$ to be the set of strictly positive real numbers. Vectors $v \in \mathbb{R}^N$ and matrices $A \in \mathbb{R}^{N\times M}$ are always denoted with lower and uppercase letters respectively. Throughout, I will use I_M to represent the identity matrix of dimension M. For any vector $v \in \mathbb{R}^N$, $\|v\|$ is the Euclidean norm. For any matrix $A \in \mathbb{R}^{N\times M}$, $\|A\|$ is the spectral norm unless stated otherwise; $\|A\|_{\max} = \max_{i,j}\left|a_{ij}\right|$ is the maximal entry norm; $\|A\|_F = (\mathrm{tr}\{A'A\})^{-1/2}$ is the Frobenius norm; $\mathrm{vec}(\cdot)$ is the vectorization operator and $\otimes$ is the Kronecker product (Lütkepohl, 2005). If a vector represents a vectorized matrix, then it will be written in bold, that is, for $A \in \mathbb{R}^{N\times M}$ I write $\mathrm{vec}(A) = \boldsymbol{a} \in \mathbb{R}^{NM}$. Let $\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_{K^2p}\}$, $\lambda_i > 0$ for all $i = 1, \dots, K^2p$. To give the partial ordering of diagonal positive semi-definite penalization matrices, let $\Lambda_1 = \mathrm{diag}\{\lambda_{1,j}\}_{j=1}^{K^2p}$ and $\Lambda_2 = \mathrm{diag}\{\lambda_{2,j}\}_{j=1}^{K^2p}$. I 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 $\lambda_{1,j} < \lambda_{2,j}$. Symbols $\xrightarrow{P}$ and $\xrightarrow{d}$ are used to indicate convergence in probability and convergence in distribution respectively.

2. RIDGE REGULARIZED VAR ESTIMATION

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

$$y_t = \nu_t + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t, \tag{1}$$

where $u_t = (u_{1t}, \dots, u_{Kt})'$ is additive noise such that u_t are i.i.d. with $\mathbb{E}[u_{it}] = 0$ and $\mathrm{Var}[u_t] = \Sigma_u$, and ν_t is a deterministic trend. For simplicity, in the remainder I shall assume that $\nu_t = 0$ so that y_t has no trend component – equivalently, y_t is a detrended series.

For a given sample size T define $Y = (y_1, \dots, y_T) \in \mathbb{R}^{K\times T}$, $z_t = (y_t', y_{t-1}', \dots, y_{t-p+1}')' \in \mathbb{R}^{Kp}$, $Z = (z_0, \dots, z_{T-1}) \in \mathbb{R}^{Kp\times T}$, $B = (A_1, \dots, A_p) \in \mathbb{R}^{K\times Kp}$, $U = (u_1, \dots, u_T) \in \mathbb{R}^{K\times T}$, and vectorized counterparts $\boldsymbol{y} = \mathrm{vec}(Y)$, $\boldsymbol{\beta} = \mathrm{vec}(B)$ and $\boldsymbol{u} = \mathrm{vec}(U)$. Accordingly, $Y = BZ + U$ and $\boldsymbol{y} = (Z' \otimes I_K)\boldsymbol{\beta} + \boldsymbol{u}$, where $\Sigma_{\boldsymbol{u}} = I_K \otimes \Sigma_u$. Importantly, throughout this article, I will assume that the cross-sectional dimension K remains fixed.

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

$$\hat{\boldsymbol{\beta}}^R(\lambda) := \arg\min_{\boldsymbol{\beta}} \frac{1}{T}\left\|\boldsymbol{y} - (Z' \otimes I_K)\boldsymbol{\beta}\right\|^2 + \lambda\|\boldsymbol{\beta}\|^2,$$

where $\lambda > 0$ is the scalar regularization parameter or regularizer. When $\lambda\|\boldsymbol{\beta}\|^2$ is replaced with quadratic form $\boldsymbol{\beta}'\Lambda\boldsymbol{\beta}$ for a positive definite matrix Λ the above is often termed Tikhonov regularization. To avoid confusion, I shall also refer to it as 'ridge', since in what follows Λ will always be assumed diagonal. As Λ does not, in general, penalize all coefficients equally, it can be used to construct an *anisotropic* ridge estimator. By solving the normal equations (see Supplementary Appendix A.1), the RLS estimator with positive semi-definite regularization matrix $\Lambda \in \mathbb{R}^{K^2p\times K^2p}$ is shown to be

$$\hat{\boldsymbol{\beta}}^R(\Lambda) = \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1} \frac{(Z \otimes I_K)\boldsymbol{y}}{T}.$$

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

$$\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) = \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1} \left(\frac{(Z \otimes I_K)\boldsymbol{y}}{T} + \Lambda\boldsymbol{\beta}_0\right). \tag{2}$$

In the context of multi-variate estimation, one has to make a further distinction between two related types of ridge estimators. I let $\hat{B}^R(\Lambda, \boldsymbol{\beta}_0)$ be the de-vectorized coefficient estimator obtained from reshaping $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ to a $K \times Kp$ matrix. But one can also consider the *matrix RLS estimator* $\hat{B}^R_{\text{mat}}(\Lambda_{Kp}, B_0)$ given by

$$\hat{B}^R_{\text{mat}}(\Lambda_{Kp}, B_0) = T^{-1}(Y + B_0\,\Lambda_{Kp})Z'\left(T^{-1}ZZ' + \Lambda_{Kp}\right)^{-1},$$

where $\Lambda_{Kp} = \text{diag}\{\lambda_1, \dots, \lambda_{Kp}\}$ and B_0 is a centering matrix. The distinction is important because the vectorized and matrix RLS estimators in general need not coincide. As discussed in Supplementary Appendix A.2, $\hat{B}^R(\Lambda, \boldsymbol{\beta}_0)$ allows for more general penalty structures compared to $\hat{B}^R_{\text{mat}}(\Lambda_{Kp}, B_0)$. I, therefore, focus on the former rather than the latter.

Remark 1. Equation (2) implies that $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ and, therefore, $\hat{\boldsymbol{\beta}}^R(\Lambda)$ provide *simultaneous* estimation of all the coefficients in $\boldsymbol{\beta}$. However, by analogy with ordinary least squares (OLS) VAR estimation, one may also consider an *equation-by-equation* ridge regression (ebe-RLS) scheme. For $k = 1, \dots, K$, let $\boldsymbol{y}_k = (Z' \otimes I_K)\boldsymbol{\beta}_k + \boldsymbol{u}_k$ be the autoregressive equation for the kth series of y_t. Then, we can define the kth equation RLS estimator to be

$$\hat{\boldsymbol{\beta}}^R_k(\Lambda, \boldsymbol{\beta}_0) = \left(\frac{ZZ'}{T} \otimes I_K + \Lambda_k\right)^{-1}\left(\frac{(Z \otimes I_K)\boldsymbol{y}_k}{T} + \Lambda\boldsymbol{\beta}_{0k}\right),$$

where $0 \preceq \Lambda_k$ and $\boldsymbol{\beta}_{0k}$ are the kth equation regularizer and centering respectively. Notice that the ebe-RLS approach allows, by construction, to penalize the estimates for one component differently than for another, and the two can be independently chosen. This provides a higher degree of freedom than the one afforded by, for example, the anisotropic lag-adapted scheme proposed in Section 3.2 or the Bayesian schemes of Section 4. However, implementing ebe-RLS in applications inherently implies that data-driven tuning of Λ_k will be significantly more computationally intensive – with costs growing linearly in K. Due to this higher complexity, in both theoretical derivation and simulations below, I will focus on studying the properties of the simultaneous RLS estimator.

Remark 2. Further regarding ebe-RLS, another way to approach estimation is through the *recursive form* of the VAR model. Let $\Sigma_u = P^{-1}DP^{-1'}$, where P^{-1} is an unitriangular matrix and D a diagonal matrix, so that we may write

$$Y = GZ - \widetilde{P}Y + D^{-1/2}E,$$

where $G = PB$, $\widetilde{P} = P - I_K$ and noise term E has identity covariance matrix. Estimation can now be performed in ebe-RLS fashion, and matrices P, B and D are recovered (Hausman, 1983). Notice, however, that in this framework the *ordering* of variables plays a role, since it also determines the decomposition of Σ_u. Indeed, even if a penalization scheme is fixed, permuting the entries may yield different penalized estimates for P, B and D, so that both slope and covariance parameter estimates are different, implying (structural) IRF estimates will also differ. However, note that this issue is somewhat mirrored in a recursive shock identification approach: after estimation, $\hat{\Sigma}_u$ is Cholesky decomposed to identify the shocks' rotation, and the ordering of variables is key and must be economically justified.

3. SHRINKAGE

Here, I discuss both the isotropic ridge penalty, i.e., the 'standard' ridge approach, and an anisotropic penalty that is better adapted to the VAR setting. An important result is that, even in simple setups with only two variables, the shrinkage induced by ridge can either increase or reduce bias, as well as the stability of autoregressive estimates.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12737

Throughout this section, I consider *fixed* design matrices and the focus will be on the geometric properties of ridge.

3.1. Isotropic Penalty

The most common way to perform a ridge regression is through isotropic regularization, that is, $\Lambda = \lambda I$ for some scalar $\lambda \geq 0$. Isotropic ridge has been extensively studied, see for example, the comprehensive review of Hastie (2020). With regard to shrinkage, an isotropic ridge penalty can be readily studied.

Proposition 1. Let $Z \in \mathbb{R}^{M\times T}$, $Y \in \mathbb{R}^{T}$ for $T > M$ be regression matrices. For $\lambda_\bullet > \lambda > 0$ and isotropic RLS estimator $\hat{\beta}^R(\lambda) := (T^{-1}ZZ' + \lambda I_M)^{-1}(T^{-1}ZY)$ it holds

$$\left\|\hat{\beta}^R(\lambda_\bullet)\right\| < \left\|\hat{\beta}^R(\lambda)\right\|.$$

Proof. Using the full singular-value decomposition (SVD), decompose $T^{-1/2}Z = UDV' \in \mathbb{R}^{M\times T}$ where U is $M \times M$ orthogonal, D is $M \times T$ diagonal and V is $T \times T$ orthogonal. Write

$$\begin{aligned}
\hat{\beta}^R(\lambda_\bullet) &= (T^{-1}ZZ' + \lambda_\bullet I_M)^{-1}(T^{-1}ZY) \\
&= (UDV'VDU' + \lambda_\bullet I_M)^{-1}UDV'(T^{-1/2}Y) \\
&= U(D^2 + \lambda_\bullet I_M)^{-1}DV'(T^{-1/2}Y) \\
&= U(D^2 + \lambda_\bullet I_M)^{-1}(D^2 + \lambda I_M)(D^2 + \lambda I_M)^{-1}DV'(T^{-1/2}Y) \\
&= \left[U(D^2 + \lambda_\bullet I_M)^{-1}(D^2 + \lambda I_M)U'\right]\hat{\beta}^R(\lambda).
\end{aligned}$$

Since $D^2 = \operatorname{diag}\{\sigma_j^2\}_{j=1}^M$, the term within brackets is $U \operatorname{diag}\{(\sigma_j^2 + \lambda)/(\sigma_j^2 + \lambda_\bullet)\}_{j=1}^M U'$. Moreover, because the spectral norm is unitary invariant and $\lambda_1 > \lambda_2$, it follows that

$$\left\|U(D^2 + \lambda_\bullet I_M)^{-1}(D^2 + \lambda I_M)U'\right\| = \left\|\operatorname{diag}\{(\sigma_j^2 + \lambda)/(\sigma_j^2 + \lambda_\bullet)\}_{j=1}^M\right\| < 1.$$

Finally, by the sub-multiplicative property it holds

$$\left\|\hat{\beta}^R(\lambda_\bullet)\right\| \leq \left\|U(D^2 + \lambda_1 I_M)^{-1}(D^2 + \lambda I_M)U'\right\| \cdot \left\|\hat{\beta}^R(\lambda)\right\| < \left\|\hat{\beta}^R(\lambda)\right\|,$$

as claimed. ∎

Proposition 1 and its proof expose the main ingredients of ridge regression. From the SVD of $T^{-1/2}Z$ used above, it is clear that ridge regularization acts uniformly along the orthogonal directions that are the columns of V. The improvement in the conditioning of the inverse comes from all diagonal factors $[(D^2 + \lambda_\bullet I_M)^{-1}D]_j = \sigma_j/(\sigma_j^2 + \lambda_\bullet)$ being well-defined, even when $\sigma_j = 0$ (as is the case in systems with collinear regressors).

However, directly applying isotropic ridge to vector autoregressive models is not necessarily the most effective estimation approach. Stable VAR models show decay in the absolute size of coefficients over lags. Thus, it is reasonable to choose a more general ridge penalty that can accommodate lag decay.

3.2. Lag-Adapted Penalty

I now consider a different form for Λ that is of interest when applying ridge specifically to a VAR model. Define family $\mathcal{F}^{(p)}$ of *lag-adapted* ridge penalty matrices as

$$\mathcal{F}^{(p)} = \{\operatorname{diag}\{\lambda_1, \ldots, \lambda_p\} \otimes I_{K^2} | \lambda_i \in \mathbb{R}_+,\ i = 1, \ldots, p\},$$

where each λ_i intuitively implies a different penalty for the elements of each coefficient matrix A_i, $i = 1, \dots, p$.[1] The family $\mathcal{F}^{(p)}$ allows imposing a ridge penalty that is coherent with the lag dimension of an autoregressive model. It is parametrized by p distinct penalty factors, meaning that the penalization is *anisotropic*.

Proposition 2. Let $Z \in \mathbb{R}^{Kp \times T}$, $\boldsymbol{y} \in \mathbb{R}^{KT}$ for $T > Kp$ be multi-variate VAR regression matrices. Given subset $\mathcal{S} \subseteq \{1, \dots, p\}$ of cardinality $s = |\mathcal{S}|$, for $\Lambda^{(p)} \in \mathcal{F}^{(p)}$ define $\hat{\boldsymbol{\beta}}^R(\Lambda^{(p)})_{[\mathcal{S}]}$ as the vector of sK^2 coefficient estimates located at indexes $1 + K^2(j-1), \dots, K^2 j$ for $j \in \mathcal{S}$. Let $\mathcal{S}^c = \{1, \dots, p\} \setminus \mathcal{S}$ be the complement of $\mathcal{S}$.

a. If $\lambda_1 \geq \lambda_2$, then $\left\| \hat{\boldsymbol{\beta}}^R(\lambda_1 I_{K^2 p})_{[\mathcal{U}]} \right\| \leq \left\| \hat{\boldsymbol{\beta}}^R(\lambda_2 I_{K^2 p})_{[\mathcal{U}]} \right\|$ for any $\mathcal{U} \subset \{1, \dots, K^2 p\}$. The inequality is strict when $\lambda_1 > \lambda_2$.
b. 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 zeros as coefficients for the lags indexed by $\mathcal{S}^c$. Similarly, let $\Lambda^{(p)}_{[\mathcal{S}]}$ be the subset of diagonal elements in $\Lambda^{(p)}$ penalizing the lags in $\mathcal{S}$. Then

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

where $\Lambda^{(p)}_{[\mathcal{S}]} \to 0$ and $\Lambda^{(p)}_{[\mathcal{S}^c]} \to \infty$ are to be intended as the element-wise convergence.

Proposition 2 shows that the limiting geometry of a lag-adapted ridge estimator is thus identical to that of a least squares regression run on the subset specified by $\mathcal{S}$. By controlling the size of coefficients $\{\lambda_1, \dots, \lambda_p\}$ it is therefore possible to obtain pseudo-model-selection. However, in the next section, I show that anisotropic penalization produces complex effects on the model's coefficient estimates.

3.3. Illustration of Anisotropic Penalization

Here, I aim to illustrate the effects of a lag-adapted ridge penalty on VAR coefficients estimates using a particular example. This further helps motivate and contextualize the results of the simulation exercises provided in Section 7. More generally, before moving on to the discussion of more sophisticated forms of ridge regression, it is important to gain some intuition regrading the properties of anisotropic penalization, which I highlight with the help of a simple bivariate VAR(2) model.

Note that, since ridge operates along principal components, there is no immediate relationship between a specific subset of the estimated coefficients and a given diagonal block of $\Lambda^{(p)}$. With regard to autoregressive modeling, three effects are of interest: the shrinkage of coefficient matrices A_i relative to the choice of λ_i; the entity of the bias introduced by shrinkage, and the impact of shrinkage on the persistence of the estimated model.

To showcase these effects, I consider the VAR(2) model

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + u_t, \quad u_t \sim \text{i.i.d. } \mathcal{N}(0, \Sigma_u),$$

where

$$A_1 = \begin{bmatrix} 0.8 & 0.1 \\ -0.1 & 0.7 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0.1 & -0.2 \\ -0.1 & 0.1 \end{bmatrix}, \quad \Sigma_u = \begin{bmatrix} 0.3 & 0 \\ 0 & 5 \end{bmatrix}.$$

[1] Note that with a lag-adapted penalty it is also possible to directly use the matrix ridge estimator since the penalty for $\hat{\boldsymbol{\beta}}^R$ is given by $\text{diag}\{\lambda_1, \dots, \lambda_p\} \otimes I_{K^2} = (\text{diag}\{\lambda_1, \dots, \lambda_p\} \otimes I_K) \otimes I_K$, see Supplementary Appendix A.2. Importantly, this kind of structure is minimal in terms of modeling the relative size of coefficients *within* each coefficient matrix A_i. If economic theory or intuition provides information about the effects of one specific variable and lag on another – say, the contemporaneous effect of the first series on the second series is zero – more structure can be integrated into the ridge penalty matrix. This would mean, however, that different ridge estimator forms are not equivalent.

DOI: 10.1111/jtsa.12737

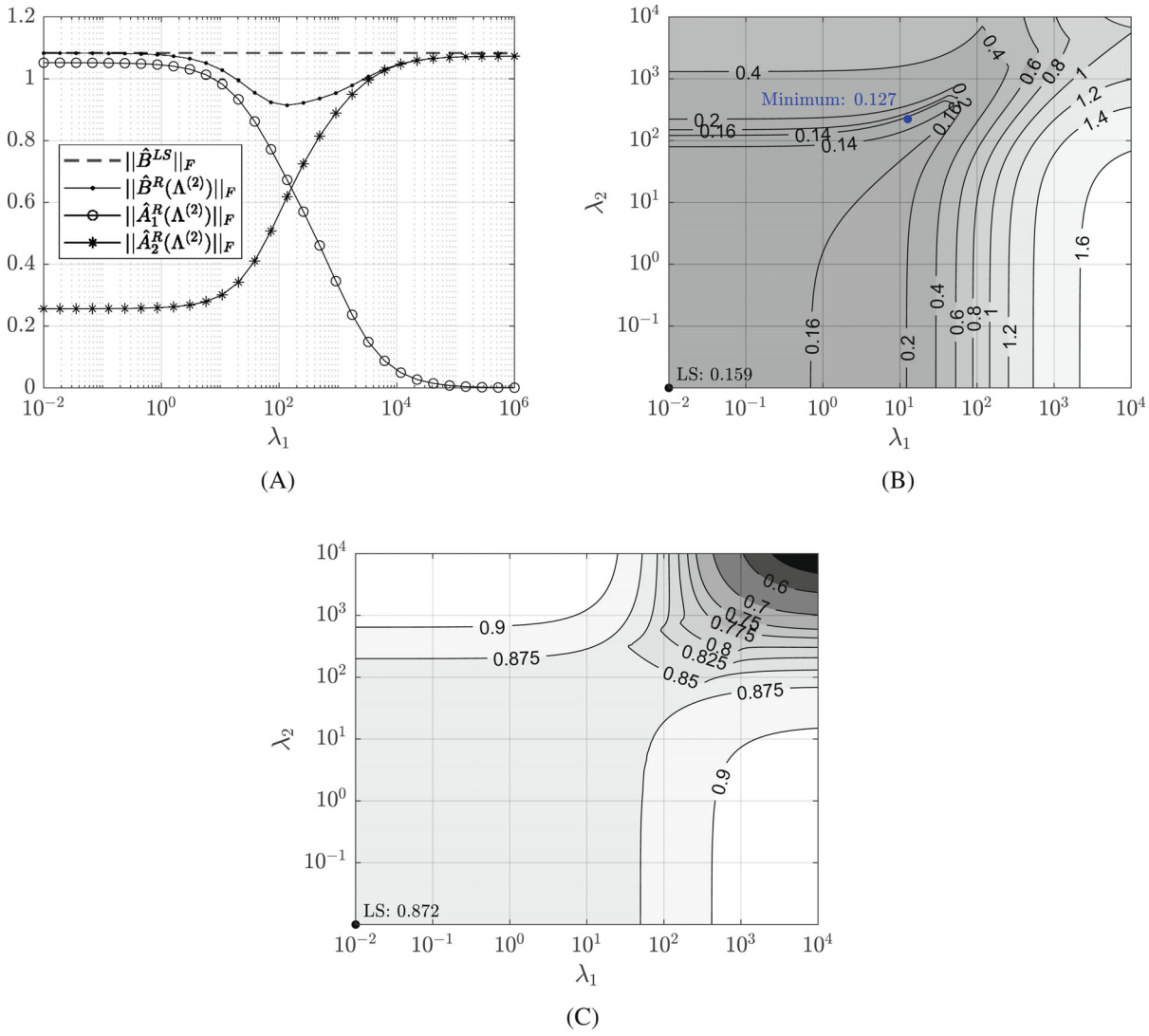

Figure 1. Shrinkage of coefficients estimate in Frobenius norm (a); bias induced by shrinkage (b); change in stability of estimated VAR model at different levels of penalization, measured by the absolute value of the largest companion form eigenvalue (c)

A single sample of length $T = 200$ is drawn, demeaned and used to estimate coefficients A_1 and A_2. The VAR(2) model is fitted using the lag-adapted ridge estimator $\hat{B}^R(\Lambda^{(2)})$, where $\Lambda^{(2)} = \text{diag}\{\lambda_1, \lambda_2\} \otimes I_2$. Note that $\hat{B}^R(\Lambda^{(2)})$ can be partitioned into estimates $\hat{A}_1^R(\Lambda^{(2)})$ and $\hat{A}_2^R(\Lambda^{(2)})$ for the respective parameter matrices.

3.3.1. Shrinkage

To illustrate shrinkage, I consider the restricted case of $\lambda_1 \in [10^{-2}, 10^6]$ and $\lambda_2 = 0$. The ridge estimator is computed for varying λ_1 over a logarithmically spaced grid. Figure 1(a) shows that $\left\|\hat{B}^R(\Lambda^{(2)})\right\|_F \approx \left\|\hat{B}^{LS}\right\|_F$ for $\lambda_1 \approx 0$, but as the penalty increases $\left\|\hat{A}_1^R(\Lambda^{(2)})\right\|_F$ decreases while $\left\|\hat{A}_2^R(\Lambda^{(2)})\right\|_F$ grows. The resulting behavior of $\left\|\hat{B}^R(\Lambda^{(2)})\right\|_F$ is non-monotonic in λ_1, although indeed $\left\|\hat{B}^R(\Lambda^{(2)})\right\|_F < \left\|\hat{B}^{LS}\right\|_F$ in the limit $\lambda_1 \to \infty$. This effect is

due to the model selection properties of lag-adapted ridge, and the resulting omitted variable bias. Therefore, in practice, it is not generally true that anisotropic ridge induces monotonic shrinkage of estimates.

3.4. Bias

Since ridge bias is hard to study theoretically, I use a simulation with the same setup of Figure 1(a), this time with $\lambda_1, \lambda_2 \in [10^{-2}, 10^4]$. The grid is logarithmic with 150 points. Figure 1(b) presents a level plot of the sup-norm ridge bias $\left\| \hat{B}^R(\Lambda^{(2)}) - B \right\|_\infty$ given multiple combinations of λ_1 and λ_2. While there can be gains compared to the least squares estimator $\hat{B}^{\text{LS}}$, they are modest. Moreover, level curves of the bias surface show that gains concentrate in a very thin region of the parameter space. Consequently, one may imagine that, in practice, any (data-driven) ridge penalty selection criterion is unlikely to yield bias improvement over least squares. Yet, in large VAR models with many lags, the reduction in variance of the ridge estimator often yields improvements over un-regularized procedures (Li *et al.*, 2024). However, the bias-variance trade-off in ridge is not a free-lunch when performing inference. Pratt (1961) showed that it is not possible to produce a test (equivalently, a CI procedure) which is valid uniformly over the parameter space and yields meaningfully smaller confidence intervals than any other valid method.

3.5. Stability

To study the stability of ridge VAR estimates, I reuse the results of the bias simulation above. Let $\mathbb{A}$ be the companion matrix of $[A_1, A_2]$, and $\hat{\mathbb{A}}^R$ the companion matrix of estimates $[\hat{A}_1^R(\Lambda^{(2)}), \hat{A}_2^R(\Lambda^{(2)})]$. For all combinations (λ_1, λ_2), I compute the largest eigenvalue $\omega_1(\hat{\mathbb{A}}^R)$ of $\hat{\mathbb{A}}^R$. Note that if $|\omega_1(\hat{\mathbb{A}})| < 1$, then the estimated VAR(2) is stable (Lütkepohl, 2005). Figure 1(c) presents the level sets for the surface of maximal eigenvalue moduli, and for comparison $|\omega_1(\hat{B}^{\text{LS}})|$ is shown at the origin.[2] While along the main diagonal there is a clear decrease in $|\omega_1(\hat{\mathbb{A}}^R)|$ as isotropic penalization increases, when λ_1 is large and $\lambda_2 \ll 1$ (or vice versa) the maximal eigenvalue increases instead. Therefore, an estimate of a VAR model obtained with anisotropic ridge may be *closer* to unit root than the least squares estimate.

4. BAYESIAN AND FREQUENTIST RIDGE

So far, I have discussed standard ridge penalization schemes. Here, I study the posterior mean of Bayesian VAR (BVAR) priors commonly applied in the macroeconometrics literature. I show that such posteriors are in fact specific GLS formulations of the ridge estimator. This comparison highlights that ridge can be seen as a way to embed prior knowledge into the least squares estimation procedure by means of centering and rescaling coefficient estimates.

4.1. Litterman–Minnesota Priors

In Bayesian time series modeling, the so-called Minnesota or Litterman prior has found great success (Litterman, 1986). For stationary processes which one believes to have reasonably small dependence, a zero-mean normal prior can be put on the VAR parameters, with non-zero prior variance. Assuming that the covariance matrix of errors Σ_u is known, the Litterman–Minnesota has posterior mean

$$\overline{\boldsymbol{\beta}}|\Sigma_u = \left[\underline{V}_{\boldsymbol{\beta}}^{-1} + (ZZ' \otimes \Sigma_u^{-1})\right]^{-1} (Z \otimes \Sigma_u^{-1})\boldsymbol{y}, \tag{3}$$

[2] If $\Lambda^{(2)} \to 0$, then by continuity of eigenvalues it follows that $|\omega_1(\hat{\mathbb{A}}^R)| \to |\omega_1(\hat{\mathbb{A}}^{\text{LS}})|$, see Supplementary Appendix A.3.

where $\underline{V}_\beta \succ 0$ is the prior covariance matrix of $\boldsymbol{\beta}$ (Lütkepohl, 2005). It is common to let $\underline{V}_\beta$ be diagonal, and often the entries follow a simple pattern which depends on lag, individual components variances, and prior hyperparameters. For example, Bańbura *et al.* (2010) suggest the following structure for the diagonal

$$v_{i,jk} = \begin{cases} \dfrac{\lambda^2}{i^2} & \text{if } j = k, \\ \theta \dfrac{\lambda^2}{i^2} \dfrac{\sigma_j^2}{\sigma_k^2} & \text{if } j \neq k, \end{cases} \tag{4}$$

where $v_{i,jk}$ is the prior variance for coefficients $(A_i)_{jk}$ for $i = 1, \dots, p$ and $j, k = 1, \dots, K$. Here, σ_j is the jth diagonal element of Σ_u, $\theta \in (0, 1)$ specifies beliefs on the explanatory importance of own lags relative to other variables' lags, while $\lambda \in [0, \infty]$ controls the overall tightness of the prior. The extreme $\lambda = 0$ yields a degenerate prior centered at $\overline{\boldsymbol{\beta}} = 0$, while $\lambda = \infty$ reduces the posterior mean to the OLS estimate $\hat{\boldsymbol{\beta}}^{\text{LS}}$. Factor $1/i^2$, which explicitly shrinks variance at higher lags, was originally introduced by De Mol *et al.* (2008), who formally developed the idea that coefficients at deeper lags should be coupled with more penalizing priors. Note that, in (4), assuming $\sigma_j^2 = \sigma_k^2$ for all $j, k = 1, \dots, K$ and setting $\theta = 1$, produces a $\underline{V}_\beta$ that has a lag-adapted structure with quadratic lag decay.[3]

Equation (3) more generally demonstrates that the Minnesota posterior mean is equivalent to a ridge procedure. It is important to notice that, while with least squares the OLS and GLS estimators of VAR coefficients coincide, this is not the case with ridge regression. Regularizing a GLS regression will yield

$$\hat{\boldsymbol{\beta}}^{\text{RGLS}}(\Lambda) := \left[\Lambda + (ZZ' \otimes \Sigma_u^{-1})\right]^{-1} (Z \otimes \Sigma_u^{-1})\boldsymbol{y}, \tag{5}$$

instead of $\hat{\boldsymbol{\beta}}^{R}$, which is equivalent to (3) under an appropriate choice of Λ. While I develop the asymptotic results for $\hat{\boldsymbol{\beta}}^{R}$ assuming a centering parameter $\underline{\boldsymbol{\beta}}_0 \neq 0$ in general, I do not directly study the properties $\hat{\boldsymbol{\beta}}^{\text{RGLS}}$. The generalization to GLS ridge employing the least squares error covariance estimator $\hat{\Sigma}_T^{\text{LS}}$ should follow from straightforward arguments. In Section 7, I focus on providing evidence on the application $\hat{\boldsymbol{\beta}}^{\text{RGLS}}$ in terms of its pointwise impulse response estimation mean-squared error.

Remark 3. In principle, ridge penalties can be designed to implement shrinkage toward nonstationary or long memory priors, too. Very recently, for example, Bauwens *et al.* (2023) have suggested a ridge-type strategy to estimate a one-lag long memory model: their penalization scheme follows naturally from the assumption that an observed AR(1) series originates from an infinite-dimensional VAR(1) with an appropriate off-diagonal structure. One may also think of applying a unit-root-centered matrix RLS estimator $\hat{B}^{R}_{\text{mat}}(\Lambda_{Kp}, B_0^{\dagger})$, where $B_0^{\dagger} := (I_K, 0_K, \dots, 0_K) \in \mathbb{R}^{K \times Kp}$. This is, in fact, exactly the centering of the Litterman–Minnesota prior (Bańbura *et al.*, 2010). Notice, however, that this type of prior imposes very strict assumptions on the *form* of the unit-root – namely, each component is unaffected by any of the others.[4] Finally, shrinkage to subspaces associated with a factor model specification has also been explored (Huber and Koop, 2023).

[3] Bańbura *et al.* (2010) also assume $\theta = 1$ in their BVAR estimation. They wish to relax the Litterman–Minnesota assumption that Σ_u is a fixed, diagonal matrix and implement estimation directly by augmenting their data with appropriately constructed dummy variables (Kadiyala and Karlsson, 1997). This approach, however, is selected primarily for computation reasons due to the size of their Bayesian model.

[4] While stationarity of autoregressive estimates can be easily enforced using the Yule–Walker estimator (Brockwell and Davis, 1991), exact unit-root behavior is inherently hard to encode via penalization due to the complex geometry of the stationary region, see the discussion by Heaps (2023).

DOI: 10.1111/jtsa.12737 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

4.2. Hierarchical Priors

Recent research on Bayesian vector autoregressions exploit more sophisticated priors compared to the Litterman–Minnesota design. Giannone *et al.* (2015) develop an advanced BVAR model by setting up hierarchical priors which entail not only model parameters, but also hyperparameters. They impose

$$\Sigma_u \sim \text{IW}(\underline{\Psi}, \underline{d}),$$
$$\boldsymbol{\beta}|\Sigma_u \sim \mathcal{N}\left(\underline{\boldsymbol{\beta}}, \lambda(\Sigma_u \otimes \underline{\Omega})\right),$$

for hyperparameters $\underline{\boldsymbol{\beta}}$, $\underline{\Omega}$, $\underline{\Psi}$, and $\underline{d}$, where IW is the Inverse-Wishart distribution. Here, too, scalar $\lambda \in [0, \infty]$ controls prior tightness. Let $\underline{B}$ be the matrix form of the VAR coefficient prior mean, so that $\text{vec}(\underline{B}) = \underline{\boldsymbol{\beta}}$. The resulting (conditional) posterior mean $\overline{B}$ is given by

$$\overline{B}|\Sigma_u = \left[(\lambda\underline{\Omega})^{-1} + ZZ'\right]^{-1}\left[ZY + (\lambda\underline{\Omega})^{-1}\underline{B}\right]. \tag{6}$$

Observe that equation (6) is effectively equivalent to a centered ridge estimator, cf. (2).

The introduction of a hierarchical prior leaves space to add informative hyperpriors on the model hyperparameters, allowing for a more flexible fit. Indeed, removing the zero centering constraint from the prior on $\boldsymbol{\beta}$ can improve estimation. It is often the case that economic time series show a high degree of correlation and temporal dependence, therefore imposing $\boldsymbol{\beta} = 0$ as in the Minnesota prior is inadequate. In fact, Giannone *et al.* (2015) show that their approach yields substantial improvements in forecasting exercises, even when hyperparameter priors are relatively flat and uninformative.

5. STANDARD INFERENCE

Here, I state the main asymptotic results 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, random variables dependent on sample size T. In particular, $\boldsymbol{\beta}_0$ may be a consistent estimator of $\boldsymbol{\beta}$.

I will impose the following assumptions.

Assumptions

A. $\{u_t\}_{t=1}^T$ is a sequence of i.i.d. random variables with $\mathbb{E}[u_{it}] = 0$, covariance $\mathbb{E}[u_t u_t'] = \Sigma_u$ non-singular positive definite and $\mathbb{E}\left|u_{it}u_{jt}u_{mt}u_{nt}\right| < \infty$, $i, j, m, n = 1, \ldots, K$.
B. There exists $\rho > 1$ such that $\det(I_K - \sum_{i=1}^p A_i z^i) \neq 0$ for all complex z, $|z| \leq \rho$.
C. There exist $0 < \underline{m} \leq \overline{m} < \infty$ such that $\underline{m} \leq \omega_K(\Gamma) \leq \omega_1(\Gamma) \leq \overline{m}$, where $\Gamma = \mathbb{E}[z_t z_t']$ is the autocovariance matrix of z_t and $\omega_1(\Gamma)$, $\omega_K(\Gamma)$ are its largest and smallest eigenvalues respectively.

Assumption A is standard and allows proving the main asymptotic results with well-known 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 but is otherwise a simple generalization, although more involved in terms of notation, see e.g., Boubacar Mainassara and Francq (2011). Assumption B guarantees that y_t has no unit roots and is stable. Of course, many setups of interest do not satisfy this assumption, the most significant ones being unit roots, cointegrated VARs, and local-to-unity settings. Incorrect identification of unit roots does not invalidate the use of LS or ML estimators (Park and Phillips, 1988, 1989; Phillips, 1988; Sims *et al.*, 1990), however inference is significantly impacted as a result (Pesavento and Rossi, 2006; Mikusheva, 2007, 2012). Assumption C is standard in the literature regarding penalized estimation and does not imply 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.

Before stating the main theorems, let

$$\hat{\Gamma} = T^{-1}ZZ',$$
$$\hat{U} = Y - \hat{B}^R Z,$$
$$\hat{\Sigma}_u^R = T^{-1}\hat{U}\hat{U}',$$

be the regression covariance matrix, regression residuals and sample innovation covariance estimator respectively.

Theorem 1. Let Assumptions A–C hold and define $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ be the centered RLS estimator as in (2). If $\sqrt{T}\Lambda \xrightarrow{P} \Lambda_0$ and $\boldsymbol{\beta}_0 \xrightarrow{P} \underline{\boldsymbol{\beta}}_0$, where Λ_0 is a positive semi-definite diagonal matrix and $\underline{\boldsymbol{\beta}}_0$ is a constant vector, then

a. $\hat{\Gamma} \xrightarrow{P} \Gamma$,
b. $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) \xrightarrow{P} \boldsymbol{\beta}$,
c. $\hat{\Sigma}_u^R \xrightarrow{P} \Sigma_u$,
d. $\sqrt{T}\left(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\Gamma^{-1}\Lambda_0(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}), \Gamma^{-1} \otimes \Sigma_u\right)$.

Theorem 1 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 distribution is clearly dependent on the unknown model parameters $\boldsymbol{\beta}$, complicating inference.

However, it is possible – under strengthened assumptions for Λ or $\boldsymbol{\beta}_0$ – for $\hat{\boldsymbol{\beta}}^R$ to have a zero-mean Gaussian limit distribution.

Theorem 2. In the setting of Theorem 1, results (a)–(c) hold and (d) simplifies to

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

if either

(i) $\Lambda = o_P\left(T^{-1/2}\right)$,
(ii) $\Lambda = O_P\left(T^{-1/2}\right)$ and $\boldsymbol{\beta}_0 - \boldsymbol{\beta} = o_p(1)$.

The Corollary 1 is immediate.

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

5.1. Joint Inference

To handle smooth transformations of VAR coefficients, such as impulse responses (Lütkepohl, 1990), I also derive a standard joint limit result for both $\hat{\boldsymbol{\beta}}^R$ and the variance estimator $\hat{\Sigma}_u^R$.

Theorem 3. Let $\hat{\boldsymbol{\sigma}}^R = \mathrm{vec}(\hat{\Sigma}_u^R)$ and $\boldsymbol{\sigma} = \mathrm{vec}(\Sigma_u)$. Under the assumptions of Theorem 1,

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

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

This result is key as it allows, under the stated assumptions on the penalizer, to construct valid asymptotic confidence intervals and, specifically, perform impulse response inference, as done in the simulations of Section 7 using the Delta Method (Lütkepohl, 2005).

5.2. Cross-validation

In practice, the choice of ridge penalty is often data-driven, and cross-validation is a very popular approach to select Λ. I now turn to the properties of CV as applied to the RLS estimator $\hat{\boldsymbol{\beta}}^R(\Lambda)$.

For simplicity, assume that y_t is an AR(p) process, that is, $K = 1$. In this setting,

$$\hat{\boldsymbol{\beta}}^R(\Lambda) = \left(\frac{ZZ'}{T} + \Lambda\right)^{-1}\frac{Z\boldsymbol{y}}{T},$$

where $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_p\}$. Following Patil *et al.* (2021), the *prediction error* of ridge estimator $\hat{\boldsymbol{\beta}}^R(\Lambda)$ given penalty Λ is

$$\text{Err}\left(\hat{\boldsymbol{\beta}}^R(\Lambda)\right) := \mathbb{E}_{\tilde{y},\tilde{z}}\left[\left(\tilde{y} - \tilde{z}'\hat{\boldsymbol{\beta}}^R(\Lambda)\right)^2 \middle| Z, \boldsymbol{y}\right],$$

where $\tilde{y}$ and $\tilde{z}$ are random variables from an independent copy of y_t. In particular, $\tilde{z}$ is the vector of p lags of $\tilde{y}$. Moreover, the error curve for Λ is given by

$$\text{err}(\Lambda) := \text{Err}\left(\hat{\boldsymbol{\beta}}^R(\Lambda)\right).$$

The prediction error is crucial because it allows to determine the oracle optimal penalization,

$$\Lambda^* := \arg\min_{\Lambda \succcurlyeq 0} \text{err}(\Lambda).$$

Clearly, err(Λ) is unavailable in practice and Λ^* must be substituted with a feasible alternative. Cross-validation proposes to construct a collection of paired, non-overlapping subsets of the sample data such that the first subset of the pair (estimation set) is used to estimate the model, while the second (validation set) is used to provide an empirical estimate of the prediction error. The CV penalty is then selected to minimize the total error over validation sets. A very popular approach to build cross-validation subsets is k-fold CV, wherein the sample is split into k blocks, so-called *folds*, of sequential observations (possibly after shuffling the data). Each fold determines a validation set, and is paired with its complement, which gives the estimation set. For more details, see e.g. Hastie *et al.* (2009).

Again with the intent of keeping complexity low – as this article is not focused on cross-validation – I will make the additional simplifying assumption that CV is implemented with two folds and one pair. Specifically, the first fold is the estimation set, where Z and $\boldsymbol{y}$ are constructed and $\hat{\boldsymbol{\beta}}^R(\Lambda)$ is estimated. The second fold is the validation set and yields $\widetilde{Z}, \widetilde{\boldsymbol{y}}$, where $\widetilde{Z} \in \mathbb{R}^{p \times \tilde{T}}$ and $\widetilde{\boldsymbol{y}} \in \mathbb{R}^{\tilde{T}}$. To account for dependence, a buffer of m observations

DOI: 10.1111/jtsa.12737

between validation and estimation folds is introduced. The last observation of y_t in the estimation set is y_T, while the first observation in the validation set is $\widetilde{y}_1 := y_{T+m+1}$, that is, the total number of available observations is $T + m + \widetilde{T} + 2p + 1$. This is a stylized version of the CV setup of Burman *et al.* (1994) – also called *m-block* or *non-dependent cross-validation* in Bergmeir *et al.* (2018) – and is effectively equivalent to an out-of-sample (OOS) validation scheme. Thus, the two-fold m-buffered CV error curve is

$$\mathrm{cv2}_m(\Lambda) := \frac{1}{\widetilde{T}} \sum_{s=1}^{\widetilde{T}} \left(\widetilde{y}_s - \widetilde{z}_s' \hat{\boldsymbol{\beta}}^R(\Lambda) \right)^2. \tag{7}$$

Theorem 4. Under Assumptions A–C, for every Λ in the cone of diagonal positive definite penalty matrices with diagonal entries in $(\lambda_{\min}, \infty)$, $\lambda_{\min} \geq 0$, it holds that

$$\mathrm{cv2}_m(\Lambda) - \mathrm{err}(\Lambda) \xrightarrow{a.s.} 0,$$

as $T, \widetilde{T} \to \infty$. Furthermore, the convergence is uniform in Λ over compact subsets of penalty matrices.

In the current setup, the joint limit $T, \widetilde{T} \to \infty$ should be thought as $\widetilde{T}/T \to \gamma \in (0, 1)$, where aspect ratio γ determines the balance of the cross-validation split.

Remark 4. Under Assumption C, $\omega_K(\hat{\Gamma}) > 0$ for T large. Therefore, the bounds derived in the proof of Theorem 4 are finite even if $\Lambda = 0$. In fact, it is easily seen that the behavior of $\mathrm{err}(\Lambda)$ and $\mathrm{cv2}_m(\Lambda)$ is consistent at the endpoints $\Lambda = 0$ and $\Lambda \to \infty$, see Patil *et al.* (2021). Observe that

$$\mathrm{cv2}_m(\Lambda) \to \Sigma_u \quad \text{and} \quad \mathrm{err}(\Lambda) \to \Sigma_u$$

as $\Lambda \to 0$, while

$$\mathrm{cv2}_m(\Lambda) \to \Gamma \quad \text{and} \quad \mathrm{err}(\Lambda) \to \Gamma,$$

as $\Lambda \to \infty$, as needed.

Theorem 4 thus shows that $\mathrm{cv2}_m(\Lambda)$ gives an asymptotically valid way to evaluate the prediction error curve, and thus tune Λ, over any compact set of diagonal positive semi-definite penalization matrices. Moreover, in Theorem C.2.1, Supplementary Appendix C.2, I show that the impact of dependence due to the VAR data generating process is exponentially small for m sufficiently large. This property of $\mathrm{cv2}_m(\Lambda)$ is desirable because it lets one choose m small also in applications with moderate sample sizes, and it theoretically justifies the prescription of Burman *et al.* (1994).

5.3. Asymptotically Valid CV

So far, I have shown that a simple two-fold CV – or, equivalently, OOS validation – correctly estimates the predictive error of the ridge estimator, even under dependence. I turn now to the question of selecting an *asymptotically valid* penalty, that is, a Λ such that condition (1) of Theorem 2 is fulfilled. This enables inference, since one is in a setting where the bias is asymptotically negligible.

The idea is to scale the ridge penalty used at the estimation step of CV by a factor $\sqrt{T}$, so that the validated penalty converges to zero at an appropriate rate as both T and $\widetilde{T}$ grow. In other words, an over-smoothed ridge regression turns out to be key when studying cross-validation. To derive this result, first let

$$\hat{\boldsymbol{\beta}}^R_{\blacklozenge}(\Lambda) := \left(\frac{ZZ'}{T} + \sqrt{T}\Lambda \right)^{-1} \frac{Z\boldsymbol{y}}{T},$$

be the *over-smoothed ridge estimator*.

DOI: 10.1111/jtsa.12737 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

Theorem 5. Under Assumptions A–C, let $\mathcal{I}_\lambda$ be the compact set of diagonal positive semi-definite penalization matrices Λ such that $\|\Lambda\|_{\max} \leq \lambda < \infty$. It holds

$$\Lambda^*_{\blacklozenge} := \arg\min_{\Lambda \in \mathcal{I}_\lambda} \mathrm{Err}\left(\hat{\boldsymbol{\beta}}^R_{\blacklozenge}(\Lambda)\right) = o_p(T^{-1/2}).$$

Remark 5. The previous theorem is stated in terms of the oracle predictive error $\mathrm{Err}\left(\hat{\boldsymbol{\beta}}^R_{\blacklozenge}\left(\widetilde{\Lambda}\right)\right)$, which equals the 2-fold CV error curve up to a factor of order $O_P(\widetilde{T}^{-1/2})$. Therefore, assuming that the CV aspect ratio γ is strictly between zero and one, the result of Theorem 5 also directly generalizes to an empirically cross-validated penalty.

6. INFERENCE WITH SHRINKAGE

Fu and Knight (2000) have argued that results such as Theorems 1 and 2 portray penalized estimators in a somewhat unfair light, because they result in asymptotic distributions showing no bias-variance trade-off. Indeed, they show that ridge shrinkage yields estimates with asymptotic variance no different from that of least squares. Of course, in finite samples shrinkage has an effect on $\Gamma^{-1} \otimes \Sigma_u$ since $\hat{\Sigma}^R_T$ is used in place of $\hat{\Sigma}^{\mathrm{LS}}_T$ to estimate the error term variance matrix. To better understand the value of ridge penalization in practice, one should therefore consider the situation where shrinkage is *not* asymptotically negligible for at least a subset of coefficients. A motivating example would be that of a VAR(∞) model derived by inverting a stable VARMA(p, q) process: for i sufficiently large, coefficient matrices A_i decay exponentially to zero.[5] One should thus be able to exploit such structural information about the autoregressive coefficients to asymptotically improve on the bias-variance trade-off. Following this intuition and the discussion of lag-adapted penalty matrices in Section 3.2, I shall now consider the empirically relevant regression setup where one assumes that a subset of VAR coefficients are small (with respect to sample size), but not necessarily zero. Thus, to have inference reflect this type of shrinkage, an asymptotic framework with non-negligible penalization of higher-order lag coefficients is in fact more appropriate than that of Theorem 1.[6]

Formally, assume that for some $0 < n \leq p$ one can partition the VAR coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^{K^2(p-n)}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{K^2 n}$, and assume that $\boldsymbol{\beta}_2 = T^{-(1/2+\delta)}\, \boldsymbol{b}_2$ for $\delta > 0$ and $\boldsymbol{b}_2 \in \mathbb{R}^{K^2 n}$ is fixed. Such ordered partitioning of $\boldsymbol{\beta}$ is without loss of generality.[7] In this setup, it is clearly desirable to penalize $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ differently when constructing the ridge penalty. Let $\Lambda = \mathrm{diag}\{(L_1', L_2')'\} \otimes I_K$ where $L_1 \in \mathbb{R}_+^{K^2(p-n)}$ and $L_2 \in \mathbb{R}_+^{K^2 n}$. Assume that

$$L_1 = o_P(T^{-1/2}) \quad \text{and} \quad L_2 \xrightarrow{P} \overline{L}_2,$$

for a fixed vector $\overline{L}_2 \in \mathbb{R}_+^{K^2 n}$. In particular, letting $\Lambda_1 = \mathrm{diag}\{L_1\}$ and $\Lambda_2 = \mathrm{diag}\{L_2\}$,

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \otimes I_K \xrightarrow{P} \overline{\Lambda} \otimes I_K \quad \text{where} \quad \overline{\Lambda} = \begin{bmatrix} 0 & 0 \\ 0 & \overline{\Lambda}_2 \end{bmatrix}, \overline{\Lambda}_2 \succ 0. \tag{8}$$

One can now develop an asymptotic result which shows non-negligible shrinkage in the limit distribution of the ridge estimator. For simplicity of exposition, here I will assume that ridge centering $\boldsymbol{\beta}_0$ is chosen to be zero.

Theorem 6. In the setting of Theorem 1, assume that, for $0 < n \leq p$,

[5] This result follows from a straightforward generalization of Lemma 1 in Supplementary Appendix C. The choice of norm to measure such decay is not fundamental, as they are equivalent given that dimension K is fixed.

[6] Such an approach to inference also follows De Mol *et al.* (2008), who argue for explicit lag penalization within BVAR priors on similar theoretical grounds. In the context of maximum-likelihood estimation, the use of appropriate and plausible model restrictions to improve efficiency by shrinkage, rather than perform hypothesis testing, has also been discussed by Hansen (2016a).

[7] 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. This choice is also without loss of generality and simplifies exposition.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12737

(i) $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ where $\boldsymbol{\beta}_1 \in \mathbb{R}^{K^2(p-n)}$ and $\boldsymbol{\beta}_2 = T^{-(1/2+\delta)}\, \boldsymbol{b}_2$ for $\delta > 0$, $\boldsymbol{b}_2 \in \mathbb{R}^{K^2 n}$ fixed.
(ii) $\Lambda = \text{diag}\{(L_1', L_2')'\}$ where $L_1 \in \mathbb{R}_+^{K^2(p-n)}$ and $L_2 \in \mathbb{R}_+^{K^2 n}$.
(iii) $L_1 = o_P(T^{-1/2})$ and $L_2 \xrightarrow{P} \overline{L}_2$ as $T \to \infty$.
(iv) $\boldsymbol{\beta}_0 = 0$.

Let $\Gamma_{\overline{\Lambda}} = \Gamma + \overline{\Lambda}$ where $\overline{\Lambda} \succcurlyeq 0$ is given by (8). Then, results (a)-(c) hold and

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

It is easy to see that indeed the term $\Gamma_{\overline{\Lambda}}^{-1}\, \Gamma\, \Gamma_{\overline{\Lambda}}^{-1}$ in Theorem 6 is weakly smaller than Γ^{-1} in the positive-definite sense. Note that

$$\begin{aligned}
\Gamma_{\overline{\Lambda}}^{-1}\, \Gamma\, \Gamma_{\overline{\Lambda}}^{-1} \prec \Gamma^{-1} &\iff (\Gamma + \overline{\Lambda})^{-1}\Gamma \prec \Gamma^{-1}(\Gamma + \overline{\Lambda}) \\
&\iff I_{K^2 p} - (\Gamma + \overline{\Lambda})^{-1}\overline{\Lambda} \prec I_{K^2 p} + \Gamma^{-1}\overline{\Lambda} \\
&\iff 0 \preceq ((\Gamma + \overline{\Lambda})^{-1} + \Gamma^{-1})\overline{\Lambda}.
\end{aligned}$$

The last inequality is true by definition of $\overline{\Lambda}$. Shrinkage gains are concentrated at the components that have non-zero asymptotic shrinkage, i.e. those penalized by $\overline{L}_2$.

Remark 6. A key point in the application of Theorem 6 is identification of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. In practice, one may then proceed in two ways. As discussed in Section 4, one can see the ridge approach as a frequentist 'counterpart' to implementing a Bayesian prior. Therefore, the researcher may split $\boldsymbol{\beta}$ into subsets of small and large parameters based on economic intuition, domain knowledge or preliminary information. Alternatively, in the following section, I show that cross-validation is able to automatically tune Λ appropriately.

Finally, it is immediate to generalize the argument of Theorem 6 to the case where $\boldsymbol{\beta}$ is not split into subsets based on the relative size of coefficients, but rather a non-zero, *partially consistent* centering sequence $\boldsymbol{\beta}_0$ is used.

Corollary 2. Consider the setup of Theorem 6, where now assumptions (i) and (iv) are replaced by

(i′) $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ where $\boldsymbol{\beta}_1 \in \mathbb{R}^{K^2(p-n)}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{K^2 n}$ are fixed.
(iv′) $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', \boldsymbol{\beta}_{02}')'$ where $\boldsymbol{\beta}_{01} \in \mathbb{R}^{K^2(p-n)}$ is such that $\boldsymbol{\beta}_{01} \neq \boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_{02} = \boldsymbol{\beta}_2 + T^{-(1/2+\delta)}\boldsymbol{b}_2$ for $\delta > 0$, $\boldsymbol{b}_2 \in \mathbb{R}^{K^2 n}$ fixed.

Then, results (a)-(c) and (d″) still hold.

6.1. Cross-validation with Partitioned Coefficients

One can use the same approach applied to derive Theorem 5 to show that cross-validating the RLS estimator with $\text{Err}(\hat{\boldsymbol{\beta}}_{\blacklozenge}^R(\Lambda))$ is also asymptotically valid under partitioning.

Corollary 3. Consider the setup of Theorem 6 and assume that the assumptions of Theorem 5 are met. It holds

$$\begin{bmatrix} \Lambda_{1,\blacklozenge} & 0 \\ 0 & \Lambda_{2,\blacklozenge} \end{bmatrix} := \arg\min_{\Lambda \in \mathcal{I}_\lambda} \text{Err}\left(\hat{\boldsymbol{\beta}}_{\blacklozenge}^R(\Lambda)\right) = \begin{bmatrix} o_p(T^{-1/2}) & 0 \\ 0 & o_P(1) \end{bmatrix}$$

Moreover, any $\Lambda_{2,\blacklozenge}$ such that $0 \preceq \Lambda_{2,\blacklozenge} \preceq \lambda I$ is asymptotically valid.

In theory, one would like to be able to quantify the gains obtained in the asymptotic shrinkage setup of Theorem 6 compared to the standard setting of Theorems 1 and 2, particularly when using cross-validation. Unfortunately, it is

in general hard to study the cross-validation error loss even in setups without dependence. Stephenson *et al.* (2021), in fact, show that the ridge leave-one-out CV loss is not generally convex. This suggests that studying the behavior of CV when penalizing with a diagonal anisotropic Λ can be a very complex task in a finite sample setup.

7. SIMULATIONS

To study the performance of ridge-regularized estimators, I now perform simulation exercises focused on impulse response functions (IRFs). Throughout the experiments I will consider structural impulse responses, and I assume that identification can be obtained in a recursive way (Kilian and Lütkepohl, 2017), which is a widely used approach for structural shock identification in macroeconometrics.

I consider two setups:

1. The three-variable VARMA(1,1) design of Kilian and Kim (2011), representing a small-scale macro model. I term this setup 'A'.
2. A VAR(5) model in levels, using the model specification of Giannone *et al.* (2015) with the dataset of Hansen (2016b) consisting of $K = 7$ variables in levels.[8] I term this setup 'B'. For the ease of exposition, in the discussion I will tabulate results only for three variables – real GDP, investment and federal funds rate – but complete tables can be found in Supplementary Appendix D.5.

The specification of Kilian and Kim (2011) has already been extensively used in the literature as a benchmark to gauge the basic properties of inference methods. On the other hand, the estimation task of Giannone *et al.* (2015) involves more variables and a higher degree of persistence. This setting is useful to evaluate the effects of ridge shrinkage when applied to realistic macroeconomic questions. It is also a suitable test bench to compare Bayesian methods with frequentist ridge.

7.1. Estimators

For frequentist methods, I include both $\hat{\boldsymbol{\beta}}^{R}$ and $\hat{\boldsymbol{\beta}}^{\text{RGLS}}$ ridge estimators as well as the local projection estimator of Jordà (2005). For Bayesian methods, I implement both the Minnesota prior approach of Bańbura *et al.* (2010) with stationary prior and the hierarchical prior BVAR of Giannone *et al.* (2015).[9] The full list of method I consider is given in Table I. To make methods comparable, I have extended the ridge estimators to include an intercept in the regression. A precise discussion regarding the tuning of penalties and hyperparameters of all methods can be found in Supplementary Appendix D.

7.2. Pointwise MSE

The first two simulation designs explore the MSE performance of ridge-type estimators versus alternatives. Let $\theta_{km}(h)$ be the horizon h structural IRF for variable k given a unit shock from variable m. To compute the MSE for each k, define

$$\text{MSE}_k(h) := \sum_{m=1}^{K} \mathbb{E}\left[\left(\hat{\theta}_{km}(h) - \theta_{km}(h)\right)^2\right],$$

which is the total MSE for the kth variable over all possible structural shocks. In simulations, I use B replications to estimate the expectation. All MSEs are normalized by the mean squared error of the least squares estimator.

[8] The dataset is supplied by the author at https://users.ssc.wisc.edu/~bhansen/progs/var.html. While the data provided by Hansen (2016b) includes releases until 2016, I do not include more recent quarterly data since this is a simulation exercise. Moreover, due to the effects of the COVID-19 global pandemic, an extended sample would likely only add data released until Q4 2019 due to overwhelming concerns of a break point.

[9] To estimate hierarchical prior BVARs I rely on the original MATLAB implementation provided by Giannone *et al.* (2015) on the authors' website at http://faculty.wcas.northwestern.edu/gep575/GLPreplicationWeb.zip.

Table I. List of estimation methods

Type	Name	Description
Frequentist	LS	Least squares estimator
	RIDGE	Ridge estimator, CV penalty
	RIDGE-GLS	GLS ridge estimator, CV penalty
	RIDGE-AS	Ridge estimator with asymptotic shrinkage, CV penalty
	LP	Local projections with Newey-West covariance estimate
Bayesian	BVAR-CV	Litterman-Minnesota Bayesian VAR, CV tightness prior
	H-BVAR	Hierarchical Bayesian VAR of Giannone *et al.* (2015)

Table II. MSE relative to OLS – Setup A

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Investment	RIDGE	0.97	0.74	0.64	0.64	0.65	0.63	0.60
growth	RIDGE-GLS	5.16	0.89	0.55	0.47	0.44	0.41	0.38
	LP	1.00	1.05	1.13	1.52	2.15	3.20	4.87
	BVAR-CV	1.55	0.84	0.70	0.70	0.71	0.70	0.66
	H-BVAR	1.80	0.66	0.53	0.52	0.54	0.53	0.50
Deflator	RIDGE	0.93	0.78	0.69	0.68	0.67	0.64	0.59
	RIDGE-GLS	2.43	0.83	0.59	0.52	0.48	0.44	0.40
	LP	1.00	1.05	1.13	1.44	1.99	2.90	4.47
	BVAR-CV	1.03	0.89	0.74	0.73	0.73	0.70	0.66
	H-BVAR	1.01	0.70	0.58	0.56	0.55	0.53	0.50
Paper rate	RIDGE	0.94	0.76	0.66	0.66	0.66	0.64	0.60
	RIDGE-GLS	1.80	0.87	0.59	0.52	0.47	0.43	0.39
	LP	1.00	1.05	1.13	1.46	1.99	2.86	4.31
	BVAR-CV	0.87	0.87	0.74	0.73	0.73	0.71	0.66
	H-BVAR	0.81	0.69	0.57	0.55	0.56	0.54	0.51

7.2.1. Setup A

A time series of length $T = 200$ is generated a number $B = 10{,}000$ of times for replication. All VAR estimators are computed using $p = 10$ lags, while LPs include $q = 10$ regression lags. Table II shows relative MSEs for this design. It is important to notice that, in this situation, GLS ridge has remarkably low performance at horizon $h = 1$ compared to other methods. The primary issue is that Σ_u features strong correlation between components, and thus the diagonal lag-adapted structure does not shrink along the appropriate directions. This is much less prominent as the horizon increases due to the fact that impulse responses eventually decay to zero, since the underlying VARMA DGP is stationary. While there is no clear ranking, the MSE of the baseline ridge VAR estimator is in between those of the BVAR and hierarchical BVAR approaches. The degrading quality of local projection estimates are mainly due to the smaller samples available in regressions at each increasing horizon (Kilian and Kim, 2011). This behavior is one of the prime reasons behind the development of LP shrinkage estimators, like that proposed in Plagborg-Møller (2016) or the SLP estimator of Barnichon and Brownlees (2019).

7.2.2. Setup B

Using the data of Hansen (2016b), I estimate and simulate a stationary but highly persistent VAR(5) model using the same sample size and number of replications as Setup A. For all methods, $p = 5$ lags are used, so that VAR estimators are correctly specified. The results can be found in Table III. In this setup, unlike in the previous experiment, one can clearly notice that impulse responses computed via cross-validated ridge show increasing MSE as horizon h grows. There are two main reasons behind this behavior. First, the chosen setup features a very persistent data generating process, as the largest root of the underlying VAR model is 0.9945. This means that the true IRFs

DOI: 10.1111/jtsa.12737 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

Table III. MSE relative to OLS – Setup B

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	RIDGE	1.11	1.08	1.16	1.06	0.90	0.89	0.94
	RIDGE-GLS	1.16	1.00	0.99	1.00	0.93	0.93	0.95
	LP	1.00	1.14	1.37	1.52	1.72	1.98	2.24
	BVAR-CV	0.90	0.87	1.04	1.01	0.92	0.92	0.98
	H-BVAR	0.83	0.62	0.78	0.73	0.62	0.62	0.68
Investment	RIDGE	1.49	1.27	1.17	0.99	0.70	0.73	1.61
	RIDGE-GLS	1.34	1.14	1.02	1.02	0.86	0.82	0.86
	LP	1.00	1.15	1.40	1.63	2.03	2.76	3.59
	BVAR-CV	1.51	1.01	0.97	0.97	0.93	1.08	1.24
	H-BVAR	1.06	0.68	0.69	0.66	0.63	0.87	1.14
Fed funds	RIDGE	2.17	1.21	0.96	0.93	1.03	4.00	53.18
rate	RIDGE-GLS	1.21	1.04	0.90	0.93	0.90	0.88	0.91
	LP	1.00	1.18	1.51	1.71	1.97	2.44	2.99
	BVAR-CV	0.92	0.94	0.91	0.90	0.86	0.87	0.92
	H-BVAR	0.75	0.77	1.32	1.38	1.25	1.15	1.20

revert to zero only over long horizons, while lag-adapted ridge estimates yields models with lower persistence and thus flatter impulse responses. Second, the dataset from Hansen (2016b) is not normalized, and the included series have markedly heterogenous variances. Since GLS ridge shrinks along covariance-rotated data, shrinkage is adjusted according to each series variance, unlike that baseline ridge estimator $\hat{\boldsymbol{\beta}}^R$. The MSE for the Fed Fund Rate impulse responses shows that the pointwise difference between baseline and GLS ridge can be severe for long horizon IRFs when the DGP is highly persistent. On short horizons, Bayesian estimators perform on par or better than baseline least squares estimates, while at longer horizons differences are less stark. It is, however, clear that the hierarchical prior BVAR of Giannone *et al.* (2015) shows the overall best results. As in the previous setup, local projections show degrading performance at larger horizons.

Remark 7. The comparison between methods in both Setup A and Setup B is largely consistent with the findings of Li *et al.* (2024), who make extensive computational simulations by simulating from synthetic DGPs. They provide a comprehensive treatment of the question of which model – VAR or LP – is best suited for IRF inference in a given scenario in terms of bias-variance trade-off. They show that a key balance of bias vs. variance exists between LP and VAR estimates of impulse responses: LPs tend to have low bias, due to their flexibility, but they also feature large variance at higher horizons. Their results allow one to better understand the trade-offs at play in Tables II and III. In particular, it is clear that ridge shrinkage is beneficial at short horizons only if the penalization scheme is well-adapted to the DGP at hand. Otherwise, as is the case for RIDGE and RIDGE-GLS methods, the induced bias can be such that ridge MSEs surpass that of OLS estimates. One also finds that the medium and long horizons MSE gains over LPs are more pronounced in cases of moderate dependence, but in the case of the Federal Funds Rate IRFs in Setup B zero-centered RIDGE estimates thoroughly mistake long-term behavior.

7.3. Confidence Intervals

I now try and evaluate whether ridge shrinkage has a negative impact on inference. There have also been recent contributions directly aimed at studying shrinkage effects. Using the same simulation setups as in the previous section, I investigate coverage and size properties of pointwise CIs constructed using the methods in Table I. All confidence intervals are constructed with nominal 90% level coverage.

In this set of simulations, I swap GLS ridge for the asymptotic shrinkage ridge estimator, $\hat{\boldsymbol{\beta}}^R_{as}$, see Section 6, since the latter allows for a partially non-negligible penalization in the limit. To implement $\hat{\boldsymbol{\beta}}^R_{as}$, one needs to

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12737

Table IV. Impulse response inference – Setup A – CI coverage

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Investment	LS	0.88	0.88	0.87	0.88	0.91	0.93	0.94
growth	RIDGE	0.90	0.92	0.94	0.93	0.94	0.95	0.95
	RIDGE-AS	0.90	0.92	0.88	0.88	0.88	0.89	0.89
	LP	0.88	0.97	0.99	0.99	0.99	0.99	0.99
	BVAR-CV	0.77	0.88	0.88	0.90	0.92	0.94	0.96
	H-BVAR	0.72	0.89	0.89	0.92	0.93	0.95	0.96
Deflator	LS	0.88	0.87	0.86	0.88	0.91	0.92	0.94
	RIDGE	0.91	0.92	0.93	0.92	0.93	0.94	0.95
	RIDGE-AS	0.91	0.91	0.88	0.88	0.87	0.87	0.88
	LP	0.88	0.97	0.99	0.99	0.99	0.99	1.00
	BVAR-CV	0.80	0.86	0.88	0.91	0.93	0.94	0.96
	H-BVAR	0.84	0.88	0.90	0.92	0.94	0.95	0.97
Paper rate	LS	0.87	0.86	0.86	0.88	0.90	0.92	0.94
	RIDGE	0.90	0.91	0.93	0.93	0.93	0.94	0.95
	RIDGE-AS	0.89	0.90	0.89	0.88	0.88	0.88	0.88
	LP	0.87	0.97	0.99	0.99	0.99	0.99	0.99
	BVAR-CV	0.82	0.84	0.87	0.90	0.92	0.93	0.95
	H-BVAR	0.88	0.88	0.90	0.92	0.93	0.95	0.96

choose a partition of $\boldsymbol{\beta}$ which identifies asymptotically negligible coefficient. To do this, I split $\boldsymbol{\beta}$ by lag and penalize all coefficients with lag orders greater than a given threshold $\overline{p}$, such that $1 < \overline{p} < p$. In setup A, I choose $\overline{p} = 6$, while in setup B I set $\overline{p} = 3$. In Bayesian methods, including the cross-validated Minnesota BVAR, I construct high-probability intervals by drawing from the posterior. Comparison between frequentist CIs and Bayesian posterior densities is not generally valid, because they are not analogous concepts. Therefore, the discussion below is intended to highlight differences in *structure* between ridge approaches.

7.3.1. Setup A

Simulations with the DGP of Kilian and Kim (2011), presented in Tables IV and V, highlight some of the advantages of applying ridge when performing inference. Focusing on estimator $\hat{\boldsymbol{\beta}}^R$, it is clear that CI coverage is in fact higher than the intervals obtained by least squares estimation in all situations. At impact, ridge CIs are larger than the LS baseline, but they shrink as horizons increase. Thus, is IRFs revert relatively quickly to zero, ridge can effectively reduce length while preserving coverage. As discussed in Section 3, these gains are inherently local to the DGP – shrinkage to zero at deep lags embodies correct prior knowledge of a weakly persistent process. For Bayesian estimators, one can note that quantile intervals at small horizons tend to be shorter compared to least squares and ridge methods.

7.3.2. Setup B

The effects of ridge shrinkage on a DGP with high persistence are much more severe, as shown in Tables VI and VII. Focusing on frequentist ridge, one can observe that close to impact ($h = 1$) ridge has similar or even higher coverage than other methods for real GDP[10] However, as the IRF horizon grows, shrinkage often leads to severe undercoverage, with asymptotic shrinkage estimator $\hat{\boldsymbol{\beta}}^R_{as}$ giving the worst results. In comparison, Bayesian methods are much more reliable at all horizons, although the only estimator that can consistently improve on the benchmark least squares VAR CIs is the hierarchical prior BVAR of Giannone *et al.* (2015). The reason behind this is simple enough: the implementation of the Minnesota-prior BVAR I have used here has a white noise prior on all variables, which in this case is far from the truth. Indeed, Bańbura *et al.* (2010) implement the same BVAR by tuning the

[10] This also is the case with consumption and compensation, see also Tables 9 and 10 in Supplementary Appendix D.5.

Table V. Impulse response inference – Setup A – CI length

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Investment	LS	2.99	5.11	5.78	5.35	4.79	4.17	3.56
growth	RIDGE	3.13	5.20	5.82	5.17	4.48	3.78	3.09
	RIDGE-AS	3.11	5.15	4.84	4.33	3.70	3.06	2.48
	LP	2.99	7.50	10.97	12.89	13.99	14.55	14.70
	BVAR-CV	2.84	4.48	4.70	4.38	3.99	3.56	3.11
	H-BVAR	2.71	4.20	4.50	4.29	3.96	3.56	3.13
Deflator	LS	1.19	1.92	2.23	2.14	1.94	1.71	1.46
	RIDGE	1.24	1.97	2.25	2.09	1.84	1.54	1.26
	RIDGE-AS	1.24	1.95	1.95	1.78	1.52	1.25	1.01
	LP	1.19	3.03	4.56	5.42	5.90	6.14	6.21
	BVAR-CV	1.03	1.69	1.87	1.80	1.67	1.50	1.31
	H-BVAR	1.01	1.64	1.83	1.79	1.67	1.51	1.33
Paper rate	LS	0.97	1.42	1.64	1.57	1.44	1.27	1.09
	RIDGE	1.01	1.44	1.65	1.53	1.36	1.16	0.95
	RIDGE-AS	1.01	1.43	1.42	1.31	1.13	0.94	0.77
	LP	0.97	2.19	3.28	3.90	4.26	4.43	4.48
	BVAR-CV	0.84	1.22	1.35	1.30	1.21	1.09	0.96
	H-BVAR	0.85	1.21	1.34	1.31	1.22	1.10	0.97

Table VI. Impulse response inference – Setup B: CI coverage

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	LS	0.87	0.81	0.75	0.72	0.71	0.72	0.73
	RIDGE	0.90	0.79	0.66	0.62	0.65	0.68	0.68
	RIDGE-AS	0.89	0.72	0.61	0.58	0.61	0.65	0.65
	LP	0.87	0.93	0.94	0.94	0.93	0.93	0.91
	BVAR-CV	0.70	0.71	0.63	0.64	0.71	0.75	0.76
	H-BVAR	0.84	0.86	0.76	0.76	0.83	0.88	0.88
Investment	LS	0.87	0.82	0.76	0.73	0.75	0.82	0.87
	RIDGE	0.85	0.79	0.65	0.62	0.73	0.80	0.81
	RIDGE-AS	0.82	0.69	0.59	0.57	0.68	0.77	0.77
	LP	0.87	0.94	0.94	0.95	0.94	0.94	0.94
	BVAR-CV	0.70	0.73	0.67	0.71	0.77	0.81	0.83
	H-BVAR	0.80	0.86	0.81	0.82	0.87	0.88	0.88
Fed funds	LS	0.85	0.83	0.80	0.78	0.77	0.79	0.80
rate	RIDGE	0.79	0.77	0.74	0.68	0.68	0.72	0.72
	RIDGE-AS	0.78	0.66	0.68	0.64	0.64	0.68	0.69
	LP	0.85	0.94	0.96	0.96	0.95	0.94	0.93
	BVAR-CV	0.76	0.72	0.76	0.77	0.77	0.81	0.83
	H-BVAR	0.87	0.86	0.74	0.73	0.78	0.84	0.87

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12737

Table VII. Impulse response inference – Setup B: CI length (rescaled ×100)

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	LS	0.71	1.56	2.07	2.31	2.32	2.24	2.15
	RIDGE	0.79	1.56	1.85	1.95	1.92	1.85	1.77
	RIDGE-AS	0.74	1.31	1.65	1.76	1.75	1.70	1.64
	LP	0.71	2.42	4.21	5.40	5.90	5.91	5.70
	BVAR-CV	0.53	1.23	1.74	2.00	2.10	2.13	2.15
	H-BVAR	0.58	1.36	1.87	2.16	2.32	2.44	2.55
Investment	LS	3.38	6.65	7.89	7.89	7.31	6.69	6.18
	RIDGE	3.79	6.81	6.93	6.46	5.79	5.19	4.73
	RIDGE-AS	3.59	5.57	6.11	5.77	5.21	4.72	4.34
	LP	3.37	10.16	16.00	18.85	19.06	18.22	17.23
	BVAR-CV	2.64	5.26	6.59	6.91	6.78	6.57	6.38
	H-BVAR	2.89	5.74	7.08	7.54	7.63	7.60	7.58
Fed funds	LS	0.25	0.39	0.43	0.43	0.41	0.38	0.35
rate	RIDGE	0.29	0.39	0.37	0.36	0.33	0.30	0.29
	RIDGE-AS	0.27	0.31	0.33	0.32	0.30	0.28	0.27
	LP	0.25	0.59	0.88	1.01	1.05	1.03	0.98
	BVAR-CV	0.21	0.31	0.36	0.37	0.36	0.35	0.34
	H-BVAR	0.23	0.36	0.42	0.44	0.45	0.45	0.46

prior to a random walk for very persistent variables in their applications. In this sense, the cross-validated BVAR considered – which is assumed centered at zero – is really the flip-side of ridge estimators. Therefore, the addition of a prior on the mean of the autoregressive parameters as done by Giannone *et al.* (2015) is a key element to perform shrinkage in high persistence setups in a way that does not systematically undermine asymptotic inference on impulse responses.

8. CONCLUSION

In this article, I have studied ridge regression and its application to vector autoregressive model estimation in detail. This appears to be the first work that provides a thorough analysis of ridge penalization in the context of time series data, including geometric as well as asymptotic properties. I have also derived results on the validity of cross-validation as a method to select the penalty intensity in practice, and I have shown that CV produces asymptotically valid penalization rates. Finally, I have compared both frequentist and Bayesian ridge formulation in simulations aimed at quantifying the applicability of ridge for impulse response inference.

The key takeaway of this article is that ridge penalization is a useful approach to VAR estimation as long as the chosen penalty structure is well-adapted to the model's structure. Bayesian ridge posteriors are especially flexible, with hierarchical priors also allowing shrinkage toward non-zero coefficient vectors. However, it is important to note that the Bayesian approach also permits the researcher to specify uninformative priors, so that the influence of the priors' hyperparameters is less pronounced. This is not the case in frequentist ridge, cf. including an explicit non-zero centering vector. However, prior knowledge or a pre-estimation procedure may be available to the researcher, so that ridge can be effectively implemented without the need to implement a BVAR.

To conclude, there are still avenues of research regarding ridge which would be interesting to develop. First and foremost, the high-dimensional setup, for which, however, it seems non-trivial to find a domain of applicability, as discussed in the introduction. Second, a more in-depth analysis of cross-validation, especially in the multi-variate case, would be extremely valuable. Moreover, both the latter and former topics should be jointly addressed in the context of mild cross-sectional dimension growth, i.e., $K \to \infty$ such that $K/T \to \rho \in (0, 1)$, which is comparable to factor model setups.

DOI: 10.1111/jtsa.12737 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

ACKNOWLEDGEMENTS

I am grateful for the comments and suggestions from Lyudmila Grigoryeva, So Jin Lee, Thomasz Olma, Oliver Pfäuti and Mikkel Plagborg-Møller, and the seminar participants at the University of Mannheim, the HKMetrics Workshop and the Young Researchers Workshop on Big and Smart Data Analysis in Finance. I am especially thankful to Claudia Noack for pointing out an important error in a previous version of this article, as well as Jonas Krampe and Carsten Trenkler for their insightful discussions which helped develop this article significantly. Lastly, I wish to thank Peter C. B. Phillips, Atsushi Inoue and many other colleagues for the suggestion to consider adding a formal analysis of cross-validation in the article. The author acknowledges support by the state of Baden–Württemberg through bwHPC. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The author reports that there are no competing interests to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in GitHub at https://github.com/giob1994/ridge_var.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

REFERENCES

Babii A, Ghysels E, Striaukas J. 2021. Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics* **40**(3):1–23.

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

Barnichon R, Brownlees C. 2019. Impulse response estimation by smooth local projections. *The Review of Economics and Statistics* **101**(3):522–530.

Bauwens L, Chevillon G, Laurent S. 2023. We modeled long memory with just one lag! *Journal of Econometrics* **236**(1):105467.

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

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

Brockwell PJ, Davis RA. 1991. *Time Series: Theory and Methods* New York: Springer Science + Business Media.

Burman P, Chow E, Nolan D. 1994. A cross-validatory method for dependent data. *Biometrika* **81**(2):351–358.

Cavaliere G, Gonçalves S, Nielsen MØ, Zanelli E. 2023. Bootstrap inference in the presence of bias. *Journal of the American Statistical Association* 1–11. https://doi.org/10.1080/01621459.2023.2284980

Davidson J. 1994. *Stochastic Limit Theory: An Introduction for Econometricians* New York: Oxford University Press.

De Mol C, Giannone D, 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.

Dobriban E, Wager S. 2018. High-dimensional asymptotics of prediction: ridge regression and classification. *Annals of Statistics* **46**(1):247–279.

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

Fuleky P (ed.). 2020. *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*. *Advanced Studies in Theoretical and Applied Econometrics*, Vol. 52 Springer International Publishing, Cham.

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

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

Goulet Coulombe P. 2023. Time-varying parameters as ridge regressions. *Working Paper*.

Goulet Coulombe P, Leroux M, Stevanovic D, Surprenant S. 2022. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics* **37**(5):920–964.
Hamilton JD. 1994. *Time Series Analysis* Princeton, NJ: Princeton University Press.
Hansen BE. 2016a. Efficient shrinkage in parametric models. *Journal of Econometrics* **190**(1):115–132.
Hansen BE. 2016b. Stein combination shrinkage for vector autoregressions. *Working Paper*.
Hastie T. 2020. Ridge Regularization: an essential concept in data science. *Technometrics* **62**(4):426–433.
Hastie T, Tibshirani R, Friedman JH, Friedman JH. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second ed. New York: Springer.
Hastie T, Montanari A, Rosset S, Tibshirani RJ. 2022. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics* **50**(2):949–986.
Hausman JA. 1983. *Chapter 7: specification and estimation of simultaneous equation models*. In *Handbook of Econometrics*, Vol. 1, Amsterdam: Elsevier; 391–448.
Heaps SE. 2023. Enforcing stationarity through the prior in vector autoregressions. *Journal of Computational and Graphical Statistics* **32**(1):74–83.
Huber F, Koop G. 2023. Subspace shrinkage in conjugate Bayesian vector autoregressions. *Journal of Applied Econometrics* **38**(4):556–576.
Inoue A, 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.
Kadiyala KR, Karlsson S. 1997. Numerical methods for estimation and inference in bayesian var-models. *Journal of Applied Econometrics* **12**(2):99–132.
Kilian L, Kim YJ. 2011. How reliable are local projection estimators of impulse responses? *Review of Economics and Statistics* **93**(4):1460–1466.
Kilian L, Lütkepohl H. 2017. *Structural Vector Autoregressive Analysis* Cambridge: Cambridge University Press.
Li D, Plagborg-Møller M, Wolf CK. 2024. Local projections vs. vars: lessons from thousands of DGPs. *Working Paper*.
Litterman RB. 1986. Forecasting with Bayesian vector autoregressions five years of experience. *Working Papers No. 274*, Federal Reserve Bank of Minneapolis.
Liu S, Dobriban E. 2020. Ridge Regression: Structure, Cross-Validation, and Sketching. In *International Conference on Learning Representations*.
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* Heidelberg: Springer Berlin.
Medeiros MC, Vasconcelos GFR, Veiga Á, 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.
Park JY, Phillips PC. 1988. Statistical inference in regressions with integrated processes: part 1. *Econometric Theory* **4**(3):468–497.
Park JY, Phillips PC. 1989. Statistical inference in regressions with integrated processes: part 2. *Econometric Theory* **5**(1):95–131.
Patil P, Wei Y, Rinaldo A, Tibshirani R. 2021. *Uniform consistency of cross-validation estimators for high-dimensional ridge regression*. In Banerjee A, & Fukumizu, K (Eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, PMLR; 3178–3186.
Pesavento E, Rossi B. 2006. Small-sample confidence intervals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics* **21**(8):1135–1155.
Phillips PCB. 1988. Regression theory for near-integrated time series. *Econometrica* **56**(5):1021–1043.
Plagborg-Møller M. 2016. Essays in Macroeconometrics. PhD thesis, Harvard University.
Pratt JW. 1961. Length of confidence intervals. *Journal of the American Statistical Association* **56**(295):549–567.
Sims CA, Stock JH, Watson MW. 1990. Inference in linear time series models with some unit roots. *Econometrica: Journal of the Econometric Society* **58**(1):113–144.
Smeekes S, Wijler E. 2018. Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting* **34**(3):408–430.
Stephenson W, Frangella Z, Udell M, Broderick T. 2021. *Can we globally optimize cross-validation loss? Quasiconvexity in ridge regression*. In Ranzato M, Beygelzimer A, Dauphin Y, Liang PS & Wortman Vaughan J. *Advances in Neural Information Processing Systems*, Vol. 34, Curran Associates, Inc; 24352–24364.

DOI: 10.1111/jtsa.12737 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.