Stein Combination Shrinkage for Vector Autoregressions

Bruce E. Hansen*
University of Wisconsin†

June 2016
Preliminary

Abstract

This paper introduces Stein combination shrinkage for vector autoregressions (VARs). The proposed methods shrink unrestricted least-squares VAR estimates towards multiple user-specified linear constraints, including lag exclusion and autoregressive models. We propose weighted combination estimators, where the weights minimize an estimate of the mean-squared error (MSE) of a vector-valued parameter of interest. Particular attention is given to impulse response estimation and multi-period point forecasting. The combination estimators are similar to Stein shrinkage estimators. Our proposed weights are specific to the horizon, which allows the degree of shrinkage to adapt across horizons.

The proposed methods are evaluated in a careful simulation experiment. The simulation evidence shows that the Stein combination methods have much lower MSE than conventional OLS and BVAR methods. We illustrate the methods with an application to a standard seven-variable system of U.S. macroeconomic aggregates.

*Research supported by the National Science Foundation. Thanks to Wei Song for excellent research assistance, and to Mark Watson for suggesting this research idea to me.

†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706.

1 Introduction

A core tool for macroeconomic evaluation is linear vector autoregressive (VAR) models. Popularized by the seminal work of Sims (1980), VAR models are routinely used for model comparison and analysis, forecasting, and impulse response analysis.

VAR models are typically estimated either by least-squares or Bayesian methods. The latter (typically referred to as Bayesian VAR or BVAR) have been particularly emphasized to counter the high-dimensionality of VAR models, especially in the presence of a moderately large number of variables. The Bayesian approach dates back to the work of Doan, Litterman and Sims (1984) and is enshrined in the popular "Minnesota Prior" which is centered at a random walk with drift. Influential papers include Sims and Zha (1998), Banbura, Giannone and Reichlin (2010) and recently Giannone, Lenza and Primiceri (2015).

Bayesian methods produce estimates with low Bayes risk, but can have unbounded classical (e.g. minimax) risk. In particular, Bayes methods are not designed to produce estimates for specific goals such as multi-step point forecasting or impulse response estimation. In contrast, classical shrinkage estimators of the James-Stein class can be designed for such purposes and are known to possess optimal minimax properties. In recent work, Hansen (2015) has shown how to develop Stein-type estimators taylored to specific user-selected loss functions. The goal of the present paper is to apply similar techniques to the VAR setting, developing Stein combination shrinkage estimators designed to minimize the weighted mean-squared error (MSE) of targeted parameters such as impulse response coefficients and multi-step point forecasting coefficients. We compare the Stein estimators with classical estimators (OLS and BVAR) in a simulation experiment calibrated to the "medium model" setting explored in Giannone, Lenza and Primiceri (2015).

Shrinkage methods depend on both a shrinkage direction (the model towards which to shrink) and a loss function (how to evaluate estimation efficiency). For our shrinkage direction, we consider constrained models based on lag restrictions and autoregressive models. This effectively shrinks the unrestricted models towards models with fewer lags, and towards an autoregression. In particular, shrinkage towards an AR(1) is important as this is similar to shrinking towards a random walk, which is the core of the Minnesota prior, and random walks are excellent forecasting defaults. One advantage of our combination shrinkage method is that the shrinkage direction adapts with the data.

Our shrinkage method is designed to minimize the mean-squared error of a targeted function of the coefficients. Leading examples include impulse response coefficients and multi-step point forecasting coefficients. Combination weights are selected to minimize an estimate of the MSE of the targeted parameter of interest. Since the combination weights depend on the parameter of interest, they (and the degree of shrinkage) will naturally vary with the latter, for example they will vary across the impulse response or forecast horizon. This is intentional, as it is desireable to impose greater shrinkage at longer horizons due to decreased estimation precision.

The methods developed here are a generalization of combination methods developed in the previous literature. Hansen (2008) proposed combination methods for one-step ahead forecasts based

on the Mallows criteria. Cheng and Hansen (2015) proposed combination methods for multi-step direct forecasts from factor models, using both Mallows and leave-h-out cross-validation criteria. Hansen (2015) developed an asymptotic theory of estimation efficiency for combination in the context of two estimators. Liu and Kuo (2016) propose a frequentist model averaging criteria for one-step-ahead forecasts. Liao and Tsay (2016) generalized Hansen's (2008) methods to VARs based on a one-step-ahead Mallows criteria. Of these papers, only Hansen (2015) considered the case where the parameters of interest are a nonlinear function of the regression coefficients, which is important for multi-step impulse responses and multi-step forecast coefficients.

This paper also relates to the growing literature on high-dimensional VARs. Shrinkage allows the consideration of larger models (more variables and more lags) than would be considered if estimation is OLS. Our proposed combination method is probably best suited to "medium" sized models, as it requires the estimation of the baseline "large" model by OLS, so the latter estimate must be feasible and reasonably accurate. Generalizations of the methods here to allow for high-dimensional models would be valuable. Some of the recent literature exploring VAR estimation with many variables includes Carriero, Galvao, and Kapetanios (2015), Koop, Korobilis and Pettenuzzo (2016), and Kapetanious, Marcellino and Venditti (2016).

The organization of the paper is as follows. Section 2 introduces the VAR model and notation. Section 3 presents least-squares estimation and its asymptotic distribution. Section 4 presents submodel and combination estimates, and the asymptotic distribution of the latter under fixed weights. Section 5 presents details on the constraint matrices. Section 6 derives the asymptotic MSE of the combination estimates, proposes an estimator of the asymptotic MSE, and shows that the estimator is an asymptotically unbiased estimate of the MSE. Section 7 proposes weight selection. Section 8 describes application of the method to multi-step point forecasting. Section 9 describes application to impulse response estimation. Section 10 is a simulation study, exploring the MSE of the impulse response estimates and the MSFE of the point forecasts in three simulation designs. The Stein combination methods are compared with unrestricted OLS and BVAR methods. Section 11 is an empirical application to a standard 7-variable model. Section 12 is a conclusion. Mathematical proofs are provided in the appendix.

Matlab code which produces the simulation and empirical work reported in the paper is posted on the author's website http://www.ssc.wisc.edu/~bhansen/

2 Model

Take a standard vector autoregressive (VAR) model with m variables and p lags

$$y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + B_0 + e_t \tag{1}$$

where y_t is an $m \times 1$ vector of endogenous variables, e_t is an $m \times 1$ of shocks, $B_1, ..., B_p$ are $m \times m$ coefficient matrices, and B_0 is $m \times 1$. Observations are available for $t = 1, ..., n$. This is a system of m equations, where each equation has $k = mp + 1$ coefficients. We assume that the shock vector

e_t satisfies $E_{t-1}e_t = 0$. We also define the covariance matrix of the shocks $\Sigma = E(e_t e_t')$.

It is convenient to write the VAR equation in the regression format

$$y_t = Bx_{t-1} + e_t$$

where

$$B = [B_1, \cdots, B_p, B_0]$$

is an $m \times k$ matrix of coefficients and

$$x_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \\ 1 \end{pmatrix}$$

is $k \times 1$.

It will also be useful to write the coefficients in the vector format $\theta = \text{vec}(B') = (\theta_1', ..., \theta_m')$, where θ_j' is the j^{th} row of B and hence is the coefficient for the j^{th} variable. Note that θ_j is $k \times 1$ and θ is $km \times 1$.

The goal is estimation of a $q \times 1$ differentiable function of the coefficients $\beta = g(\theta)$. Examples include multi-step point forecasts and impulse response estimation, as will be described in detail in Sections 8 and 9.

3 Least-Squares Estimation

The least-squares estimate of B, θ, β and Σ are

$$\widehat{B} = \left(\sum_{t=1}^{n} y_t x_{t-1}'\right)\left(\sum_{t=1}^{n} x_{t-1}x_{t-1}'\right)^{-1}$$

$$\widehat{\theta} = \text{vec}(\widehat{B}') = \begin{pmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_m \end{pmatrix}$$

$$\widehat{\beta} = g\left(\widehat{\theta}\right)$$

$$\widehat{\Sigma} = \frac{1}{n-k}\sum_{t=1}^{n} \widehat{e}_t \widehat{e}_t'$$

where $\widehat{e}_t = y_t - \widehat{B}x_{t-1}$ are the least-squares residuals.

Define the components of the asymptotic variance

$$\begin{aligned} V &= \left(I_m \otimes Q^{-1}\right) \Omega \left(I_m \otimes Q^{-1}\right) \\ Q &= E\left(x_{t-1} x_{t-1}'\right) \\ \Omega &= E\left(e_t e_t' \otimes x_{t-1} x_{t-1}'\right) \end{aligned}$$

and the least-squares estimates

$$\begin{aligned} \widehat{V} &= \left(I_m \otimes \widehat{Q}^{-1}\right) \widehat{\Omega} \left(I_m \otimes \widehat{Q}^{-1}\right) \\ \widehat{Q} &= \frac{1}{n} \sum_{t=1}^{n} x_{t-1} x_{t-1}' \\ \widehat{\Omega} &= \frac{1}{n-k} \sum_{t=1}^{n} \left(\widehat{e}_t \widehat{e}_t' \otimes x_{t-1} x_{t-1}'\right). \end{aligned}$$

The asymptotic distribution of the estimates is known (e.g. Hamilton (1994)), but the following formulation will be convenient for our development.

Theorem 1. *If y_t is strictly stationary, $E\left\|y_t\right\|^4 < \infty$, $E_{t-1} e_t = 0$ and $G(\theta) = \frac{\partial}{\partial \theta'} g(\theta)$ is continuous at the true value θ then*

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \to_d Z \sim N\left(0, V\right),$$

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \to_d G' Z,$$

$$\widehat{V} \to_p V$$

where $G = G(\theta)$.

4 Sub-Model and Combination Estimates

A sub-model of the VAR model (1) is any restriction on the coefficients, including variable exclusions, lag exclusions, order of integration constraints, and random walk constraints. We focus on linear constraints. These can be written as

$$R'\theta = a \tag{2}$$

where R is $k \times q$ and a is $q \times 1$. Each constraint can be viewed as a model (a parametric restriction on the VAR). We will index the models by r, and write the constraint matrices as $(R(r), a(r))$ for $r = 1, ..., M$ so that there are M models. We describe the form of the restriction matrices $R(r)$ for our recommended sub-models in Section 4.

For each sub-model the constrained least-squares estimator of θ can be written as

$$\widehat{\theta}(r) = \widehat{\theta} - \widehat{W}_\theta^{-1} R(r) \left(R(r)' \widehat{W}_\theta^{-1} R(r)\right)^{-1} \left(R(r)' \widehat{\theta} - a(r)\right)$$

where $\widehat{W}_\theta = I_m \otimes \widehat{Q}$. More generally for a arbitrary weight matrix $\widehat{W}_\theta$, $\widehat{\theta}(r)$ is the minimum distance estimator of θ under the constraint (2).

The constrained estimator $\widehat{B}(r)$ for B is constructed by stacking the estimates

$$\widehat{B}(r) = \begin{pmatrix} \widehat{\theta}_1(r)' \\ \widehat{\theta}_2(r)' \\ \vdots \\ \widehat{\theta}_m(r)' \end{pmatrix},$$

and that for β is $\widehat{\beta}(r) = g\left(\widehat{\theta}(r)\right)$. The residuals are $\widehat{e}_t(r) = y_t - \widehat{B}(r)x_{t-1}$ and an estimate of Σ is

$$\widehat{\Sigma}(r) = \frac{1}{n - \overline{k}} \sum_{t=1}^{n} \widehat{e}_t \widehat{e}_t'$$

where $\overline{k}$ is the average number of coefficients in each equation. In our applications all equations have the same number of coefficients but otherwise this an ad hoc degree of freedom adjustment.

A combination estimator assigns a weight $w(r)$ to each model where $w(r) \geq 0$ and $\sum_{r=1}^{M} w(r) = 1$. Set $\mathbf{w} = (w(1), ..., w(M))$. The combination estimator is

$$\widehat{\beta}(\mathbf{w}) = \sum_{r=1}^{M} w(r)\widehat{\beta}(r).$$

For an asymptotic distribution theory, we assume that the coefficients θ are local to the restrictions, thus $R(r)'\theta = a(r) + n^{-1/2}\delta(r)$

Theorem 2. *Under the assumptions of Theorem 1, plus* $R(r)'\theta = a(r) + n^{-1/2}\delta(r)$

$$\sqrt{n}\left(\widehat{\beta}(\mathbf{w}) - \beta\right) \to_d G' \left[\left(I_m - D(\mathbf{w})\right) Z + \delta(\mathbf{w})\right]$$

where

$$\begin{aligned} D(\mathbf{w}) &= \sum_{r=1}^{M} w(r) W_\theta^{-1} R(r) \left(R(r)' W_\theta^{-1} R(r)\right)^{-1} R(r)' \\ \delta(\mathbf{w}) &= \sum_{r=1}^{M} w(r) W_\theta^{-1} R(r) \left(R(r)' W_\theta^{-1} R(r)\right)^{-1} \delta(r) \\ W_\theta &= I_m \otimes Q \end{aligned}$$

Theorem 2 shows that that the combination estimator with fixed weights is asymptotototically normal with a bias component. The weights affect the balance between variance and bias. Efficient estimation requires a careful choice for these weights, and this is explored in Sections 6 and 7.

5 Constraint Matrices

The sub-model estimates $\widehat{\theta}(r)$ can easily be calculated by least-squares on regressor subsets. However, for some purposes the explicit constraint matrices $R(r)$ are required. We discuss two classes of contraints which we recommend for applications: lag restrictions and autoregressive restrictions. These are both exclusion restrictions for which $a(r) = 0$.

Lag restrictions correspond to VAR(r) models with $r < p$. The constraint matrices equal

$$\begin{aligned} R(r) &= I_m \otimes \overline{R}(r) \\ \overline{R}(r) &= \begin{bmatrix} 0_{mr \times m(p-r)} \\ I_{m(p-r)} \\ 0_{1 \times m(p-r)} \end{bmatrix}. \end{aligned}$$

Autoregressive restrictions correspond to AR(r) models with $r \leq p$. We can write the constraint matrices as

$$\begin{aligned} R(r) &= \operatorname{diag}\{R_1(r), ..., R_m(r)\} \\ R_\ell(r) &= \begin{pmatrix} I_r \otimes C_\ell & 0_{mr \times m(p-r)} \\ 0_{m(p-r) \times (m-1)r} & I_{m(p-r)} \\ 0_{1 \times (m-1)r} & 0_{1 \times m(p-r)} \end{pmatrix} \end{aligned}$$

where C_ℓ is the $m \times (m-1)$ matrix equalling the identity I_m with the ℓ^{th} column deleted.

The combination estimator $\widehat{\beta}(\mathbf{w})$ allows inclusion of a large set of constrained models. We recommend using the following models

1. VAR(1) through VAR(p)
2. AR(1) through AR(p)

Constrained sub-models allows the estimator to shrink the unrestricted VAR(p) specification towards the constrained model, which effectively shrinks the less-precisely-estimated coefficients. Inclusion of lag constraints allows the estimator to shrink the larger lag matrices. Inclusion of the autoregressive models allows the estimator to shrink towards an autoregression (which often produce excellent point forecasts). In particular, the AR(1) model allows shirnkage towards a model close to the random walk, which has been a successful approach in BVAR applications via the Minnesota prior.

6 Mean-Squared Error

The combination estimate $\widehat{\beta}(\mathbf{w})$ can be evaluated based on weighted squared error

$$S_n(\mathbf{w}) = n\left(\widehat{\beta}(\mathbf{w}) - \beta\right)' W_\beta \left(\widehat{\beta}(\mathbf{w}) - \beta\right)$$

where W_β is a $q \times q$ weight matrix. In some examples (such as point forecasting) the weight matrix will be naturally determined. In other cases, it can be selected by the user. If the components of $\widehat{\beta}$ have similar variance we can set $W_\beta = I_q$. If they have differing variances, or if it is desirable to render the criterion invariant to re-scaling, it is advisable to set $W_\beta = (G'VG)^{-1}$ the inverse of the asymptotic variance of the least-squares estimator for β. This is identical to the weight matrix choice used for efficient minimum distance estimation. We will recommend specific chocies for forecasting (Section 7) and impulse response estimation (Section 8).

Technically, the mean-squared error $E\left(S_n(\mathbf{w})\right)$ is difficult to evaluate so we analyze instead the asymptotic trimmed MSE:

$$R(\mathbf{w}) = \lim_{\zeta\to\infty} \liminf_{n\to\infty} E \min\left(S_n(\mathbf{w}), \zeta\right).$$

Theorem 3. *Under the assumptions of Theorem 2*

$$R(\mathbf{w}) = \delta(\mathbf{w})' G W_\beta G' \delta(\mathbf{w}) + \mathrm{tr}\left(W_\beta G' D(\mathbf{w}) V D(\mathbf{w})' G\right) - 2\sum_{r=1}^{M} w(r) K(r) + \mathrm{tr}\left(W_\beta G' V G\right)$$

where

$$K(r) = \mathrm{tr}\left(W_\beta G' W_\theta^{-1} R(r) \left(R(r)' W_\theta^{-1} R(r)\right)^{-1} R(r)' V G\right)$$

Theorem 3 shows that the mean-squared error of the combination estimator $\widehat{\beta}(\mathbf{w})$ can be approximated by $R(\mathbf{w})$. The first component represents the squared bias due to constrained estimation. The remaining components represent estimation variance.

Ideally, the optimal weights $\mathbf{w}$ should be selected to minimize $R(\mathbf{w})$. However the latter is unknown so such weights are infeasible. We thus propose estimation of $R(\mathbf{w})$, and then select weights which minimize this estimate. Our proposed estimator for $R(\mathbf{w})$ is

$$\widehat{R}(\mathbf{w}) = n\left(\widehat{\beta}(\mathbf{w}) - \widehat{\beta}\right)' \widehat{W}_\beta \left(\widehat{\beta}(\mathbf{w}) - \widehat{\beta}\right) - 2\sum_{r=1}^{M} w(r)\widehat{K}(r) + \mathrm{tr}\left(\widehat{W}_\beta \widehat{G}' \widehat{V} \widehat{G}\right) \tag{3}$$

with

$$\begin{aligned} \widehat{K}(r) &= \mathrm{tr}\left(\widehat{W}_\beta \widehat{G}' \widehat{W}_\theta^{-1} R(r) \left(R(r)' \widehat{W}_\theta^{-1} R(r)\right)^{-1} R(r)' \widehat{V} \widehat{G}\right) \\ \widehat{G} &= G(\widehat{\theta}) \end{aligned} \tag{4}$$

and $\widehat{W}_\beta$ is an estimate of W_β .

Theorem 4. *Under the assumptions of Theorem 2 and in addition* $\widehat{W}_\beta \to_p W_\beta$ *then*

$$\lim_{\zeta\to\infty} \liminf_{n\to\infty} E \min\left(\widehat{R}(\mathbf{w}), \zeta\right) = R(\mathbf{w})$$

Theorem 4 shows that $\widehat{R}(\mathbf{w})$ is an asymptotically unbiased estimate of the MSE. The weight

vector $\mathbf{w}$ which minimizes $\widehat{R}(\mathbf{w})$ is thus an estimate of the infeasible optimal weight vector $\mathbf{w}$ which minimizes $R(\mathbf{w})$.

7 Weight Estimation

We propose selecting the combination weights by minimizing the MSE estimate $\widehat{R}(\mathbf{w})$. Since the third term in (3) does not depend on the weight vector $\mathbf{w}$ it can be omitted. Taking the first two terms and rewriting in matrix notation we obtain

$$\widehat{R}(\mathbf{w}) = \mathbf{w}'\mathbf{J}\mathbf{w} - 2\mathbf{w}'\widehat{\mathbf{K}} \tag{5}$$

where $\mathbf{J}$ is the $M \times M$ matrix

$$\begin{aligned}
\mathbf{J} &= n\left(\overline{\beta} - \widehat{\beta}\right)' \widehat{W}_\beta \left(\overline{\beta} - \widehat{\beta}\right) \\
\overline{\beta} &= \left[\widehat{\beta}(1), ..., \widehat{\beta}(M)\right] \\
\widehat{\mathbf{K}} &= \left[\widehat{K}(1), ..., \widehat{K}(M)\right]'
\end{aligned}$$

Thus $\widehat{R}(\mathbf{w})$ is a simple quadratic in the weight vector $\mathbf{w}$.

The weight vector which minimizes $\widehat{R}(\mathbf{w})$ solves the minimization problem

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \widehat{R}(\mathbf{w}) \tag{6}$$

subject to $w(r) \geq 0$ and $\sum_{r=1}^{M} w(r) = 1$. This is a quadratic programming problem, for which algorithms are widely available. For example, in Matlab the command is "quadprog". Given $\widehat{\mathbf{w}}$ the feasible combination estimator of β is

$$\widehat{\beta}^* = \widehat{\beta}(\widehat{\mathbf{w}}) = \overline{\beta}\widehat{\mathbf{w}}.$$

This estimator is completely feasible and does not depend on tuning parameters.

We call $\widehat{\beta}^*$ the Stein combination estimator as it is a multi-model generalization of the classic Stein estimator.

8 VAR Point Forecasts

We now describe the precise form of the combination estimator for multi-step point forecasting from the VAR.

It is convenient to write the model in the first-order Markov form

$$x_t = Px_{t-1} + u_t \tag{7}$$

where P is the $k \times k$ matrix

$$P = \begin{bmatrix} B \\ J \end{bmatrix}, \qquad J = \begin{bmatrix} I_{m(p-1)} & 0_{m(p-1)\times m} & 0_{m(p-1)\times 1} \\ 0_{1\times m(p-1)} & 0_{1\times m} & 1 \end{bmatrix}$$

and

$$u_t = \begin{bmatrix} e_t \\ 0 \end{bmatrix}.$$

Iterating the Markov equation h times after the final observation n we obtain

$$x_{n+h} = P^h x_n + P^{h-1} u_{n+1} + \cdots + u_{n+h}.$$

Since the shocks $e_{n+\ell}$ for $\ell \geq 1$ are unforecastable at time n, the optimal point forecast for the j^{th} variable $y_{j,n+h}$ is the j^{th} element of $P^h x_n$, or $\beta_1' x_n$ where

$$\beta_1 = g_1(\theta) = \left(P^h\right)' S_j \tag{8}$$

and S_j is a $k \times 1$ selector vector with a 1 in the j^{th} place (and the remainder 0). Thus multi-step forecasting requires an estimate of the k-vector $\beta_1 = g_1(\theta)$ as defined in (8). This coefficient β_1 varies across forecast variable j and horizon h.

Theorem 5. *For $g_1(\theta)$ defined by (8),*

$$G_1 = \frac{\partial}{\partial \theta'} g_1(\theta) = \sum_{\ell=1}^{h} \left(\overline{S}_{k,m}' \left(P^{h-\ell}\right)' S_j \otimes P^{\ell-1} \right)$$

where we have used the shorthand $P^0 = I_k$ and

$$\overline{S}_{k,m} = \begin{pmatrix} I_m \\ 0_{(k-m)\times m} \end{pmatrix}.$$

The matrix G_1 is specific to the forecast horizon h and forecast variable j.

The least-squares estimates of P, β_1 and G_1 are

$$\begin{aligned} \widehat{P} &= \begin{bmatrix} \widehat{B} \\ J \end{bmatrix} \\ \widehat{\beta}_1 &= \left(\widehat{P}^h\right)' S_j \\ \widehat{G}_1 &= \sum_{\ell=1}^{h} \left(\overline{S}_{k,m}' \left(\widehat{P}^{h-\ell}\right)' S_j \otimes \widehat{P}^{\ell-1} \right), \end{aligned}$$

and the least-squares forecast for $y_{j,n+h}$ is $\widehat{y}_{j,n+h} = \widehat{\beta}_1' x_n$. This is equivalent to the point forecast $\widehat{y}_{j,n+h}$ found by iterating the estimated VAR.

The sub-model estimates are

$$\widehat{P}(r) = \begin{bmatrix} \widehat{B}(r) \\ J \end{bmatrix}$$
$$\widehat{\beta}_1(r) = \left(\widehat{P}(r)^h\right)' S_j$$

and the sub-model forecast for $y_{j,n+h}$ is $\widehat{y}_{j,n+h}(r) = \widehat{\beta}_1(r)' x_n$.

The mean-squared forecast error from any estimator $\widetilde{\beta}_1$ is

$$\begin{aligned} MSFE_n(\mathbf{w}) &= E\left(y_{j,n+h} - \widetilde{\beta}_1' x_n\right)^2 \\ &= E\left(\left(y_{j,n+h} - x_n'\beta_1\right) - x_n'\left(\widetilde{\beta}_1 - \beta_1\right)\right)^2 \\ &= E\left(x_n'\left(\widetilde{\beta}_1 - \beta_1\right)\right)^2 + E\left(y_{j,n+h} - x_n'\beta_1\right)^2 \\ &\simeq E\left(\left(\widetilde{\beta}_1 - \beta_1\right)' Q \left(\widetilde{\beta}_1 - \beta_1\right)\right) + E\left(y_{j,n+h} - x_n'\beta_1\right)^2. \end{aligned}$$

The first component on the right-hand-side is the weighted MSE of the estimate $\widetilde{\beta}_1$ with the weight matrix $W_\beta = Q$. The second component on the right-hand-side is the variance of the infeasible optimal error which is independent of the estimator $\widetilde{\beta}_1$. It follows that the estimator which minimizes the MSFE is the estimator which minimizes the weighted estimation MSE with the weight matrix $W_\beta = Q$. Hence for point forecasting we recommend setting $W_\beta = Q$ and $\widehat{W}_\beta = \widehat{Q}$.

In summary, for h-step point forecasting of the j^{th} variable $y_{j,n+h}$, the recommended combination forecast is

$$\widehat{y}^*_{j,n+h} = \sum_{r=1}^{M} \widehat{w}_1(r)\widehat{y}_{j,n+h}(r)$$

where

$$\begin{aligned} \widehat{\mathbf{w}}_1 &= \underset{\mathbf{w}}{\operatorname{argmin}}\, \widehat{R}_1(\mathbf{w}) \\ \widehat{R}_1(\mathbf{w}) &= \mathbf{w}'\mathbf{J}_1\mathbf{w} - 2\mathbf{w}'\widehat{\mathbf{K}}_1 \\ \mathbf{J}_1 &= n\left(\overline{\beta}_1 - \widehat{\beta}_1\right)' \widehat{Q}\left(\overline{\beta}_1 - \widehat{\beta}_1\right) \\ \overline{\beta}_1 &= \left[\widehat{\beta}_1(1), ..., \widehat{\beta}_1(M)\right] \\ \widehat{\mathbf{K}}_1 &= \left[\widehat{K}_1(1), ..., \widehat{K}_1(M)\right]' \\ \widehat{K}_1(r) &= \operatorname{tr}\left(\widehat{Q}\widehat{G}_1'\widehat{W}_\theta^{-1}R(r)\left(R(r)'\widehat{W}_\theta^{-1}R(r)\right)^{-1}R(r)'\widehat{V}\widehat{G}_1\right) \end{aligned}$$

Our forecast combination weights are specific for the forecast horizon h and forecast variable j, so should be calculated separately for each h and j.

Alternatively the weights could be aggregated across forecast variables and/or forecast horizons. This would be achieved by stacking the desired forecast coefficients β_1. Our recommendation to calculate the weights specific to the forecast variable and forecast horizon is because we expect that the best-fitting model is likely to vary across forecast variable and (especially) forecast horizon.

9 Impulse Response Functions

For impulse response analysis it is typical to decompose the equation error from model (1) into structural shocks as $e_t = H\varepsilon_t$ where H is $m \times m$ and identified and $E\varepsilon_t\varepsilon_t' = I_m$. In our application we will focus on the recursive case where H is lower triangular but other identifying structures can be used.

The impulse responses can be calculated from model (1) setting $B_0 = 0$. It is convenient to write the model in the Markov format (7) with the matrix P replaced by

$$P_0 = \begin{bmatrix} B_1 & \cdots & B_p & 0 \\ & & J & \end{bmatrix}. \tag{9}$$

Iterating the equation $h+1$ steps we find

$$x_{n+h} = P_0^{h+1}x_{n-1} + P_0^h \begin{pmatrix} H\varepsilon_n \\ 0 \end{pmatrix} + P_0^{h-1} \begin{pmatrix} H\varepsilon_{n+1} \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} H\varepsilon_{n+h} \\ 0 \end{pmatrix}.$$

The h-step impulse response of y_t with respect to the shock vector ε_t is the $m \times m$ derivative

$$\Gamma = \frac{\partial}{\partial \varepsilon_n'} y_{n+h} = \overline{S}_{k,m}' P_0^h \overline{S}_{k,m} H.$$

The $j^{th} - i^{th}$ element Γ_{ji} is the impulse response of the the j^{th} variable $y_{j,n+h}$ with respect to the i^{th} shock $\varepsilon_{i,n}$. We can write this as an $m^2 \times 1$ coefficient vector as

$$\beta_2 = g_2(\theta) = \text{vec}(\Gamma'), \tag{10}$$

This impulse response coefficient varies across horizon h.

Theorem 6. *For $g_2(\theta)$ defined by (10),*

$$G_2 = \frac{\partial}{\partial \theta'} g_2(\theta) = \sum_{\ell=1}^{h} \left(\overline{S}_{k,m}' \left(P_0^{h-\ell} \right)' \overline{S}_{k,m} \otimes P_0^{\ell-1} \overline{S}_{k,m} H \right).$$

The least-squares estimates of P_0, Γ, β_2 , and G_2 are

$$\begin{aligned}
\widehat{P}_0 &= \begin{bmatrix} \widehat{B}_1 & \cdots & \widehat{B}_p & 0 \\ & & J & \end{bmatrix} \\
\widehat{\Gamma} &= \overline{S}_{k,m}' \widehat{P}_0^h \overline{S}_{k,m} \widehat{H} \\
\widehat{\beta}_2 &= \operatorname{vec}\left(\widehat{H}' \overline{S}_{k,m}' \left(\widehat{P}_0^h \right)' \overline{S}_{k,m} \right) \\
\widehat{G}_2 &= \sum_{\ell=1}^{h} \left(\overline{S}_{k,m}' \left(\widehat{P}_0^{h-\ell} \right)' \overline{S}_{k,m} \otimes \widehat{P}_0^{\ell-1} \overline{S}_{k,m} \widehat{H} \right),
\end{aligned}$$

where $\widehat{\Sigma} = \widehat{H}\widehat{H}'$ is the structural decomposition of $\widehat{\Sigma}$.

The sub-model estimates are

$$\begin{aligned}
\widehat{P}_0(r) &= \begin{bmatrix} \widehat{B}_1(r) & \cdots & \widehat{B}_p(r) & 0 \\ & & J & \end{bmatrix} \\
\widehat{\Gamma}(r) &= \overline{S}_{k,m}' \widehat{P}_0(r) \overline{S}_{k,m} \widehat{H}(r) \\
\widehat{\beta}_2(r) &= \operatorname{vec}\left(\widehat{H}(r)' \overline{S}_{k,m}' \left(\widehat{P}_0(r)^h \right)' \overline{S}_{k,m} \right).
\end{aligned}$$

where $\widehat{\Sigma}(r) = \widehat{H}(r)\widehat{H}(r)'$ is the structural decomposition of $\widehat{\Sigma}(r)$.

For the weighted mean-squared error, we recommend setting $\widehat{W}_\beta = \left(\widehat{G}_2' \widehat{V} \widehat{G}_2 \right)^{-1}$ as a normalization.

Our theory for weight selection is developed for the case where the parameter of interest is a function of the regression coefficients $\widehat{\theta}$. Impulse responses are a bit more complicated in that they are also functions of the shock matrix $\widehat{H}$. Thus our theory technically applies only in the case where H is known and not estimated. For now, we effectively ignore the estimation error in $\widehat{H}$.

In summary, for h-step impulse response analysis the recommended combination estimate is

$$\widehat{\Gamma}^* = \sum_{r=1}^{M} \widehat{w}_2(r) \widehat{\Gamma}(r)$$

where

$$\begin{aligned}
\widehat{\mathbf{w}}_2 &= \underset{\mathbf{w}}{\operatorname{argmin}}\, \widehat{R}_2(\mathbf{w}) \\
\widehat{R}_2(\mathbf{w}) &= \mathbf{w}'\mathbf{J_2}\mathbf{w} - 2\mathbf{w}'\widehat{\mathbf{K}}_2 \\
\mathbf{J}_2 &= n\left(\overline{\beta}_2 - \widehat{\beta}_2\right)' \left(\widehat{G}_2'\widehat{V}\widehat{G}_2\right)^{-1} \left(\overline{\beta}_2 - \widehat{\beta}_2\right) \\
\overline{\beta}_2 &= \left[\widehat{\beta}_2(1), ..., \widehat{\beta}_2(M)\right] \\
\widehat{\mathbf{K}}_2 &= \left[\widehat{K}_2(1), ..., \widehat{K}_2(M)\right]' \\
\widehat{K}_2(r) &= \operatorname{tr}\left(\left(\widehat{G}_2'\widehat{V}\widehat{G}_2\right)^{-1} \widehat{G}_2'\widehat{W}_\theta^{-1}R(r)\left(R(r)'\widehat{W}_\theta^{-1}R(r)\right)^{-1} R(r)'\widehat{V}\widehat{G}_2\right)
\end{aligned}$$

These weights are specific for the horizon h, so should be calculated separately for each h.

Alternatively the weights could be aggregated across horizons, or calculated separately by row or column of Γ. We do not recommend aggregating across horizons as we expect the best-fitting models to vary considerably across horizons. We also do not recommend calculating the weights separately by row or column of Γ unless m is large.

10 Simulation

10.1 Impulse Response Analysis

We carefully explore the performance of our proposed impulse response estimator using simulation designs motivated by typical applied VARs. We take the VAR model (1) with $m = 7$ variables, $p = 5$ lags, and $n = 200$ observations. We compare impulse response estimates at horizons $h = 1, 4, 8, 12, 16$, and 20. We compare three methods:

1. OLS
2. Default BVAR posterior mode from Giannone, Lenza and Primiceri (2015) (we use their provided MATLAB code).
3. Stein combination of models AR(1) through AR(p) and VAR(1) through VAR(p)

We focus on the BVAR method from Giannone, Lenza and Primiceri (2015) as this is the state-of-the-art in the published BVAR literature, they documented impressive performance in their forecasting experiments, and they have tested MATLAB code for implementation.

We consider three simulation designs. The first generates each variable as an independent AR(1), thus $B_1 = I_m a$ and $B_2 = B_3 = B_4 = B_5 = 0$. The error e_t is generated as iid $N(0, \sigma^2 I_m)$ with $\sigma = 0.027$ to match the innovation variance of $4\ln(GDP)$ (to be consistent with the prior of GLP.) We vary a among [0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.98]. The results are calculated by simulation using 10,000 replications.

We compare the methods at the forecast horizons $h = 1, 4, 8, 12, 16$, and 20 by unweighted MSE, defined as

$$MSE(h) = \sum_{i=1}^{m} \sum_{j=1}^{m} E\left(\left(\widehat{\Gamma}_{ji} - \Gamma_{ji}\right)^2\right)$$

where Γ_{ji} is the true impulse response and $\widehat{\Gamma}_{ji}$ is an estimate. We report the results as the ratios of the square root of the MSE of each method relative to that of OLS. Thus root MSE ratios less than one indicate better performance than OLS, and root MSE ratios over one indicate worse performance than OLS. The results are presented in Table 1

Table 1: Impulse Response Estimates, Root MSE Relative to OLS, Design 1

a		$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$
0.5	BVAR	0.71	0.57	0.42	0.27	0.20	0.16
	Stein	0.56	0.38	0.05	0.01	0.00	0.00
0.6	BVAR	0.67	0.72	0.83	1.37	2.54	4.64
	Stein	0.58	0.43	0.12	0.05	0.02	0.01
0.7	BVAR	1.31	2.41	4.10	7.67	13.0	21.7
	Stein	0.60	0.49	0.26	0.12	0.07	0.05
0.8	BVAR	0.96	1.75	2.82	4.56	7.00	10.3
	Stein	0.63	0.56	0.46	0.26	0.19	0.16
0.9	BVAR	0.51	0.62	0.91	1.21	1.59	2.02
	Stein	0.66	0.64	0.66	0.62	0.58	0.57
0.95	BVAR	0.48	0.44	0.61	0.80	1.02	1.26
	Stein	0.68	0.69	0.77	0.79	0.80	0.81
0.98	BVAR	0.45	0.27	0.28	0.33	0.39	0.47
	Stein	0.70	0.74	0.84	0.87	0.89	0.91

There are several striking features in Table 1. First, the Stein method uniformly has lower root MSE than OLS. In many cells the root MSE ratio is less than one-half, and is as low as 0.00. Second, the root MSE of the Stein method relative to OLS decreases with the horizon h for low a, but increases with h when the persistence parameter a is close to one. Also the root MSE of Stein relative to OLS increases with a, indicating that the estimation problem is more difficult. Third, the root MSE of of the Stein method is typically much smaller than the BVAR method, except when the persistence parameter a is close to one. Except for $a = 0.98$, the Stein method has uniformly smaller root MSE than BVAR for $h \geq 12$. Fourth, the BVAR method in many cases has higher root MSE than OLS. It has lower root MSE for lower h and high a but can have much higher root MSE in other cases. Fifth, in some cases the BVAR method has incredibly high root MSE. The worst case is $a = 0.7$ and $h = 20$, where the root MSE of BVAR is 22 times that of OLS (yet the Stein method has a relative root MSE of 0.05). This occurs because in this setting, the BVAR algorithm has high probability (about one-third) of placing its posterior mode on the pure random walk model. Apparantly, the process looks close enough to a random walk that the BVAR

shrinks it all the way to this model. While this may be acceptable for one-step horizons (for which the likelihood is calibrated) it produces very poor multi-step impulse response estimates (1 versus $.7^h$ for the own responses).

In recent work, Giannone, Lenza and Primiceri (2016) have introduced cointegration priors into their BVAR method which improves performance at long horizons. It is possible, though not likely, that cointegration priors will rectify the problems with the BVAR method revealed in Table 1.

Our second design is the following VAR(5):

$$(I_m - aL)(I_m - bL)(I_m - cDL)\left(I_m + dL^2\right) y_t = e_t$$

where L is the lag operator and D is an $m \times m$ matrix of ones. Equivalently

$$\begin{aligned}
y_t &= B_1 y_{t-1} + B_2 y_{t-2} + B_3 y_{t-3} + B_4 y_{t-4} + B_5 y_{t-5} + e_t \\
B_1 &= (a+b) I_m + cD \\
B_2 &= -(ab+d) I_m - (a+b) cD \\
B_3 &= (a+b) dI_m + (ab+d) cD \\
B_4 &= -abdI_m - (a+b) cdD \\
B_5 &= abcdD
\end{aligned}$$

We set $b = 0.3$, $c = 0.1$ and $d = 0.3$ and again vary a. (Some experimention showed that a is the key parameter which affects the results.) The results are presented in Table 2.

Table 2: Impulse Response Estimates, Root MSE Relative to OLS, Design 2

a		$h=1$	$h=4$	$h=8$	$h=12$	$h=16$	$h=20$
0.5	BVAR	0.99	0.69	0.58	0.50	0.46	0.44
	Stein	0.85	0.61	0.38	0.18	0.09	0.05
0.6	BVAR	1.04	0.74	0.65	0.60	0.58	0.62
	Stein	0.85	0.68	0.46	0.23	0.11	0.06
0.7	BVAR	1.07	0.78	0.76	0.81	0.94	1.19
	Stein	0.86	0.74	0.55	0.34	0.21	0.14
0.8	BVAR	1.07	0.81	0.85	0.99	1.21	1.53
	Stein	0.87	0.79	0.69	0.51	0.41	0.34
0.9	BVAR	1.01	0.77	0.79	0.92	1.15	1.55
	Stein	0.88	0.82	0.83	0.76	0.70	0.65
0.95	BVAR	1.03	0.76	0.66	0.65	0.72	0.83
	Stein	0.88	0.85	0.89	0.87	0.85	0.84
0.98	BVAR	1.05	0.76	0.62	0.53	0.50	0.50
	Stein	0.89	0.87	0.93	0.93	0.92	0.92

The results in Table 2 are qualitatively similar to those in Table 1, but with much less extremes. With the more complicated correlation patterns, the BVAR method is less likely to put the posterior mode on the pure random walk, and hence is better behaved. Still, the overall patterns show that major efficiency improvements are achieved by the Stein combination method, and the performance of the BVAR method is sensitive and dependent on the parameterization.

The third design is based on a common empirical VAR. The seven-variable system used in Giannone, Lenza and Primiceri (2015) is estimated by OLS as a VAR(5) on an updated sample (quarterly data, 1959-2015, discussed in Section 11). The OLS coefficients estimates are used to parameterize a VAR(5), with the errors drawn as i.i.d. $N(0, \Sigma)$ where Σ is set equal to the estimated error covariance matrix. In this design we report the MSE of the impulse response estimates for each variable separately. Thus we calculate

$$MSE(h, j) = \sum_{i=1}^{m} E\left(\left(\widehat{\Gamma}_{ji} - \Gamma_{ji}\right)^2\right)$$

for each horizon h and variable j. Again we report the ratios of the root MSE of the BVAR and Stein estimates relative to the OLS estimates. The results are reported in Table 3.

Table 3: Impulse Response Estimates, Root MSE Relative to OLS, Design 3

		$h=1$	$h=4$	$h=8$	$h=12$	$h=16$	$h=20$
Real GDP	BVAR	0.83	0.85	1.24	1.33	1.41	1.64
	Stein	0.95	0.94	0.96	0.97	0.96	0.97
GDP Deflator	BVAR	1.26	1.91	2.22	2.40	2.48	2.49
	Stein	0.97	0.97	0.97	0.97	0.96	0.97
Real Comsumption	BVAR	0.99	1.02	1.26	1.30	1.45	1.72
	Stein	0.94	0.94	0.97	0.97	0.96	0.97
Real Investment	BVAR	0.88	0.70	1.03	1.04	1.04	1.30
	Stein	0.95	0.95	0.97	0.98	0.96	0.93
Hours	BVAR	0.90	1.14	1.67	1.83	1.89	2.11
	Stein	0.97	0.95	0.95	0.97	0.96	0.93
Real Compensation	BVAR	1.08	1.23	1.84	2.35	2.66	2.80
	Stein	0.94	0.94	0.95	0.96	0.96	0.98
Fed Funds Rate	BVAR	1.00	1.41	2.01	12.44	2.75	2.96
	Stein	0.96	0.94	0.95	0.94	0.96	0.94

There are several notable features about the results in Table 3. First, once again the Stein method has uniformly lower root MSE than OLS, though the differences are small. This appears to be because the DGP is highly persistent, which is similar to the $a = 0.98$ case in the design 2. Second, the Stein method has uniformly lower root MSE than the BVAR method for $h \geq 8$, but the two methods are competitive for $h = 1$. Third, for large h, the BVAR method once again displays quite large root MSE.

10.2 Multi-Step Point Forecast

We compare the point forecast accuracy of our proposed Stein combination method using the same simulation design as for impulse response estimation. We express the results as the the ratio of the root mean-squared forecast error (MSFE) of the Stein and BVAR point forecasts relative to the OLS point forecasts. The results are calculated by simulation using 10,000 replications.

For simulation designs 1 and 2, we compare the methods at the forecast horizons $h = 1, 4, 8$, and 12 by unweighted MSFE, defined as

$$MSFE(h) = \sum_{j=1}^{m} E\left(y_{j,n+h} - \widehat{y}_{j,n+h}\right)^2$$

where $y_{j,n+h}$ is out-of-sample observation and $\widehat{y}_{j,n+h}$ is the point forecast.

The results for design 1 are presented in Table 4.

Table 4: Point Forecasts, Root MSFE Relative to OLS, Design 1

a		$h=1$	$h=4$	$h=8$	$h=12$
0.5	BVAR	0.93	0.95	0.99	1.00
	Stein	0.91	0.95	0.99	1.00
0.7	BVAR	0.96	1.08	1.20	1.25
	Stein	0.91	0.93	0.96	0.98
0.9	BVAR	0.91	0.92	0.96	1.00
	Stein	0.91	0.90	0.91	0.92
0.95	BVAR	0.90	0.89	0.90	0.93
	Stein	0.90	0.89	0.89	0.89
0.98	BVAR	0.89	0.85	0.83	0.84
	Stein	0.90	0.89	0.88	0.88

The results in Table 4 are similar to those in Table 1, but milder. First, the Stein method uniformly has lower root MFSE than OLS. The ratio of the root MSFE is about 0.90 in most cells, but close to 1 for smaller a and high h. The root MSFE of the Stein and BVAR methods is similar in many cells, though there are some settings ($a = 0.7$ most notably) where the BVAR method has quite high root MSFE. For $a = 0.98$, the BVAR method has the lowest root MSFE.

The results for design 2 are presented in Table 5. The results in Table 5 are qualitively similar to those in Table 4, but with less extremes. Both the BVAR and Stein combination methods have lower MSFE than OLS, and neither one dominates the other.

Table 5: Point Forecasts, Root MSFE Relative to OLS, Design 2

a		$h=1$	$h=4$	$h=8$	$h=12$
0.5	BVAR	0.96	0.95	0.98	0.99
	Stein	0.94	0.94	0.98	0.99
0.7	BVAR	0.97	0.95	0.97	0.99
	Stein	0.94	0.94	0.95	0.98
0.9	BVAR	0.97	0.93	0.94	0.98
	Stein	0.94	0.92	0.91	0.91
0.95	BVAR	0.97	0.90	0.89	0.91
	Stein	0.93	0.91	0.90	0.89
0.98	BVAR	0.97	0.88	0.85	0.84
	Stein	0.93	0.92	0.90	0.89

The results for design 3 are presented in Table 6. For this design we report the root MSFE separately for each variable.

Table 6: Point Forecasts, Root MSFE Relative to OLS, Design 3

		$h=1$	$h=4$	$h=8$	$h=12$
Real GDP	BVAR	0.98	1.00	0.99	0.98
	Stein	0.97	0.98	0.99	0.99
GDP Deflator	BVAR	1.03	1.15	1.22	1.25
	Stein	0.99	1.00	1.00	1.00
Real Comsumption	BVAR	0.98	0.99	0.99	1.00
	Stein	0.97	0.99	0.99	0.99
Real Investment	BVAR	0.97	0.98	0.98	0.97
	Stein	0.98	0.99	0.99	0.99
Hours	BVAR	0.99	1.04	1.07	1.05
	Stein	0.97	0.98	0.99	0.99
Real Compensation	BVAR	0.98	1.03	1.13	1.20
	Stein	0.97	0.98	0.98	0.99
Fed Funds Rate	BVAR	1.03	1.14	1.27	1.35
	Stein	0.97	0.97	0.98	0.98

There are several notable features about the results in Table 6. First, the Stein method has uniformly lower root MSFE than OLS, though the improvements are quite modest. Second, the Stein method has nearly uniformly smaller root MSFE than the BVAR. The only exception is for Real Invesment, though the differences are modest. Third, the BVAR method has higher MSFE than OLS in many cases, especially for longer horizons. The differences are particularly large for the GDP deflator and the Fed Funds rate.

10.3 Summary of Simulation Evidence

The simulation carefully explored the MSE of the impulse response estimates and MSFE of the point forecasts. The simulation show that the performance of the BVAR method is quite sensitive to the parameterization. For parameterizations close to a random walk it can perform quite well, but for other parameterizations it can perform particularly poorly, especially at longer horizons. This performance is consistent with excessive shrinkage and hard thresholding. It is not a recommended empirical procedure.

The simulations also show that the Stein combination method achieves much lower MSE and MSFE than OLS, uniformly in the parameterization, and in some cases dramatically so. The evidence shows that the method is much preferred to OLS.

11 Empirical Application

The methods are illustrated using a standard empirical VAR. We estimate the seven-variable (medium) model of Giannone, Lenza and Primiceri (2015) with updated data. The sample are U.S. macroeconomic variables, quarterly, 1959:1 to 2016:1, extracted from the FRED database. The variables are listed in Table 7, along with their FRED labels and data transformations.

Table 7

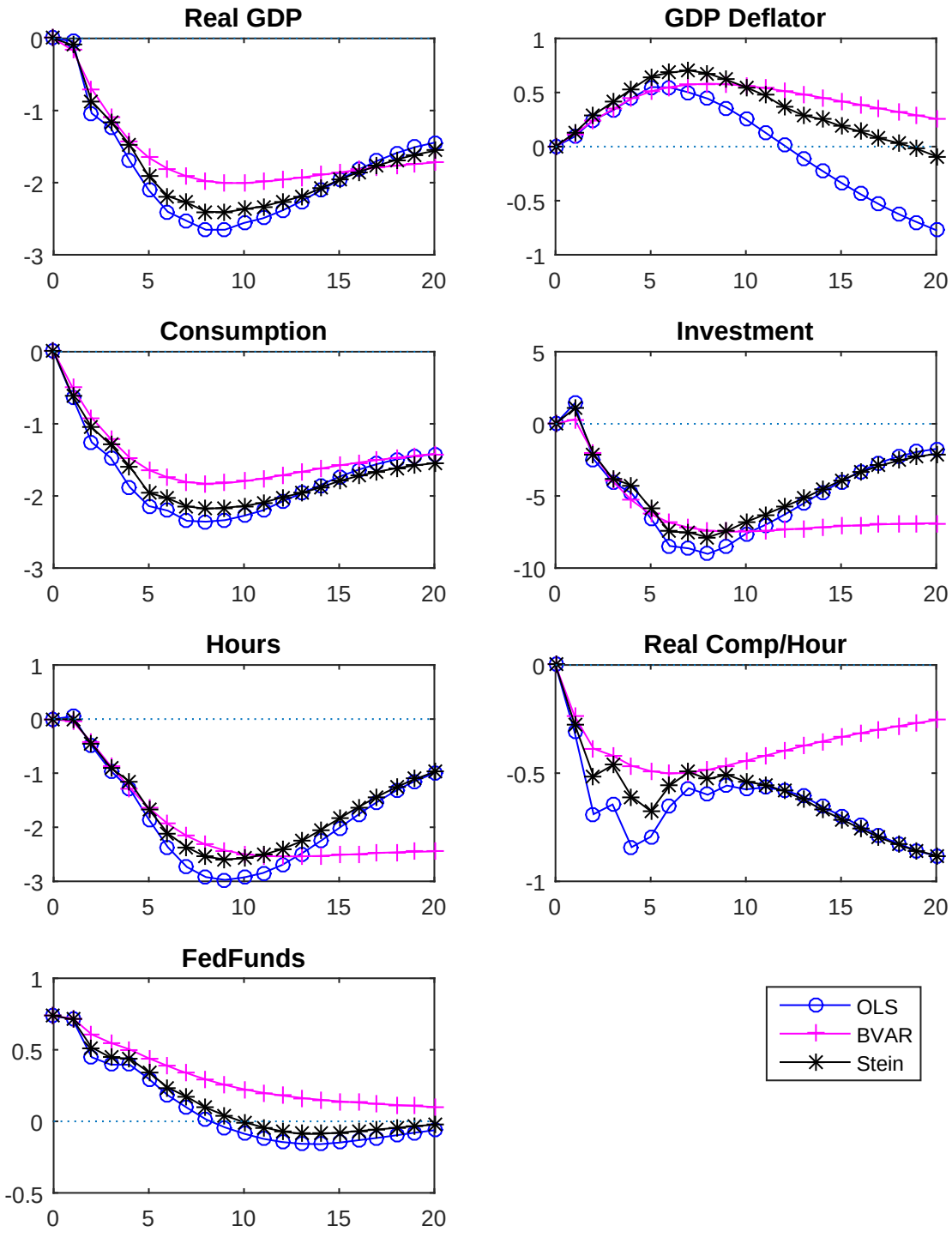
Description	FRED	Transformation
Real Gross Domestic Product	GDPC96	$4 \cdot \log$
GDP Implicit Price Deflator	GDPDEF	$4 \cdot \log$
Real Personal Consumption Expenditure	PCECC96	$4 \cdot \log$
Real Gross Private Domestic Investment	GPDIC1	$4 \cdot \log$
Hours Worked: Nonfarm Business Sector	HOANBS	$4 \cdot \log$
Real Compensation per Hour	COMPRNFB	$4 \cdot \log$
Federal Funds Rate	FF	$\div 100$

Following the prior literature, we use a VAR(5) model in the seven variables.

We first compare impulse response estimates of the variables in response to a monetary (fed funds) shock. The VAR shocks are identified recursively in the order listed in Table 7. We compare three methods: (1) OLS with 5 lags, (2) the default Bayesian posterior mode of Giannone, Lenza and Primiceri (2015), and (3) our Stein combination estimator. The estimates are displayed in Figure 1. Several features can be observed. First, the three estimation methods produce quite similar estimates for short horizons, but diverge at the longer horizons. Second, the impulse responses of some variables (Real GDP, Consumption) are similar across methods, but others (GDP Deflator, Investment, Hours, Real Compensation per Hour) are quite different.

Second, we compare the point forecasts generated by the methods. In Figure 2 we display point forecasts for the seven variables for 2016:2 through 2019:1 (1 through 12 quarters), comparing forecasts obtained by OLS with a VAR(5), the default BVAR posterior mode point forecast, and

Figure 1: Impulse Response Estimates Due to a Monetary (Fed Funds) Shock



the Stein combination point forecast. The forecasts for all variables except the Fed Funds rate are expressed as cumulative percentage changes relative to 2016:1. The forecasts for the Fed Funds rate is displayed as a percentage rate.

We can see from the Figure that the three forecasts are quite similar for the short horizons, but diverge at longer horizons. In general, the BVAR forecasts are the most optimistic (highest real growth and lowest inflation), the OLS forecasts most pessimistic (lowest real growth and highest inflation) and the Stein forecasts intermediate. The Stein forecasts for the next 12 quarters are for Real GDP to increase by about 5.4%, the GDP Deflator to increase by 5.6%, Real Consumption to increase by 5.8%, Real Investment to increase by 6.3%, Total Hours to increase by 1.2%, Real Compensation to increase by 1.4%, and the Fed Funds rate to increase to 1.2% .

To gain further insight, in Figures 3 and 4 we plot the combination weights for the impulse responses (Figure 3) and for the real GDP point forecasts (Figure 4) as a function of the horizon h. For simplicity we aggregate the five autoregressive models together. In FIgure 3 we can see that for impulse response estimation most weight (for most horizons about 0.7) is put on the full VAR(5) model. The VAR(3) receives the second most weight for the longer horizons. In Figure 5 we see that for forecasting real GDP there is a difference between short and long horizons. For short horizons the largest weight is put on the VAR(2) model. As the horizon increases the weight on VAR(2) falls, and the weights on VAR(3) and the autoregressive models increase, with each receiving about weight 0.5 for $h = 12$. The AR models which receive the weight are the AR(3) for short horizons and the AR(1) for the long horizons.

12 Conclusion

This paper has developed a new Stein combination criterion for Vector Autoregressions. The criterion is an estimate of the risk (expected squared error) of a vector-valued parameter of interest. While our criterion is appropriate for any nonlinear function of the coefficients, we pay particular attention to impulse response estimation and point forecasting. The criterion is quadratic, so weight selection is a simple quadratic programming problem. In three simulation experiments, we show that the method produces impressive reductions in MSE relative to OLS, and also show that BVAR estimates have erratic MSE. Finally, we illustrate the results on a standard macroeconomic VAR.

While the results are promising, more tasks need to be done. The current methods rely on estimation of a default or baseline model, which limits application to only medium-sized VARs, effectively excluding the "large VAR" models currently being explored. It would be useful to explore the performance in higher dimensional systems, and to develop alternative criteria which do not rely on a baseline model. The methods also only produce point forecasts: no theory of inference is currently available.

Figure 2: Quarterly Point Forecasts, 1 to 12 steps

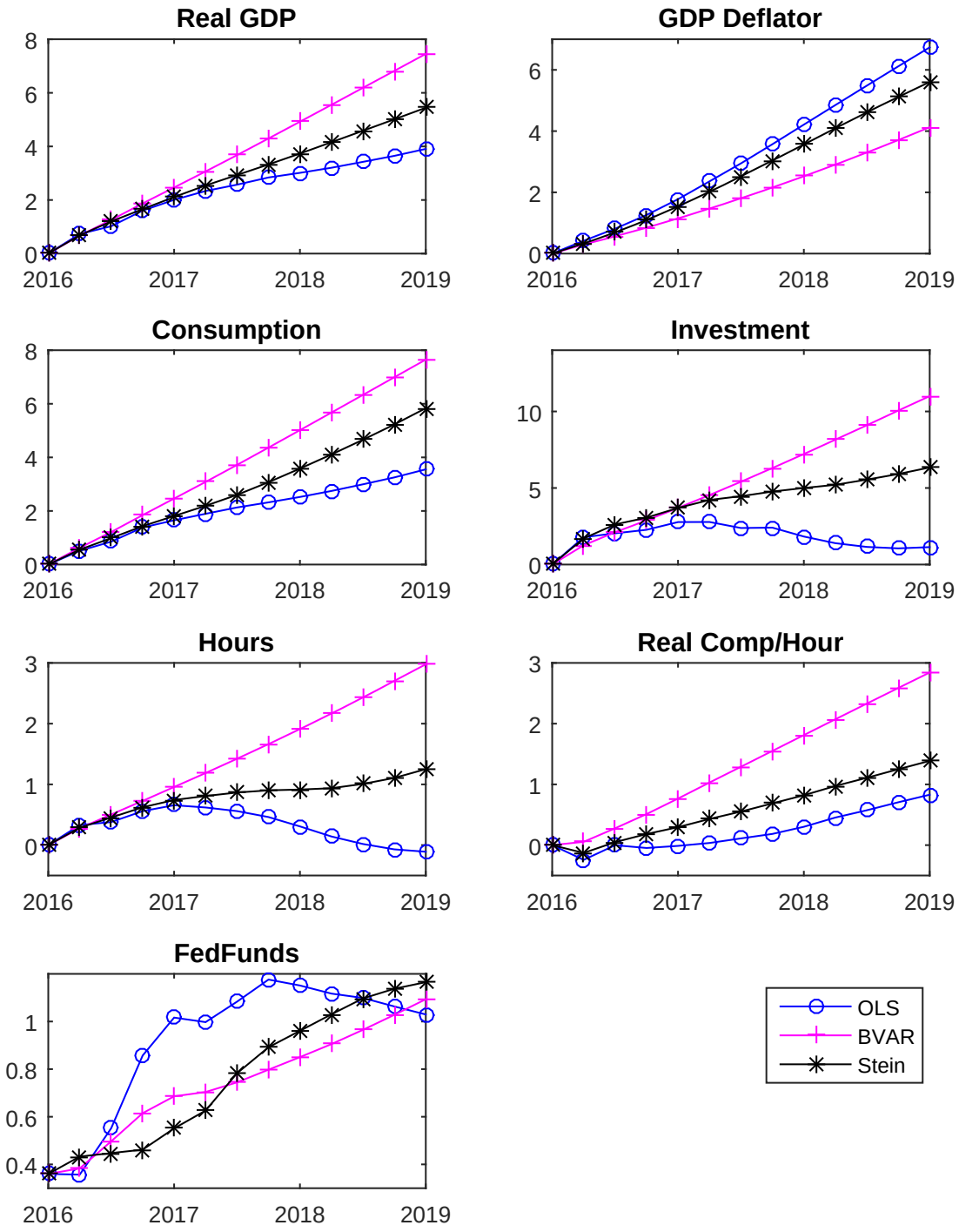


Figure 3: Impulse Response Combination Weights

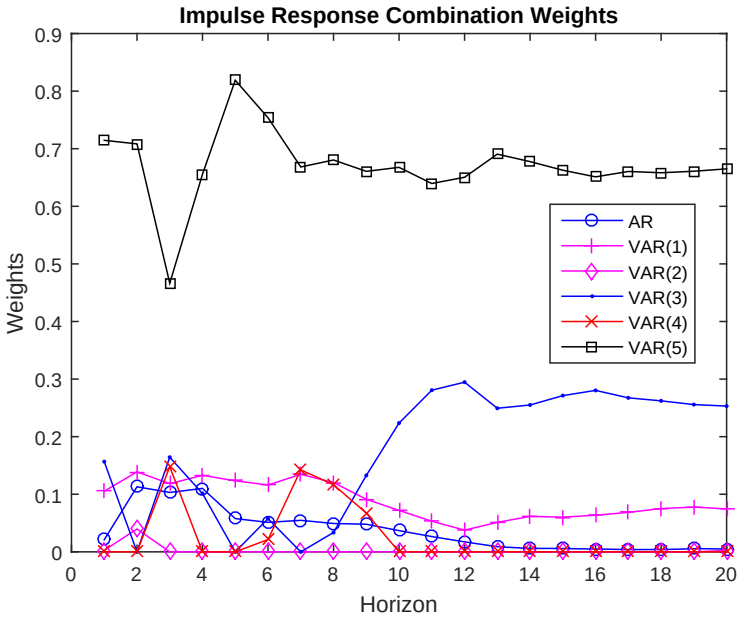
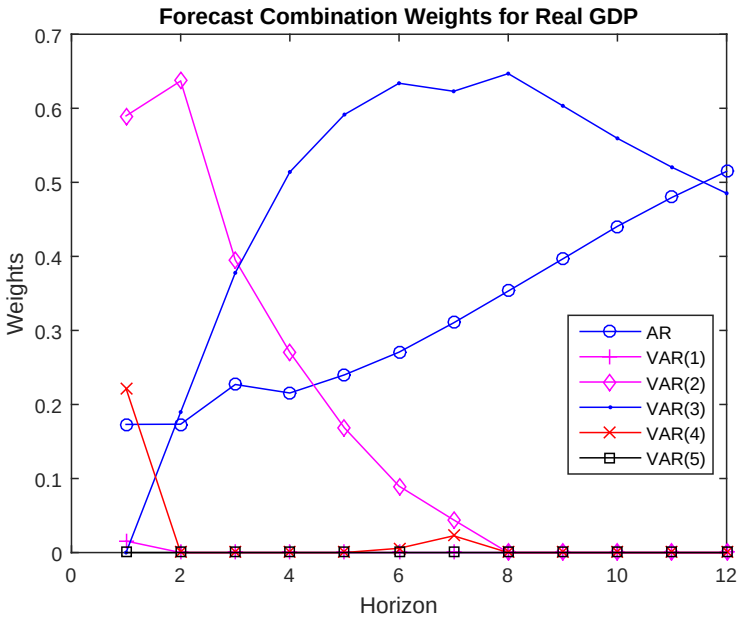


Figure 4: GDP Forecast Combination Weights



13 Mathematical Proofs

Proof of Theorem 1. Set $\eta_t = \operatorname{vec}(x_{t-1}e_t') = e_t \otimes x_{t-1}$. Observe that $E_{t-1}\eta_t = 0$ and $E\eta_t\eta_t' = E\left(e_te_t' \otimes x_{t-1}x_{t-1}'\right) = \Omega$. By the CLT for square integrable MDS

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\eta_t \to_d N(0,\Omega)$$

and the WLLN

$$\widehat{Q} = \frac{1}{n}\sum_{t=1}^{n}x_{t-1}x_{t-1} \to_p Q.$$

Using the fact $\operatorname{vec}(ABC) = (C' \otimes A)\operatorname{vec}(B)$, the above WLLN and CLT

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}-\theta\right) &= \operatorname{vec}\left(\left(\frac{1}{n}\sum_{t=1}^{n}x_{t-1}x_{t-1}\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}x_{t-1}e_t'\right)\right)\\ &= \left(I_m \otimes \widehat{Q}^{-1}\right)\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\eta_t\right)\\ &\to_d Z = \left(I_m \otimes Q^{-1}\right)N(0,Q) = N(0,V).\end{aligned}$$

Since $\beta = g(\theta)$ is a smooth function of θ, by the Delta method

$$\sqrt{n}\left(\widehat{\beta}-\beta\right) = G'\sqrt{n}\left(\widehat{\theta}-\theta\right) + o_p(1) \to_d G'Z.$$

By standard manipulations, it can also be shown that $\widehat{\Omega} \to_p \Omega$ and thus $\widehat{V} \to_p V$. ■

Proof of Theorem 2. Since $a(r) = R(r)'\theta - n^{-1/2}\delta(r)$,

$$\widehat{\theta}(r) = \widehat{\theta} - \widehat{W}_\theta^{-1}R(r)\left(R(r)'\widehat{W}_\theta^{-1}R(r)\right)^{-1}\left(R(r)'\left(\widehat{\theta}-\theta\right) + n^{-1/2}\delta(r)\right).$$

Note that

$$\widehat{W}_\theta^{-1} = \left(I_m \otimes \widehat{Q}^{-1}\right) \to_p \left(I_m \otimes Q^{-1}\right) = W_\theta.$$

Using Theorem 1

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}(r)-\theta\right) &= \left(I_k - \widehat{W}_\theta^{-1}R(r)\left(R(r)'\widehat{W}_\theta^{-1}R(r)\right)^{-1}R(r)'\right)\sqrt{n}\left(\widehat{\theta}-\theta\right)\\ &\quad + \widehat{W}_\theta^{-1}R(r)\left(R(r)'\widehat{W}_\theta^{-1}R(r)\right)^{-1}\delta(r)\\ &\to_d \left(I_k - W_\theta^{-1}R(r)\left(R(r)'W_\theta^{-1}R(r)\right)^{-1}R(r)'\right)Z\\ &\quad + W_\theta^{-1}R(r)\left(R(r)'W_\theta^{-1}R(r)\right)^{-1}\delta(r),\end{aligned}$$

and this convergence is uniform across r. By the delta method

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}(r)-\beta\right) &= G'\sqrt{n}\left(\widehat{\theta}(r)-\theta\right)+o_p(1)\\ &\to_d G'\left(I_k - W_\theta^{-1}R(r)\left(R(r)'W_\theta^{-1}R(r)\right)^{-1}R(r)'\right)Z\\ &\quad + G'W_\theta^{-1}R(r)\left(R(r)'W_\theta^{-1}R(r)\right)^{-1}\delta(r).\end{aligned}$$

The derivative matrix G is the same as in Theorem 1 because the function $\beta = g(\theta)$ is unchanged.

By linearity,

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}(\mathbf{w})-\beta\right) &= \sum_{r=1}^{M} w(r)\sqrt{n}\left(\widehat{\beta}(r)-\beta\right)\\ &\to_d \sum_{r=1}^{M} w(r)G'\left(I_k - W_\theta^{-1}R(r)\left(R(r)'W_\theta^{-1}R(r)\right)^{-1}R(r)'\right)Z\\ &\quad + \sum_{r=1}^{M} w(r)G'W_\theta^{-1}R(r)\left(R(r)'W_\theta^{-1}R(r)\right)^{-1}\delta(r)\\ &= G'\left[\left(I_m - D(\mathbf{w})\right)Z + \delta(\mathbf{w})\right]\end{aligned}$$

as stated. ■

Proof of Theorem 3. Theorem 2 implies

$$\begin{aligned}S_n(\mathbf{w}) &= n\left(\widehat{\beta}(\mathbf{w})-\beta\right)' W_\beta\left(\widehat{\beta}(\mathbf{w})-\beta\right)\\ &\to_d \left[Z'\left(I_m - D(\mathbf{w})'\right)+\delta(\mathbf{w})'\right]GW_\beta G'\left[\left(I_m - D(\mathbf{w})\right)Z+\delta(\mathbf{w})\right].\end{aligned}$$

By the standard properties of asymptotic trimmed moments

$$\begin{aligned}R(\mathbf{w}) &= \lim_{\zeta\to\infty}\liminf_{n\to\infty} E\min\left(S_n(\mathbf{w}),\zeta\right)\\ &= E\left(\left[Z'\left(I_m - D(\mathbf{w})'\right)+\delta(\mathbf{w})'\right]GW_\beta G'\left[\left(I_m - D(\mathbf{w})\right)Z+\delta(\mathbf{w})\right]\right)\\ &= \delta(\mathbf{w})'GW_\beta G'\delta(\mathbf{w}) + \operatorname{tr}\left(W_\beta G'\left(I_m - D(\mathbf{w})\right)V\left(I_m - D(\mathbf{w})'\right)G\right)\\ &= \delta(\mathbf{w})'GW_\beta G'\delta(\mathbf{w}) + \operatorname{tr}\left(W_\beta G'D(\mathbf{w})VD(\mathbf{w})'G\right) - 2\operatorname{tr}\left(W_\beta G'D(\mathbf{w})VG\right) + \operatorname{tr}\left(W_\beta G'VG\right)\\ &= \delta(\mathbf{w})'GW_\beta G'\delta(\mathbf{w}) + \operatorname{tr}\left(W_\beta G'D(\mathbf{w})VD(\mathbf{w})'G\right) - 2\sum_{r=1}^{M} w(r)K(r) + \operatorname{tr}\left(W_\beta G'VG\right)\end{aligned}$$

as stated. ■

Proof of Theorem 4. Theorem 2 implies

$$\sqrt{n}\left(\widehat{\beta}(\mathbf{w})-\widehat{\beta}\right)\to_d G'\left[\left(I_m - D(\mathbf{w})\right)Z+\delta(\mathbf{w})\right] - G'Z = -G'D(\mathbf{w})Z + G'\delta(\mathbf{w}).$$

Hence

$$\widehat{R}(\mathbf{w}) \to_d \widetilde{R}(\mathbf{w}) = (-D(\mathbf{w})Z + \delta(\mathbf{w}))' GW_\beta G' (-D(\mathbf{w})Z + \delta(\mathbf{w})) - 2\sum_{r=1}^{M} w(r)K(r) + \mathrm{tr}\left(W_\beta G' V G\right)$$

which has expectation

$$\begin{aligned} E\left(\widetilde{R}(\mathbf{w})\right) &= \delta(\mathbf{w})' G W_\beta G' \delta(\mathbf{w}) + \mathrm{tr}\left(W_\beta G' D(\mathbf{w}) V D(\mathbf{w})' G\right) - 2\sum_{r=1}^{M} w(r)K(r) + \mathrm{tr}\left(W_\beta G' V G\right) \\ &= R(\mathbf{w}). \end{aligned}$$

as stated. ■

For Theorems 5 and 6 the following result is convenient.

Lemma 1. *If B is $m \times L$, J is $(L-m) \times L$ and $P = \begin{bmatrix} B \\ J \end{bmatrix}$ is $L \times L$ then for any conformable matrices A and C*

$$\frac{\partial}{\partial \,\mathrm{vec}(B')} \left(\mathrm{vec}\left(A'\left(P^h\right)' C\right)\right)' = \sum_{\ell=1}^{h} \left(\overline{S}'_{L,m}\left(P^{h-\ell}\right)' C \otimes P^{\ell-1} A\right)$$

where we denote $P^0 = I_L$, and where

$$\overline{S}_{L,m} = \begin{pmatrix} I_m \\ 0_{(L-m)\times m} \end{pmatrix}.$$

Proof: For any conformable matrices D and F

$$\left(\mathrm{vec}\left(DP'F\right)\right)' = \left(\mathrm{vec}\left(P'\right)\right)' \left(F \otimes D'\right).$$

Hence

$$\frac{\partial \left(\mathrm{vec}\left(DP'F\right)\right)'}{\partial \,\mathrm{vec}\left(P'\right)} = \left(F \otimes D'\right).$$

Thus by the product rule

$$\begin{aligned} \frac{\partial \left(\mathrm{vec}\left(A'\left(P^h\right)' C\right)\right)'}{\partial \,\mathrm{vec}\left(P'\right)} &= \frac{\partial \left(\mathrm{vec}\left(A'P' \cdots P'C\right)\right)'}{\partial \,\mathrm{vec}\left(P'\right)} \\ &= \sum_{\ell=1}^{h} \left.\frac{\partial \left(\mathrm{vec}\left(A'\left(P^{\ell-1}\right)' X' \left(P^{h-\ell}\right)' C\right)\right)'}{\partial \,\mathrm{vec}\left(X'\right)}\right|_{X=P} \\ &= \sum_{\ell=1}^{h} \left(\left(P^{h-\ell}\right)' C \otimes P^{\ell-1} A\right). \end{aligned}$$

Since $P' = [B', J']$ then

$$\begin{aligned}\frac{\partial \left(\text{vec}\left(A'\left(P^h\right)'C\right)\right)'}{\partial \,\text{vec}\left(B'\right)} &= \left[\begin{array}{cc} I_{mL} & 0_{mL\times L(L-m)}\end{array}\right]\frac{\partial \left(\text{vec}\left(A'\left(P^h\right)'C\right)\right)'}{\partial \,\text{vec}\left(P'\right)} \\ &= \left(\overline{S}'_{L,m}\otimes I_L\right)\sum_{\ell=1}^{h}\left(\left(P^{h-\ell}\right)'C\otimes P^{\ell-1}A\right). \\ &= \sum_{\ell=1}^{h}\left(\overline{S}'_{L,m}\left(P^{h-\ell}\right)'C\otimes P^{\ell-1}A\right)\end{aligned}$$

as stated. ■

Proof of Theorem 5. $\beta_f = \left(P^h\right)'S_j = \text{vec}\left(\left(P^h\right)'S_j\right)$ takes the form of Lemma 1 with $L = k$, $A = I_k$ and $C = S_j$. ■

Proof of Theorem 6. $\beta_{ir} = \text{vec}(\Gamma') = \text{vec}\left(H'\overline{S}'_{k,m}\left(P_0^h\right)'\overline{S}_{k,m}\right)$ takes the form of Lemma 1 with $L = k$, $A = \overline{S}_{k,m}H$ and $C = \overline{S}_{k,m}$. ■

References

[1] Banbura, Marta, Domenico Giannone, and Lucrezia Reichlin (2010): "Large Bayesian VARs," *Journal of Applied Econometrics*, 25, 71-92.

[2] Carriero, Andrea, Ana Beatriz Galvao, and George Kapetanios (2015): "A comprehensive evaluatin of macroeconomic forecasting models," working paper.

[3] Cheng, Xu and Bruce E. Hansen (2015): "Forecasting with factor-augmented regression: A frequentist model averaging approach," *Journal of Econometrics*, 186, 280-293.

[4] Doan, Thomas, Robert Litterman, and Christopher A. Sims (1984): "Forecasting and conditional projection using realistic prior distributions," *Econometric Reviews*, 3, 1-100.

[5] Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2015): "Prior selection for vector autoregressions," *The Review of Economics and Statistics*, 97, 436-451.

[6] Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2016): "Prior for the long run," working paper.

[7] Hamilton, James D. (1994): *Time Series Analysis*. Princeton: Princeton University Press.

[8] Hansen, Bruce E. (2008): "Least squares forecast averaging," *Journal of Econometrics*, 146, 342-350.

[9] Hansen, Bruce E. (2014): "Model averaging, asymptotic risk, and regressor groups," *Quantitative Economics*, 5, 495-530.

[10] Hansen, Bruce E. (2015): "Efficient shrinkage in parametric models," *Journal of Econometrics*, 190, 115-132.

[11] Kapetanios, George, Massimiliano Marcellino, and Fabrizio Venditti (2016): "Large time-varying parameter VARs: A non-parametric approach," working paper.

[12] Koop, Gary, Dimitris Korobilis, and Davide Pettenuzzo (2016): "Bayesian compressed vector autoregression," working paper.

[13] Liao, Jen-Che and Wen-Jen Tsay (2016): "Multivariate least squares forecasting averaging by vector autoregressive models," working paper.

[14] Liu, Chu-An and Biing-Shen Kuo (2016): "Model averaging in predictive regressions," *Econometrics Journal*, forthcoming.

[15] Sims, Christopher A. (1980): "Macroeconomics and reality," *Econometrica*, 48, 1-48.

[16] Sims, Christopher A. and Tao Zha (1998): "Bayesian methods for dynamic multivariate models," *International Economic Review*, 39, 949-968.

[17] Stock, James H. and Mark W. Watson (2008): "Forecasting in dynamic factor models subject to structural instability," in Jennifer Castle and Neil Shephard, eds., *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*, New York: Oxford University Press.