Notes on Selected Papers in Time Series Econometrics

Mikkel Plagborg-Møller*
Harvard University
plagborg@fas.harvard.edu

March 1, 2013

DISCLAIMER:

These notes were written in preparation for a second-year PhD exam. They are only meant as rough summaries and can't substitute for actually reading the papers. I would be very happy to correct any errors and misinterpretations as well as entirely removing references to papers.

Contents

*I thank Jim Stock for helpful explanations and for devising the reading list. I am solely responsible for any errors.

1 Inference and model selection in linear time series models

1.1 Berk (1973): "Consistent Autoregressive Spectral Density Estimation"

Summary States assumptions for consistency and asymptotic normality of the autoregressive estimator of the spectral density at a finite number of frequencies.

Theory The univariate process in question is a causal AR(∞)

$$x_t = B(L)e_t \Rightarrow A(L)x_t = e_t,$$

with i.i.d. innovations e_t that have bounded fourth moments. Let $\hat{a}_{1k}, \dots, \hat{a}_{kk}$ be LS estimates from a fitted AR(k), RSS $\hat{\sigma}_k^2$. The autoregressive spectral density estimator is then

$$\hat{f}_k(\lambda) = \frac{\hat{\sigma}_k^2}{2\pi}|\hat{A}_k(e^{i\lambda})|^{-2}, \quad \hat{A}_k(z) = 1 + \hat{a}_{1k}z + \cdots + \hat{a}_{kk}z^k.$$

To get consistency, need $k^3/n \to 0$ and $k^{1/2}(|a_{k+1}| + |a_{k+2}| + \dots) \to 0$. Asymptotic normality obtains if in addition $k \to \infty$. The asymptotic variance is the same as that for truncated periodogram estimators, i.e., $(n/k)^{1/2}(\hat{f}_k(\lambda) - f(\lambda))$ has asymptotic variance $2f^2(\lambda)$ $(0 < \lambda < \pi)$ and is asymptotically independent of estimators at other frequencies.

2 Structural breaks

2.1 Andrews and Ploberger (1994): "Optimal Tests When a Nuisance Parameter Is Present Only Under the Alternative"

Summary Some tests have the feature that the likelihood depends on certain nuisance parameters π under the alternative but not the null. For example, for structural break tests, π is the timing of the break. Such tests are non-standard, so the classical asymptotic optimality results for Wald, LM and LR tests do not hold. The authors introduce a new class of test statistics that maximize weighted average (over values of π) power against local alternatives (to the parameter of interest). The special case of testing for structural breaks is treated in detail.

Theory The likelihood $f_T(\theta, \pi) \equiv f_T(Y_T; \theta, \pi)$ (wrt. the measure μ_T) is assumed well-specified. A parameter $\theta = (\beta', \delta')'$ is present under both the null and the alternative. $\beta \in \mathbb{R}^p$ is the parameter of interest, so the null hypothesis is $H_0 \colon \beta = 0$. The parameter vector π is only present under the alternative (i.e., the likelihood does not depend on π when $\beta = 0$). The new test is

$$\textit{Exp-LM}_T = (1+c)^{-p/2} \int \exp\left(\frac{1}{2}\frac{c}{1+c}LM_T(\pi)\right) dJ(\pi),$$

where $LM_T(\pi)$ is the standard LM test statistic (using the restricted ML estimator) given π and $J(\cdot)$ is a weight function (c.d.f.). Analogous tests are defined for the Wald and LR statistics. The local alternatives to H_0 are of the form $f_T(\theta_0 + B_T^{-1}h, \pi)$. The goal is to maximize weighted average power

$$\overline{\lim}_{T\to\infty} \int \left[\int \varphi_T f_T(\theta_0 + B_T^{-1}h, \pi) d\mu_T\right] dQ_\pi(h) dJ(\pi)$$

for tests φ_T of size α. By Fubini, this can also be seen as maximizing power against a single alternative density given by an integral. A particular form for Q_π is chosen, namely a singular multivariate normal distribution concentrated on the orthogonal complement to the linear space of parameters θ satisfying $\beta = 0$: $Q_\pi = N(0, c\Sigma_\pi)$. The variance is scaled by c. High c correspond to distant (local) alternatives.

Theorem 1 gives the asymptotic distribution of $\textit{Exp-LM}_T$. Theorem 2 states that it achieves the maximal weighted average power for the particular choice of Q_π.

Application The authors specialize to structural break tests by considering m-th order Markov data with an unknown break date. π is the timing of the break as a fraction of the sample size. In this case, the asymptotic distribution of the test statistics do not depend on δ_0, so it's straightforward to simulate critical values.

Proof intuition By Neymann-Pearson, the optimal test against the alternative density $\int f_T(\theta_0 + B_T^{-1}h, \pi)dQ_\pi(h)dJ(\pi)$ is given by the likelihood ratio test

$$LR_T = \frac{\int f_T(\theta_0 + B_T^{-1}h, \pi)dQ_\pi(h)dJ(\pi)}{f_T(\theta_0)}$$

with appropriate critical value. The proof proceeds by showing that

$$\textit{Exp-LM}_T - LR_T \xrightarrow{p} 0,$$

both under the null and under local alternatives. The exponential function appears due to normality of Q_π and a Taylor expansion of $\log[f_T(\theta_0 + B_T^{-1}h, \pi)/f_T(\theta_0)]$ in terms of the score.

2.2 Jushan Bai (REStat 1997): "Estimation of a Change Point in Multiple Regression Models"

Summary Considers the multivariate regression model with one break. The break point is estimated by LS. The estimator is shown to be consistent and asymptotic distribution theory is provided. To get an asymptotically pivotal distribution, the break size needs to tend to zero with the sample size.

Theory The model is

$$Y = X\beta + Z_0\delta + \epsilon,$$

where $Z_0 = (0, \ldots, 0, x_{k_0+1}, \ldots, x_T)'R$ for a fixed, known matrix R (i.e., allows for the researcher to know which subset of coefficient exhibits a break). The break date $k_0 = [\tau T]$ is estimated by, for each possible k, regressing Y on X and $Z_2(k) = (0, \ldots, 0, x_{k+1}, \ldots, x_T)'R$ and then choosing $\hat{k} = \arg\min_k S_T(k)$, where $S_T(k)$ is the SSR. It is shown that

$$\hat{k} = \arg\max_k W_T(k), \quad W_T(k) = \frac{\hat{\delta}_k'(Z_2 M Z_2)\hat{\delta}_k}{\hat{\sigma}_k^2}, \quad M = I - P_X,$$

i.e., the estimator can be thought of as first performing a sup-type test for structural breaks and then picking the break to be where the sup is attained.

Proposition 1 shows that under general conditions on the regressors (that may feature a time trend) and disturbances (which can be martingales or mixingales), the estimator is consistent. This is also the case if the parameter break δ tends to zero at a sub-$\sqrt{T}$ rate. Corollary 1 shows that β and δ can be estimated consistently by the corresponding LS estimators for $k = \hat{k}$.

If the break δ is fixed, Proposition 2 gives the asymptotic distribution, which is non-pivotal. To get a pivot, it is necessary to let $\delta_T \to 0$. Proposition 3 then gives conditions under which a proper scaling of $(\hat{k} - k_0)$ converges in distribution to a pivotal functional of Wiener processes. The limiting distribution is not symmetric in general. It is derived in closed form in Appendix B.

In the applications section, Bai discusses a procedure for detecting multiple breaks. First test for the presence of a break, then split the sample at the located break date and test parameter constancy on each subsample, etc. Any estimated change point should be reestimated if it is from a subsample with more than one break. Note that the sup-Wald test loses power with more than one break; intuitively, the estimated variance in the denominator is not consistent under breaks and it will tend to be too large when breaks actually occur.

2.3 Bai (EcmT 1997): "Estimating multiple breaks one at a time"

Summary Considers structural break tests in a linear regression model with multiple mean shifts. Gives consistency and asymptotic normality for a sequential testing procedure, for which a break point is first estimated by OLS, then the data is split in subsamples and additional breaks are estimated on the subsamples, etc. For each generated subsample, parameter constancy is tested. Simulation show that break points are estimated well when the number of breaks is known, and the sequential testing procedure is at least competitive when the number of breaks is unknown.

Theory Most of the paper focuses on the (almost WLOG) two-break model

$$Y_t = \mu_t + X_t,$$

where $\mu_t = \mu_1$ for $t \leq k_1^0$, $\mu_t = \mu_2$ for $k_1^0 + 1 \leq t \leq k_2^0$ and $\mu_t = \mu_3$ for $t \geq k_2^0 + 1$. Here $k_i^0 = [\tau_i^0 T]$. The break point is estimated by LS (Bai, REStat 1997)

$$\hat{k} = \arg\min_k S_T(k), \quad S_T(k) = \sum_{t=1}^{k} (Y_t - \bar{Y}_k)^2 + \sum_{k+1}^{T} (Y_t - \bar{Y}_k^*)^2.$$

Let $\hat{\tau} = \hat{k}/T$. Under some conditions, $\hat{\tau}$ is consistent at rate T: $\hat{\tau} - \tau_1^0 = O_p(T^{-1})$. This of course requires that the first break is more pronounced than the second in population. This rate of convergence is equal to the one obtained for simultaneous estimators (Bai and Perron, 1994) that search over all possible break dates. The T-rate is crucial for the subsequent results, since it implies that $\hat{k} - k_1^0 = O_p(1)$, so we get the order of the break point right. This allows us to consistently estimate k_2^0 (at rate T) by the LS estimator $\hat{k}_2$ restricted to the dates $[\hat{k}, T]$.

The asymptotic distribution of the initial estimator $\hat{k}$ is non-standard, depends on the distribution of X_t and is skewed. The skew results from the estimator being based on inconsistent estimates of the μ_i, as it imposes one break point when in fact there are two. The second break point estimator $\hat{k}_2$ does not suffer from this problem, so its limiting distribution is symmetric. Bai advocates "repartitioning," i.e., reestimating on the subsamples $[1, \hat{k}]$ (call the estimator $\hat{k}_1$) and $[\hat{k}_1, T]$. This yields symmetric asymptotic distributions for both. As in Bai (REStat 1997), letting

the regime means tend to zero at a rate strictly between 0 and $T^{-1/2}$ yields pivotal asymptotic distributions for the above estimators. The convergence rate for $\hat{\tau}_i$ is slower than T, though, due to the vanishing magnitude of the breaks.

With more than two breaks, an analogous sequential estimation procedure can be used. However, it is necessary to estimate the number of breaks. This may be done in a sequential fashion, where for each subsample the sup F-test is used to evaluate the reduction in the SSR from introducing the possibility of a break. The (pivotal) limiting distribution of the sup F-test is given. Bai shows that the sequential testing procedure consistently estimates the true number of breaks if the significance level is taken to vanish with the sample size at rate at most T^{-1}. Intuitively, so as to not underestimate the number of breaks, the critical values must not grow too large. So as to not overestimate, it is necessary that the critical values tend to infinity.

2.4 Elliott and Müller (2006): "Efficient Tests for General Persistent Time Variation in Regression Coefficients"

Summary It is argued that a time-varying parameter model with persistent variations in the parameter is a natural and useful way to model structural breaks. The authors show that for a large class of persistent probability laws for the time-varying parameter β_t under the alternative, the point-optimal test of the no-break hypothesis against said alternative is asymptotically equivalent to a test statistic $\widetilde{LR}_t$ which only depends on the long-run variance of $T\Delta\beta_t$. Consequently, *all* optimal tests tailored to some specific time-varying alternative are equivalent. An easy-to-compute, feasible test statistic is proposed based on $\widetilde{\mathrm{LR}}_t$. Simulations show that it does well in finite samples.

Motivation The authors argue that deterministic constant break models are very similar to time-varying parameter models. Optimal weighted average power tests for the former can be seen as providing a probability model for the break dates, which is tantamount to a time-varying break process. In most examples, whether there is *persistent* variation in β_t is really the hypothesis of interest, so tests should attempt to direct power against persistent local alternatives.

Theory The model is

$$y_t = X_t'\beta_t + Z_t'\delta + \varepsilon_t,$$

and under the null, $\beta_t \equiv \bar{\beta}$. Condition 1 outlines the assumptions on the class of alternatives. In particular, $\Delta\beta_t$ is on the scale T^{-1}, and $T\Delta\beta_t$ has a non-singular long-run covariance matrix Ω. Note that the (random) number of breaks must necessarily grow with the sample size, but otherwise the alternatives are very general. They allow for smooth adjustment to the new level after the break.

By Neymann-Pearson, a point-optimal test against any of the alternatives is given by the likelihood ratio statistic LR_t. Under the assumption that the errors are conditionally normal, Theorem 1 shows that this statistic is asymptotically equivalent (under the null and the alternative) to another (infeasible) statistic $\widetilde{\mathrm{LR}}_t$, which only depends on the probability law of β_t through Ω. This latter statistic can in turn be explicitly computed thanks to normality. It is equivalent to a test that has a quadratic form, for which a feasible version $\widehat{\mathrm{qLL}}$ can be defined (see p. 914–915). The feasible test only uses $k+1$ OLS regressions and is therefore computationally attractive relative to tests that require the researcher to search over all possible break locations. Critical values are obtained from functionals of Wiener processes.

Theorem 1 is proved by first showing the asymptotic equivalence result under the null. Then a contiguity argument is used (as in Andrews and Ploberger, 1994) to also obtain the result under the alternative.

One remaining issue is the dependence of the optimal test on Ω. If one had a particular choice of alternative in mind, a value of Ω would be implied, and so basing the test on a generic choice for Ω would result in a loss of power. However, the authors argue that if rotational invariance is desired, then $\sigma^{-2}\Sigma_X^{1/2}\Omega\Sigma_X^{1/2} =: \Omega^* = a^2 I$, so the choice of Ω reduces to the choice of a (akin to the choice of c in Andrews and Ploberger, 1994). Plots show that the local power envelopes for various choices of a are extremely close.

Theorem 4 shows that the $\widehat{\text{qLL}}$ test is valid under general conditions on regressors and errors (in particular, normality is not needed).

3 HAC

3.1 Sun, Phillips and Jin (2008): "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing"

Summary Andrews (1991) determined the optimal bandwidth M for HAC estimation based on an asymptotic MSE criterion. SPJ argue that in most cases, the long-run variance is not the main object of interest; instead, it is only an ingredient entering into hypothesis testing in regression models. The authors focus on a Gaussian location model and derive higher-order expansions of the non-standard KVB distribution of the t-statistic, as well as of the finite-sample distribution of the t-statistic (this latter derivation requires Gaussianity of the errors). Both expansions are developed for $b \to 0$, and the distributions are approximated both under the null and under local $(1/\sqrt{T})$ alternatives. The leading order terms of the approximated KVB limiting distribution and the finite-sample distribution coincide, which provides analytic support for the superiority of the KVB approach. As a bonus, the finite-sample expansion implies (using a Cornish-Fischer inversion of the Edgeworth expansion) an expression for higher-order correct critical values if one were to base inference on the normal distribution.

The expansions yield expressions for the type I and type II errors (the latter under local alternatives), under the assumption that the higher-order corrected critical value (or, equivalently to high order, the KVB limit theory) is used. The authors consider a loss function that is a convex combination of these two errors, with weight $w_T/(1 + w_T)$ on the former. Loosely, for time series with positive serial correlation, the type I error increases as the bandwidth decreases (when oversmoothing, fewer autocovariances receive weight in the estimator, so the LRV estimate is too small); the type II error generally decreases with the bandwidth. For most time series models, the optimal choice of M is $O(T^{1/(q+1)})$, where q is the Parzen characteristic exponent. This is larger than (i.e., undersmooths relative to) the MSE-minimizing choice $M = O(T^{1/(2q+1)})$. Hence, the optimal bandwidth allows for greater variance in order to reduce bias. If the weight w_T on the type I error diverges with T, the optimal choice of $b = M/T$ satisfies $b = O((w_T/T^q)^{1/(q+1)})$, so the fixed-$b$ rule of KVB can be interpreted as attaching a large weight $w_T = O(T^q)$ on the type I error. The optimal bandwidth choice achieves a strictly better combined type I and II loss, expressed as a rate in T.

Model Location model $y_t = \beta + u_t$. u_t has LRV

$$\omega^2 = \gamma_0 + 2\sum_{j=1}^{\infty} \gamma_j, \quad \gamma_j = E[u_t u_{t-j}].$$

HAC estimators of the form

$$\hat{\omega}_b^2 = \sum_{j=-T+1}^{T-1} k(j/bT)\hat{\gamma}_j$$

are considered. We want to test $H_0\colon \beta = \beta_0$ against the two-sided alternative. Focus is on the t-statistic $\sqrt{T}(\hat{\beta} - \beta)/\hat{\omega}_b$. An FCLT is assumed for the partial sum of u_t, so that the KVB limit theory kicks in when b is fixed.

As in Andrews (1991), the Parzen characteristic exponent is

$$q = \max\left\{q_0 \in \mathbb{Z}_+\colon \lim_{x\to 0} \frac{1-k(x)}{|x|^q} < \infty\right\}.$$

This is 1 for the Bartlett kernel and 2 for the QS and Parzen kernels.

3.2 Sun and Phillips (2009): "Optimal Bandwidth Choice for Interval Estimation in GMM Regression"

Summary The paper considers a linear GMM regression model $y_t = x_t'\beta_0 + u_t$, with instruments that satisfy $E[u_t z_t] = 0$. The two-step efficient GMM estimator of β_0 is employed, so the LRV of $v_t = z_t u_t$ is needed for inference. The authors derive Edgeworth expansions for the finite-sample distribution of the t-statistic that uses a non-parametric HAC estimate. These expansions lead to expressions for the coverage probability error (CPE) of one- and two-sided confidence intervals. The optimal bandwidth M and higher-order corrected critical values for minimizing the CPE are then derived (using asymptotic normality rather than KVB limit theory). It turns out that minimizing the CPE requires choosing M to balance the asymptotic bias and variance of the HACE. For two-sided CIs, the bias is $O(M^{-q})$ and the variance $O(M/T)$, so the optimal bandwidth is generally $O(T^{1/(q+1)})$. This contrasts with the asymptotic MSE criterion in Andrews (1991), in which the *square* of the bias is balanced with the variance, leading to $M = O(T^{1/(2q+1)})$. As a side result, the Sun and Phillips show that the QS kernel is not optimal for their purposes, as it doesn't minimize the CPE conditional on use of the optimal bandwidth (they don't determine which kernel is in fact optimal).

A data-driven algorithm for computing the optimal bandwidth is provided. The model is estimated in a first step, which gives estimates $\hat{v}_t$ that are then fitted to some low-order parametric model like a VAR(1). This model is used to compute approximation to the necessary inputs into the optimal M formula. The resulting automatic bandwidth is then used for estimation in the last step. Simulations show that the optimal bandwidth formula with corrected critical values clearly outperforms the Andrews (1991) approach. There is no clear ranking of the kernels.

Finally, the authors also consider a different optimality criterion, namely maximizing the power against local alternatives (minimizing the probability of false coverage), subject to the true CPE being below a threshold. This involves specifying a prior distribution over the set of local alternatives considered. The Edgeworth expansions imply formulas for the probabilities of true and false coverage, so the optimization in M is straight-forward (conceptually) for any given prior distribution.

Intuition Stock and Watson (2008 NBER mini course) give the following intuition for why the non-squared bias matters for the CPE. Let $z \sim N(0, \sigma^2)$ and let $\hat{\sigma}^2$ be an estimator that is independent of z. Then

$$\begin{aligned} P(z^2/\hat{\sigma}^2 < c) &= E[\mathbf{1}_{\{z^2 < c\hat{\sigma}^2\}}] = E[g(\hat{\sigma}^2)] \\ &\approx E[g(\sigma^2)] + E[\hat{\sigma}^2 - \sigma^2]E[g'(\sigma^2)] + \frac{1}{2}E[(\hat{\sigma}^2 - \sigma^2)^2]E[g''(\sigma^2)] \\ &= F_{\chi^2_1}(c) + \text{Bias}(\hat{\sigma}^2)E[g'(\sigma^2)] + \frac{1}{2}\text{MSE}(\hat{\sigma}^2)E[g''(\sigma^2)]. \end{aligned}$$

(To make the above somewhat meaningful, we probably need to let g be a smooth approximation to the indicator function.) We see that the expression above involves the un-squared bias, as well as the squared bias and the variance (from the MSE). So more weight is placed on bias than in an MSE criterion.

4 Weak identification

4.1 Stock and Wright (2000): "GMM with Weak Identification"

Summary Develop non-standard asymptotic approximation for IV with non-linear moment conditions when some of the parameters are weakly identified. The parameter vector is written as $\theta = (\alpha', \beta')'$, where α is weakly and β strongly identified. To formally define weak identification, the mean of the sample moment condition is split up into three parts, one evaluated at the true parameters (α_0, β_0), one that only depends on β and one that depends on both (α, β). The last component is modeled as a function (independent of the sample size) divided by $\sqrt{T}$, such that this component stays bounded as $T \to \infty$, leading to the weak identification of α. The limiting distributions for the GMM objective function and the traditional GMM estimator are given. Both $\hat{\alpha}$ and $\hat{\beta}$ have non-standard limiting distributions, as $\hat{\beta}$ is influenced by the inability to consistently estimate α even if β_0 were known, so the moments are not evaluated within a local neighborhood of α_0 in large samples. However, $\hat{\beta}$ remains $\sqrt{T}$-consistent. Due to the non-standard limit distributions, traditional LR and Wald statistics will not be valid. Instead, the authors show that the CUE objective function evaluated at the true θ_0 has a limiting χ^2 distribution, which suggests inverting this test statistic to obtain valid confidence regions. Furthermore, to reduce the d.f. by n_β, one can concentrate out the strongly identified parameters β. The resulting robust confidence regions are called S-sets. The authors caution that S-sets could be small either because (1) the model is misspecified and the data relatively uninformative or (2) the model is correct and the precisely estimated.

The authors specialize the results to the one-step, two-step and CUE GMM estimators. The asymptotics suggest a measure of identification that is analogous to the concentration parameter. If this measure is small, the two-step estimator will tend to be biased toward the NLS estimator, just as 2SLS is biased towards OLS in linear IV. It is shown that the non-linear set-up reduces to the Staiger and Stock (1997) asymptotics for linear IV under appropriate modeling choices. Finally, a simulation study and an empirical exercise based on the CCAPM model are performed.

Model The G conditional moment conditions are $E[h(Y_t, \theta_0)|F_t] = 0$, and the K-dimensional vector of instruments is Z_t, leading to the GK-dimensional moment vector $\phi_t(\theta) = h(Y_t, \theta) \otimes Z_t$.

Let $\tilde{m}_T(\alpha,\beta) = E[T^{-1}\sum_{t=1}^{T}\phi_t(\alpha,\beta)]$. This is split up into

$$\tilde{m}_T(\alpha,\beta) = \tilde{m}_T(\alpha_0,\beta_0) + \tilde{m}_{1T}(\alpha,\beta) + \tilde{m}_{2T}(\beta),$$

where $\tilde{m}_{1T}(\alpha,\beta) = \tilde{m}_T(\alpha,\beta) - \tilde{m}_T(\alpha_0,\beta)$ and $\tilde{m}_{2T}(\beta) = \tilde{m}_T(\alpha_0,\beta) - \tilde{m}_T(\alpha_0,\beta_0)$. Since β is considered to be strongly identified, the authors set $\tilde{m}_{2T}(\beta) = m_2(\beta)$ for a bounded function that satisfies the usual identification conditions. However, since α is supposed to be weakly identified, they model $\tilde{m}_{1T}(\alpha,\beta) = m_1(\alpha,\beta)/\sqrt{T}$, so that even asymptotically, the population objective function is finite globally in α.

Let $\mu(\alpha) = \Omega_{\alpha,\beta_0}^{-1/2} m_1(\alpha,\beta_0)$, where Ω_{α,β_0} is the asymptotic variance of the moment condition vector evaluated at (α,β_0). The above-mentioned analog to the concentration parameter is $\mu(\alpha)'\mu(\alpha)$. In general, the dependence of $\mu(\alpha)$ on α may be complicated, so unlike in the linear case, a full characterization of the extent of weak identification requires global knowledge of $\mu(\alpha)'\mu(\alpha)$.

4.2 Kleibergen and Mavroeidis (2011): "Inference on subsets of parameters in linear IV without assuming identification"

Summary The authors aim to show that subset versions of the standard weak IV robust tests are valid. The model is the usual linear homoskedastic IV regression model. The parameter vector is partitioned as $\theta = (\beta',\gamma')'$. To test the hypothesis $H_0\colon \beta = \beta_0$, the authors propose to plug the LIML estimator of γ under $\beta = \beta_0$ into the AR, KLM, JKLM or MQLR tests. They claim to show that these subset test statistics are stochastically dominated by their strong-instrument asymptotic distributions regardless of the size of the first-stage coefficient matrix.

Model Linear IV regression model

$$y = X\beta + W\gamma + \varepsilon,$$

$$X = Z\Pi_X + V_X, \quad W = Z\Pi_W + V_W.$$

The null hypothesis is $H_0\colon \beta = \beta_0$ with two-sided alternative. Let

$$\hat{\gamma}(\beta_0) = \arg\min_{\gamma} \frac{(y - X\beta_0 - W\gamma)'P_Z(y - X\beta_0 - W\gamma)}{(y - X\beta_0 - W\gamma)'M_Z(y - X\beta_0 - W\gamma)}$$

be the LIML estimator under $\beta = \beta_0$. Then the subset AR statistic is

$$AR(\beta_0) = \frac{(y - X\beta_0 - W\tilde{\gamma}(\beta_0))'P_Z(y - X\beta_0 - W\tilde{\gamma}(\beta_0))}{(y - X\beta_0 - W\tilde{\gamma}(\beta_0))'M_Z(y - X\beta_0 - W\tilde{\gamma}(\beta_0))}.$$

The subset Kleibergen LM statistic (KLM) is the same expression, except Z is replaced with $Z(\tilde{\Pi}_W(\beta_0),\tilde{\Pi}_X(\beta_0))$ in the projection matrices, where $\tilde{\Pi}_W(\beta_0)$, say, is the ML estimator of Π_W in the linear system under H_0. The subset JKLM statistic is the difference between the AR and KLM statistics. Finally, the MQLR statistic is a non-linear function of the AR and KLM statistics as well as a statistic that tests for reduced rank of Π_X and Π_W.

Results If Π_W is fixed and has full rank, such that the nuisance parameter γ is well identified, previous results in the literature showed that the various statistics have limiting χ^2 distributions.

For the full-parameter AR statistic, we have $AR(\beta_0, \gamma_0) = KLM_\gamma(\beta_0, \gamma_0) + JKLM_\gamma(\beta_0, \gamma_0)$, where KLM_γ is the KLM statistic that tests $\gamma = \gamma_0$ given $\beta = \beta_0$. The KLM_γ statistic is a quadratic form in the derivative of the full-parameter AR statistic wrt. γ, so it vanishes at $\tilde{\gamma}(\beta_0)$. The authors show that $AR(\beta_0, \gamma)$ and $JKLM_\gamma(\beta_0, \gamma)$ are both minimized at $\tilde{\gamma}(\beta_0)$, so

$$AR(\beta_0) = AR(\beta_0, \tilde{\gamma}(\beta_0)) = JKLM_\gamma(\beta_0, \tilde{\gamma}(\beta_0)) \leq JKLM_\gamma(\beta_0, \gamma_0).$$

It follows from Kleibergen (2002) that the RHS has a χ^2 limiting distribution, so the desired stochastic dominance result follows.

For the other statistics, the above proof strategy doesn't work. Instead the authors write those statistic in ways that resemble the AR statistic and then use some asymptotic independence arguments to purportedly prove the results.

4.3 Chen and Guggenberger (2011): "On the Asymptotic Size of Subvector Tests in the Linear Instrumental Variables Model"

Summary Consider the linear homoskedastic IV model with two endogenous variables and k_2 instruments. The null hypothesis concerns the coefficient of one of the two endogenous variables. The authors seek to establish the asymptotic size of the subvector AR and LM tests. This is done by using a result from Andrews, Chen and Guggenberger (2011, ACG) that relates the asymptotic size to the supremum of the null rejection probabilities for the limiting distributions under a set of drifting parameter sequences. In this case, the latter limiting distributions depend on a finite-dimensional parameter, whose dimension may be reduced to a point where it's feasible to numerically calculate the supremum using simulations. Extensive numerical calculations indicate that the subset AR statistic does indeed have correct size. However, contrary to the claim in Kleibergen and Mavroeidis (2011), the subset LM statistic is oversized for $k_2 > 3$, and by orders of magnitude for $k_2 \geq 10$.

Model and results The model is

$$y = Y\beta + X\zeta + u, \quad Y = Z\pi + X\phi + V,$$

where Z has column dimension k_2. The subvector AR and LM statistics are as in Kleibergen and Mavroeidis (2011). Let $\lambda \in \Lambda$ denote a parameterization of the model under the null $H_0 \colon \beta = \beta_0$ (i.e., the parameters excluding β_0). Let $RP_n(\lambda)$ denote the rejection probability of a certain test under the parameters λ and with sample size n. The *asymptotic size* is defined as

$$AsySz = \limsup_{n\to\infty} \sup_{\lambda\in\Lambda} RP_n(\lambda).$$

The authors specify a certain convenient parameterization of the model. Λ is defined so that weak identification is allowed (π and ϕ are left unrestricted) but homoskedasticity is imposed. Let h denote the limit of an (appropriately scaled) arbitrary drifting sequence of parameters λ_n, and let H be the set of all h that are limits of such drifting sequences. The authors show that the limiting distributions of the AR and LM statistics under the drifting parameter sequence λ_n depend only

on $h = \lim_{n\to\infty} \text{scaling}(n) \times \lambda_n$. ACG (2011) showed that the asymptotic size of the AR test can then be determined as

$$AsySz = \sup_{h\in H} P(AR_h > \chi^2_{k_2-1,1-\alpha}),$$

and similarly for the LM test (with 1 d.f.). It is further possible to reduce the dimensionality of h by establishing that the limiting distributions of the two statistics only depend on a dimension-reducing transformation of h.

The asymptotic size can now be calculated using numerical methods by simulating, over a grid of h values, the rejection probability of the AR_h and LM_h tests (because not all h values can be checked, the result is a lower bound on the asymptotic size). As a bonus, size-corrected critical values for the LM statistic can be obtained by figuring out numerically which critical value would render the above supremum equal to the nominal size.

5 Modeling of and inference for persistent time series

5.1 Mikusheva (2007): "Uniform Inference in Autoregressive Models"

Summary Because the behavior of autoregressive processes changes dramatically in the neighborhood of a unit root, it is argued that uniform inference is desirable, i.e., tests that control size uniformly in $\rho \in [0, 1]$ as $T \to \infty$. Focusing first on AR(1) processes, the author considers tests of $H_0\colon \rho = \rho_0$ that depend on two particular scalar statistics. Confidence sets are obtained by inverting a test statistic. Tests in this class include (1) Andrews' (1993) finite sample parametric simulation-based approach, (2) Stock's (1991) test that bases the critical value on local-to-unity asymptotics, (3) Hansen's (1999) grid bootstrap that resamples residuals, and (4) Romano and Wolf's (2001) subsampling method. It is shown that the three first methods are uniformly valid. This is done by splitting the parameter space up into two overlapping sample-size dependent regions: the stationary region and the near-unit root region. It's typically standard to show that convergence in distribution of the test statistic is uniform in the stationary region, but Skorokhod embeddings and stochastic process theory are necessary to treat the near-unit root region. The equi-tailed Romano-Wolf subsampling procedure is not uniformly valid (but pointwise valid) because it's possible to construct a drifting parameter sequence ρ_T for which the test is asymptotically oversized. Monte Carlo results confirm the validity of the three first tests and the subpar performance of the subsampling test. Finally, the results are extended to the AR(p) case.

Theory The model is

$$y_j = c + x_j, \quad x_j = \rho x_{j-1} + \varepsilon_j, \quad j = 1, \ldots, T, \quad x_0 = 0.$$

$\{\varepsilon_j\}$ is a MDS with finite moments of order $r \in (2, 4]$. Let $\Theta = (-1, 1)$. A confidence set $C(Y)$ is said to be asymptotically valid if

$$\liminf_{T\to\infty} \inf_{\rho\in\Theta} P_\rho\{\rho \in C(Y)\} \geq 1 - \alpha.$$

This is contrasted with pointwise validity, which only requires

$$\inf_{\rho\in\Theta} \lim_{T\to\infty} P_\rho\{\rho \in C(Y)\} \geq 1 - \alpha.$$

Because convergence at some values of ρ can be much slower than at other values, uniformity is desirable.

The tests considered in this paper are of the following form. Let $\varphi(Y, T, \rho_0)$ be a test statistic for $H_0\colon \rho = \rho_0$ with lower and upper critical values $c_1(T, \rho_0), c_2(T, \rho_0)$. Then define the set

$$C(Y) = \{\rho \in \Theta\colon c_1(T, \rho) \leq \varphi(Y, T, \rho_0) \leq c_2(T, \rho)\}.$$

Typically (and in particular for the Andrews, Hansen and Stock tests), the critical values are quantiles of a distribution that asymptotically approximates the distribution of $\varphi(Y, T, \rho)$. If this asymptotic approximation is uniform over Θ, then $C(Y)$ is a uniformly valid confidence set (Lemma 1). Mikusheva further focuses on test statistics φ that can be expressed as functions of

$$(S(T, \rho), R(T, \rho)) = \left(\frac{1}{\sqrt{g(T, \rho)}} \sum_{j=1}^{T} y_{j-1}^{\mu}(y_j - \rho y_{j-1}), \frac{1}{g(T, \rho)} \sum_{j=1}^{T} (y_{j-1}^{\mu})^2 \right),$$

where $y_j^{\mu} = y_j - T^{-1} \sum_{i=1}^{T} y_{i-1}$, and $g(T, \rho) = E_{\rho}[\sum_j (y_{j-1}^{\mu})^2]$ is a normalizing function. This choice for $g(T, \rho)$ ensures that the rate is correct to cover both stationary and (near-)unit root cases. There are some further technical restrictions on the class of test statistics φ considered.

The main proof idea is to split the slightly extended parameter set $\Theta_T = [-1 - \theta/T, 1 + \theta/T]$ (for a constant $\theta > 0$) into two *overlapping* regions $\mathcal{A}_T$ and $\mathcal{B}_T$, $\mathcal{A}_T \cup \mathcal{B}_T = \Theta_T$. The stationary region $\mathcal{B}_T$ is separated from the unit root by a neighborhood that is contracting at a rate slower than $1/T$. In this region, the usual asymptotic normality of $S(T, \rho)$ obtains, and $R(T, \rho)$ converges to a variance estimate. The near-unit region $\mathcal{A}_T$ is contracting towards 1 at an even slower speed. Approximating the asymptotic distribution of the test statistic in this region is harder and requires stochastic process theory.

Mikusheva first establishes the asymptotic validity of Andrews' (1993) method. He suggests obtaining the quantiles $c_1(T, \rho)$ and $c_2(T, \rho)$ from Monte Carlo simulations of the finite-sample distribution of the usual t-statistic for ρ, under the assumption that the disturbances were in fact i.i.d. Gaussian. The proof of uniform validity proceeds by establishing uniform convergence of the distribution of $\varphi(Y, T, \rho)$ on $\mathcal{B}_T$; this follows from available results in the literature, since here ρ is bounded away from a $O(1/T)$ region around unity. On $\mathcal{A}_T$ the idea is that the MDS disturbances ε_j that determine the actual statistic φ are not that different from the simulated Gaussian disturbances. Indeed, define the partial sum $S_j = \sum_{i=1}^{j} \varepsilon_i$ and the normalized one $\eta_T(t) = (1/\sqrt{T}) S_{[tT]}$. By Skorokhod's embedding scheme, the probability space can be enlarged to obtain a sequence of BM's w_T s.t. for all $\varepsilon > 0$,

$$\sup_{0 \leq t \leq 1} |\eta_T(t) - w_T(t)| = o(T^{-1/2+1/r+\varepsilon}) \quad \text{a.s.}$$

Define now the error terms $e_{T,j}/\sqrt{T} = w_T(j/T) - w_T((j-1)/T)$. These are i.i.d. standard normally distributed. Defining $z_{T,j}(\rho) = \rho z_{T,j-1}(\rho) + e_{T,j}$, we obtain a sequence $\{z_{T,j}(\rho)\}$ with the *same* distribution as Andrews' proposed simulated AR(1). Hence, the proof now just has to establish that the actual $\{y_j, \varepsilon_j\}$ are "close" to $\{z_{T,j}(\rho), e_{T,j}\}$ using the Skorokhod embedding.

Mikusheva considers a modification of Stock's (1991) proposal. Under local-to-unity asymptotics $\rho_T = \exp(c/T)$, the statistics $(S(T, \rho), R(T, \rho))$ have weak limits given by integrals of an Ornstein-Uhlenbeck process, indexed by c. Let $c(T, \rho) = T \log \rho$. Stock's method is to obtain the critical

values c_1 and c_2 as the quantiles of the statistic $\varphi(S^{c(T,\rho)}, R^{c(T,\rho)}, T, \rho)$. By construction, the set has correct coverage on the local-to-unity sequence ρ_T. Furthermore, Phillips (1987) showed that as $c \to \infty$, the above-mentioned Ornstein-Uhlenbeck integrals converge to the usual normal distribution and constant (for the variance estimate). This suggests that there shouldn't be a problem in the stationary region $\mathcal{A}_T$ on which $c(T, \rho)$ is very negative. Formally, the proof of asymptotic validity proceeds by approximating the distribution of $\varphi(S^{c(T,\rho)}, R^{c(T,\rho)}, T, \rho)$ to the distribution of Andrews' simulated statistic (then the previous result implies that the distribution is also close to the actual statistic computed from the data). Again, this is accomplished through a Skorokhod embedding.

Hansen (1999) proposed to bootstrap the distribution F_T of the residuals by drawing with replacement from the sample residuals $\hat{\varepsilon}_j = y_j - \hat{\rho} y_{j-1}$, where $\hat{\rho}$ is the OLS estimator. Mikusheva also considers a bootstrap that samples residuals by imposing the null $\rho = \rho_0$. The grid bootstrap then computes the critical values c_1 and c_2 as the quantiles of the bootstrap analogs of S and R. Note that Andrews' method is just a parametric grid bootstrap.

The subsampling procedure proposed by Romano and Wolf works as follows. Let $\hat{\rho}(T)$ be the OLS estimate. Let $b = b_T$ be a block size depending on the sample size, with $b_T \to \infty$ but $b_T/T \to 0$. For each (sequentially overlapping) block of b observations $\{y_j, \ldots, y_{j+b-1}\}$, compute the OLS t-statistic for ρ on this subsample, $\hat{t}_j(b) = (\hat{\rho}_j(b) - \hat{\rho}(T))/\sigma(\hat{\rho}_j(b))$, imposing $\hat{\rho}(T)$ as the null. On each of these subsamples, the observations are drawn from the actual distribution of $\{y_j\}$, so normally the empirical distribution function $L_{T,b}(x) = (T - b + 1)^{-1} \sum_{j=1}^{T-b+1} \mathbf{1}_{\{\hat{t}_j(b) \leq x\}}$ should be a good approximation to the actual distribution function for the full-sample t-statistic. A pointwise valid confidence interval for ρ can be constructed based on the *equi-tailed* quantiles of $L_{T,b}(x)$. However, this confidence interval is not uniformly valid (it would be, however, if the confidence interval were symmetric). The proof sets $\rho_T = 1 + c/b_T$. Because $1/b_T$ goes to zero slower than $1/T$, the usual standard normal asymptotic distribution theory for the full sample of size T would be appropriate. But the subsamples of size b_T are sufficiently small that they should be handled by local-to-unity asymptotics. Hence, the asymptotic coverage of the confidence interval is less than the nominal level.

5.2 Phillips (2011): "Folklore Theorems, Implicit Maps, and Indirect Inference"

Summary For many problems in econometrics involving sample size dependent simulations, the standard delta method and CMT don't apply. Phillips provides some general results that extend the delta method to sample-size dependent transformations, and he also discusses a useful CMT for implicitly defined variables. His main focus is the Indirect Inference Estimator (IIE) for the autoregressive coefficient ρ in an AR(1). The finite-sample bias of the MLE is discussed and the IIE is motivated as a means for bias reduction. Detailed asymptotic bias expansions for the MLE are provided. These expansions highlight a peculiar characteristic of the AR(1) example, namely that the mean of the MLE is continuous in the true ρ for finite n, but the asymptotic distribution changes dramatically around unity. It is then shown how the bias expansions may be used to obtain the asymptotic distribution of the IIE, which is given by an implicit inverse map of the MLE. This transformation turns out to concentrate and generally alter the shape of the asymptotic distribution around the true value relative to the MLE.

Theory: Mapping theorems The standard delta method says that if $d_n(T_n - \theta) \Rightarrow T$ as $n \to \infty$ and the map $\varphi\colon \mathbb{R}^m \to \mathbb{R}^p$ is continuously differentiable at θ will derivative matrix φ'_θ, then

$$d_n(\varphi(T_n) - \varphi(T)) \Rightarrow \varphi'_\theta T.$$

The central idea is that $d_n(\varphi(T_n) - \varphi(T))$ should behave asymptotically like a linear functional $\varphi'_\theta T$. If the function $\varphi = \varphi_n$ also depends on the sample size, a new result is needed. Such situations arise for example in simulation-based estimation. Theorem 1 gives conditions (for the scalar case) under which a result

$$\frac{d_n}{\varphi'_n(\theta)}(\varphi_n(T_n) - \varphi_n(\theta)) \Rightarrow T$$

may be established. The assumption needed on φ_n is that it be relatively equicontinuous in balls of shrinking radius $1/s_n$, where $s_n \to \infty$ but $s_n/d_n \to 0$. This ensures that the expression on the LHS in the above display is asymptotically linear in T_n in a wide enough neighborhood of θ.

Suppose $X_n \Rightarrow X$ on a probability space. The Topsoe-Rubin CMT says that if the set $E = \{x\colon g_n(x_n) \to g(x)\ \forall x_n \to x\}$ has probability 1 under the limit measure P, then $g_n(X_n) \Rightarrow g(X)$. Phillips applies this result to implicit maps. Often times, we encounter variables Y_n that are given implicitly by $X_n = f_n(Y_n)$ or $h_n(X_n, Y_n) = 0$. Lemma 2 and the ensuing discussion gives conditions on the derivatives of h_n for there to exist a sequence of inverse maps g_n such that $Y_n = g_n(X_n)$. If these inverse maps satisfy the Topsoe-Rubin condition, one obtains an implicit function CLT.

Theory: Implicit Inference Estimation The idea of indirect inference (II) is to use simulated data to map the dependence of moments, say, on underlying parameters of interest. Consider a parametric model with parameter θ. We can generate H simulated data trajectories $\{\tilde{y}^h\}_{h=1}^H$ of the same sample size as the actual data y. Let $Q_n(\beta; y)$ be a criterion function depending on data y and a pseudoparameter β. Suppose we estimate the pseudoparameter by

$$\hat{\beta}_n = \arg\min_\beta Q_n(\beta; y).$$

For each simulated path h, we can get

$$\tilde{\beta}_n^h(\theta) = \arg\min_\beta Q_n(\beta; \tilde{y}^h(\theta)).$$

Indirect inference now seeks to calibrate the parameter of interest θ to match the simulated values of $\tilde{\beta}_n^h$ to $\hat{\beta}_n$, for example by the criterion

$$\breve{\theta}_{n,H} = \arg\min_\theta \left\| \hat{\beta}_n - \frac{1}{H}\sum_{h=1}^H \tilde{\beta}_n^h(\theta) \right\|.$$

When H is made arbitrarily large, we get $H^{-1}\sum_{h=1}^H \tilde{\beta}_n^h(\theta) \xrightarrow{p} E\tilde{\beta}_n^h(\theta) \equiv b_n(\theta)$; this is called the *binding function*. Then the IIE can be written

$$\breve{\theta}_n = \arg\min_\theta \|\hat{\beta}_n - b_n(\theta)\|.$$

The estimator is thus implicitly determined by the binding function and $\hat{\beta}_n$. The simple delta method won't suffice for asymptotic theory. Note that in many applications, β and θ refer to the same parameter in the model (e.g., the autoregressive coefficient) and $\hat{\beta}_n$ could be the MLE, which is then being bias-corrected by calibrating θ such that the simulated MLEs, given θ, match the computed MLE from the data.

Theory: First-order Autoregression The model is $y_t = \rho y_{t-1} + \varepsilon_t$, $t = 1, \ldots, n$, where u_t is i.i.d. $\mathcal{N}(0, \sigma^2)$. The MLE is $\hat{\rho}_n = \sum_t y_t y_{t-1} / \sum_t y_{t-1}^2$. For $|\rho| \leq 1$, we can use invariance principles such that asymptotic theory holds more generally than for Gaussian errors. If $|\rho| > 1$, the limiting behavior of $\hat{\rho}_n$ will be distribution dependent.

The binding function in this case is $b_n(\rho) = E_\rho \hat{\rho}_n$. Sections 4.2–4.3 develop exact integral representations and asymptotic expansions for b_n for general ρ, using the theory of ratios of quadratic forms in Gaussian variables. These formulas generalize the classic bias formulas for AR(1) estimation. Figure 1 plots the bias as a function of ρ and for various n. For $|\rho| < 1$ the MLE is biased toward 0, with the bias increasing as $|\rho|$ gets closer to unity. However, just around unity, the bias rapidly decreases. Hence, a linear approximation to b_n around unity is not sufficient, and the more detailed asymptotic expansions are needed. The binding function formulas show that b_n is continuous through $\rho = 1$ for fixed n. However, terms like ρ^{2n} appear in the formulas, so as $n \to \infty$ the relative magnitudes of different terms in the expansions depend drastically on whether $|\rho|$ is exceeded by or exceeds unity.

Theorem 4 gives the asymptotic expansions for the bias of the MLE. We have

$$b_n(\rho) = \begin{cases} \rho - 2\rho/n + O(n^{-2}), & |\rho| \leq 1 \\ \pm 1 \mp 1.7814/n + O(n^{-2}), & \rho = \pm 1 \\ \rho + O(|\rho|^{-n}), & |\rho| > 1 \end{cases}$$

The number 1.7814 is the mean of the limit distribution of $n(\hat{\rho}_n - 1)$ when $\rho = 1$. The theorem also gives a formula for the local-to-unity case $\rho = 1 + c/n$. These formulas are now used to obtain asymptotic distribution theory for the IIE $\breve{\rho}$, which is implicitly given by $\hat{\rho}_n = b_n(\breve{\rho})$. First, Phillips shows that the binding function $b_n(\cdot)$ is monotonic, so that its inverse $f_n = b_n^{-1}$ exists.

For the *stationary case*, the binding function formula implies $b_n'(\theta) = 1 + O(n^{-1})$. Since $f_n' = 1/b_n'$, this gives that the asymptotic local equicontinuity condition for Theorem 1 holds, so that the extended delta method may be applied. Thus,

$$\sqrt{n}(\breve{\rho} - \rho) \sim \frac{1}{b_n'(\rho)} \sqrt{n}(\hat{\rho}_n - \rho) \sim \sqrt{n}(\hat{\rho}_n - \rho) \Rightarrow \mathcal{N}(0, 1 - \rho^2).$$

For the *unit root and local-to-unity* cases, the situation is more difficult, because higher order derivatives don't vanish asymptotically (this corresponds to the rapidly changing derivative of the bias around unity). Instead the asymptotic bias expressions may be used. Let $\xi_n^{\text{ml}} = n(\hat{\rho}_n - 1)$ and $\xi_n^{\text{ii}} = n(\breve{\rho} - 1)$. If we insert $\breve{\rho}$ into the bias formula for the local-to-unity case, these formulas can be rewritten in terms of ξ_n^{ml}, ξ_n^{ii} and a remainder that goes to zero, so $\xi_n^{\text{ml}} \sim h(\xi_n^{\text{ii}})$. From the literature we know what the limiting distribution of ξ_n^{ml} is. We then obtain the limiting distribution of ξ_n^{ii} by applying the implicit CMT and inverting h^{-1}. Figure 4 shows the densities of ξ_n^{ml} and ξ_n^{ii} for a large n and $\rho = 1$. The latter distribution is much more concentrated around 0, with more mass above 0 than for the centered and rescaled MLE. Hence, the IIE achieves bias reduction. The implicit transformation is seen to alter the shape of the asymptotic distribution in a very nontrivial way.

5.3 Jansson and Moreira (2006): "Optimal Inference in Regression Models with Nearly Integrated Regressors"

Summary Constructs optimal invariant one- and two-sided tests on β in the predictive regression $y_t = \alpha + \beta x_t + u_t$, where $x_t = \gamma x_{t-1} + v_t$ and γ is local-to-unity. First, the Gaussian finite-sample

problem is considered. Only tests that are invariant to location shifts of y_t are considered. The maximal invariant has a four-dimensional sufficient statistic. Two of these elements are specific ancillary statistics, i.e., their joint distribution does not depend on the parameter of interest β, only on γ. By conditioning on these elements, the likelihood turns out to be a linear exponential family, so standard optimal testing theory can be used. The authors restrict attention to conditionally similar (wrt. γ) tests and derive the test that is UMP in this class. After having motivated the approach for finite samples, the authors show that similar results can be derived in the asymptotic Gaussian case under local-to-unity asymptotics for γ and local alternatives to β_0. This is done by showing that the limiting experiment has a very similar structure to the finite sample. Finally, the Gaussianity and no-serial-correlation assumptions are dropped, and it is shown that the Gaussian asymptotic power envelope can be obtained asymptotically by a feasible test. The conditional critical values are non-standard and require numerical integration. The last chapter provides some simulation evidence that the new test performs well in terms of size control and is competitive with the Campbell-Yogo (2005) tests in terms of power.

Finite-sample theory The most basic predictive regression model is

$$y_t = \alpha + \beta x_{t-1} + \varepsilon_t^y, \quad x_t = \gamma x_{t-1} + \varepsilon_t^x, \quad \varepsilon_t = (\varepsilon_t^x, \varepsilon_t^y)' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma).$$

The covariance matrix Σ is assumed to be known for now. If β and α are variation free, testing problems that involve β are invariant under location transformations of y_t of the form $y_t \to y_t + a$, so the authors restrict attention to tests that are invariant under such transformations. The maximal invariant is then $M_T = (y_2 - y_1, \ldots, y_T - y_1, x_1, \ldots, x_T)'$. Its log likelihood may be written in a form that depends quadratically on (β, γ) and linearly on a four-dimensional statistic $S = (S_\beta, S_\gamma, S_{\beta\beta}, S_{\gamma\gamma})'$ (p. 684). The latter statistic is therefore sufficient for M_T. The object is therefore to construct an optimal test $\phi\colon \mathbb{R}^4 \to [0,1]$ of $H_0\colon \beta = \beta_0$, where the rejection probability ϕ is a function of S.

The distribution of S is a so-called *curved* exponential family as the dimension of $(\beta, \gamma)'$ is less than that of the minimal sufficient statistic S. However, the two univariate statistics $S_{\beta\beta}$ and $S_{\gamma\gamma}$ only depend on $\{x_t\}$, and so their joint distribution does not depend on β. They are therefore called *specific ancillary statistics*, as their distribution only depends on the nuisance parameter γ. A conditionality argument suggests that we should condition on them, as this only discards information about γ. It just so turns out that the distribution of $(S_\beta, S_\gamma)'$ given $(S_{\beta\beta}, S_{\gamma\gamma})'$ is a linear exponential family. This means that standard optimal testing procedures can be employed.

Consider the one-sided testing problem $H_0\colon \beta = \beta_0$ vs. $H_1\colon \beta > \beta_0$. It is said that the test $\phi(\cdot)$ is conditionally η-unbiased if

$$E_{\beta_0,\gamma}[\phi(S)|S_{\beta\beta}, S_{\gamma\gamma}] \leq \eta \quad \forall \gamma,$$
$$E_{\beta,\gamma}[\phi(S)|S_{\beta\beta}, S_{\gamma\gamma}] \geq \eta \quad \forall \beta > \beta_0,\ \gamma.$$

Any such test is *conditionally η-similar*, i.e.,

$$E_{\beta_0,\gamma}[\phi(S)|S_{\beta\beta}, S_{\gamma\gamma}] = \eta \quad \forall \gamma. \tag{1}$$

The theory of testing in exponential families gives that a test is UMP among conditionally η-similar tests only if it is conditionally η-unbiased. It follows that a test is conditionally UMP η-unbiased iff. it is UMP among conditionally η-similar tests. Hence, the authors search for a UMP test among

those that satisfy condition (1) (note that such tests will also be unconditionally similar, so the class of unconditionally similar tests is larger).

It remains to be shown what the UMP test is. Consider

$$\phi_\eta^*(s) = \mathbf{1}_{\{s_\beta > C_\eta(s_\gamma, s_{\beta\beta}, s_{\gamma\gamma})\}},$$

where the critical value function C_η is implicitly defined by the requirement that $\phi_\eta^*(s)$ satisfy

$$E_{\beta_0}[\phi_\eta^*(S)|S_\gamma, S_{\beta\beta}, S_{\gamma\gamma}] = \eta.$$

Here it is being used that the distribution of S_β conditional on $S_\gamma, S_{\beta\beta}, S_{\gamma\gamma}$ is independent of γ. By construction, the test satisfies (1), and the authors can show using standard techniques that the test is UMP in this class (Theorem 2).

Similar arguments are used to construct a UMP conditionally η-unbiased test for a two-sided hypothesis on β.

The authors compare their approach to Stock and Watson (1996). They wanted to maximize weighted average power

$$\int E_{\beta,\gamma}[\phi(S)]dG(\beta,\gamma)$$

subject to

$$E_{\beta_0,\gamma}[\phi(S)] \leq \eta, \quad \forall\gamma.$$

They showed that the optimal test depends on the weight function G, implying that there does not exist a UMP test of size η. Jansson and Moreira view their approach as complimentary to Stock and Watson's: The former authors derive a stronger conclusion (existence of a UMP test) by confining attention to a strict subset of the test considered by the latter authors.

Asymptotic theory The Gaussianity assumption is maintained at first, but now the authors let $T \to \infty$. The testing problem is standard if γ is bounded away from 1, so they assume local-to-unity asymptotics $\gamma = \gamma_T(c) = 1 + T^{-1}c$ for an unknown constant c. It is known in the literature, that standard t-statistics are no longer asymptotically pivotal in this set-up. Existing tests that are asymptotically valid in the local-to-unity case are all asymptotically biased (since many of them are not asymptotically similar). Hence, they have power less than size for β close to the null.

Local alternatives $\beta = \beta_T(b)$ are considered, where $\beta_T(b) - \beta_0 \propto T^{-1}b$ and b is a fixed constant. This scaling preserves contiguity. The null hypothesis is then $b = 0$. It is shown that the likelihood can be written in a very similar way to the finite sample, with a four-dimensional sufficient statistic R. This statistic has a limiting distribution in terms of functionals of Brownian motion and (b, c). The joint limiting distribution has a similar form to the finite sample situation, and again it is possible to condition on specific ancillaries to get rid of the statistical curvature. Analogously to the finite-sample case, attention is restricted to tests $\pi_T(r)$ that are locally asymptotically conditionally η-similar,

$$\lim_{T\to\infty} E_{\beta_T(0),\gamma_T(c)}[(\pi_T(R) - \eta)g(R_{\beta\beta}, R_{\gamma\gamma})] \quad \forall c, g \in C_b(\mathbb{R}^2),$$

where $C_b(\mathbb{R}^2)$ is the space of bounded, continuous, real-valued functions on $\mathbb{R}^2$ (the reason for using arbitrary functions g instead of conditional expectations is that it is technically simpler than having to work with conditional weak convergence). Then the desired test can be constructed precisely as in the finite sample, except that the test is now based on the realized values of R, while the

critical value function is computed based on the limiting distribution of R. Theorem 5 shows that this test is UMP locally asymptotically conditionally similar, with the power envelope depending on b, c and the correlation ρ between the errors.

Finally, the authors relax the assumptions of Gaussianity and no serial correlation for the error terms (the error in the y_t equation must still be an MDS). It is shown that if a feasible version of R (i.e., one in which the covariance matrix and all other nuisance parameters are estimated) is inserted into the asymptotically optimal test described in the previous paragraph, it attains the Gaussian asymptotic power envelope. They also give integral formulas that aid the numerical computation of the critical value function used for the test.

6 SVARs

6.1 Sims (1980): "Macroeconomics and Reality"

Summary Criticizes large-scale macroeconomic models for their piecewise, intuitive approach to identification, their ad hoc treatment of expectations and their indiscriminate classification of variables as endogenous and exogenous. Suggests using VARs to model the relationship between variables without imposing *a priori* restrictions. Sims shows that the estimated VARs may be used to test interesting economic theories phrased in terms of the exogeneity of certain groups of variables, and it is shown how impulse responses may be generated from a Cholesky decomposition of the error covariance.

Motivation Large-scale behavioral macroeconomic models often rely on incredible restrictions, often based on equation-by-equation intuition. Variables are often treated as exogenous simply because their endogenous determination would be hard to model, or because they are thought of as policy variables. Models that incorporate expectations typically replace the expectation with a distributed lag of past realizations in an ad hoc manner. However, rigorous identification of the expectation terms in an RE model often requires very strong assumptions, since variation in expected future values of a variable is always less rich than variation in the past. Despite their lack of identification, large-scale models serve useful purposes for forecasting and policy analysis. For the former, the incredible restrictions placed on the model may well help improve its out-of-sample performance due to shrinkage effects. For the latter, while the Lucas critique must be taken seriously, policy analysis in reduced-form econometric models is most often concerned with the effects of carefully steering certain policy levers, rather than introducing large policy shifts that may alter the underlying structural model.

Empirics Sims suggests that researchers instead use VARs to estimate dynamic relationships between variables without imposing *a priori* restrictions. Of course, some kind of "smoothness" assumption must be imposed, such as the choice of lag length. After this has been imposed and the model estimated, hypotheses with economic content can be tested, and if these hypotheses appear reasonable, their restrictions may be imposed.

As an example, Sims used quarterly West German and U.S. data on money, real GNP, unemployment, wages, the price level and import prices. Four lags are deemed sufficient by conducting an LR test relative to an eight-lag specification. Sims discusses finite-sample problems arising from

VARs with many free parameters. The sample is split at certain dates to test for stability of the VARs. It is found that the import price equation experienced instability.

Sims does not think the VAR coefficients provide much information about the underlying relationships. He suggests using the moving average representation (MAR) instead (i.e., impulse responses). Because the residuals are correlated, they must be orthogonalized, which here is accomplished by a Cholesky decomposition with the six variables ordered in the way mentioned above. He describes how this triangular representation is to be interpreted as variables sequentially affecting each other.

Money is found to be non-neutral in the short run. The Fed's reaction function is evident in the response of money to unemployment innovations, although this isn't the case in Germany. Price innovations are much larger in Germany than in the U.S. They tend to negatively affect real GNP and unemployment, suggesting that they can be interpreted as adverse supply shocks. A variance decomposition is conducted. Since none of the variables have more than 60% of their variance accounted for by their own innovations, there is much evidence that they all are endogenous.

Sims gives an example of an RE neomonetarist model that implies restrictions on the VAR system. Suppose the system is split into (y, m), where y is quantities and relative prices, whereas m is money. With rational, utility maximizing consumers who *don't* have money balances in their utility function, and if prices are flexible, changes in the money supply can only affect the real economy by changing agents' expectations of the future path of real shocks. If one is willing to assume that persistent taste, production and endowment dynamics impart a lot of serial correlation in the real shocks, then money has a role in influencing the real sector. However, if one is not willing to rely on ad hoc assumptions about these mechanisms, it follows from the model (as in Hall's consumption model) that stationary real variables should be serially uncorrelated. Hence, if the RE theory is to explain the economy without resorting to ad hoc assumptions about tastes and technology, the real sector should be *exogenous*, i.e., causally prior to money in Granger's sense.

This motivates a test for block exogeneity of the real sector. However, Sims notes that old-school monetarists, who reject short-run price flexibility and RE, would instead expect to find causality running from money to the real sector. And an old-school Keynesian who rejects the importance of money all-together would expect y to be causally prior to m. Thus, a test of exogeneity (Granger causality) in the VAR may help inform the debate between warring schools of thought. The test of block exogeneity of real GNP and unemployment is forcefully rejected both in the U.S. and Germany.

6.2 Cochrane and Piazzesi (2002): "The Fed and Interest Rates–A High-Frequency Identification"

Summary Monetary policy shocks identified from monthly SVARs may have a large anticipated component because the VAR is too slow to catch up with changes in the economic environment. High-frequency (daily) interest rate data circumvents the identification problem by looking at the response of the yield curve immediately following a change in the Fed funds target. A regression of changes in the Fed funds target on interest rates immediately before the change suggests that the Fed responds to expected inflation and output embodied in the interest rates. Using two high-frequency measures of unexpected shocks, impulse responses of employment and the CPI are constructed. The response of employment goes the wrong way, although the standard errors are large, as they are for convential SVAR estimates. They can't reject that Fed funds shocks don't influence prices; in fact, the point estimates again go the wrong way. This may explain why long-

term bond yields respond positively to an unexpected Fed funds hike (if contractionary policy lowered future inflation, one would expect longer-term yields to fall). The authors conclude that identifying true shocks is hard. Since most of the Fed's actions are responses to publicly known events, there may not even be any true shocks.

Empirics The authors plot the behavior of short- and long-term interest rates and the Fed funds target over 2001 in Figure 1. It is clear form the picture that some target changes are fully anticipated, while others aren't. The conclusions gleaned from short- and long-term rates in this respect are similar. If measured conventionally with a VAR, all these changes would be ascribed an unexpected component.

Two high-frequency measures of shocks are generated. The first is simply the change in the one-month eurodollar rate (from two days before to the day after) concurrent with a change in the Fed funds rate. The second regresses changes in the target rate on the previous target and interest rates immediately prior to the change. The prior target gets a small negative coefficient, indicating slow mean reversion. Long-term rates predict changes in the target much better than short-term rates, which is evidence against the expectations hypothesis but evidence for a Taylor rule. The Fed thus responds to interest rates (particularly the 5-to-2 year spread) because they embody information about expected inflation and output.

For each shock measure, the monthly horizon-j impulse response of employment and CPI are constructed by regressing the change $y_{t+j} - y_t$ on the monetary policy shock ε_{t+1}. There are no measured shocks in months without target changes. Conventional orthogonalized VAR impulse responses are provided for comparison. Standard errors are large for either of the impulse responses. As for point estimates, the high-frequency shock measures indicate *positive* effects of a target hike on employment (the conventional VAR IRF is downward-sloping). The regression shock measure also indicates a positive effect of a target hike on the price level. This would explain the puzzling observation that long-term yields tend to rise following rate hikes. The high-frequency shock measures seem to affect yields a lot even at long horizons, whereas the conventional VAR shocks only have a transitory effect.

6.3 Rigobon (2003): "Identification Through Heteroskedasticity"

Summary In a model where observed variables are influenced by common and idiosyncratic shocks, identification of the coefficients may be achieved if the system undergoes regime changes in volatility but the coefficients stay constant across regimes. This is a simple matter of theoretically computing the covariance matrix across regimes and counting equations and unknowns. If the number of regimes is large enough, there are overidentifying restrictions, so the hypothesis of coefficient constancy may be tested. The method is illustrated with a data set on emerging market sovereign bond yields, for which standard SVAR identification procedures are not appropriate. The onset of international crises serves to identify the breaks in heteroskedasticity.

Intuition Consider the estimation of demand and supply curves. Due to simultaneity, OLS is biased, as illustrated in the top panel of Figure 1. The cloud of data points is an ellipsis which does not trace out any of the curves very well. However, suppose there are two time periods, and the supply shocks are of larger magnitude in the second period. The cloud of data points, now primarily traced out by the volatile supply curve, will look more like a tilted ellipsis around the

demand curve, cf. the bottom panel of Figure 1. The rotation of the ellipsis thus helps identify the slope of the demand curve. In other words, it serves as a probabilistic instrument.

If, however, the second time period saw an equal increase in the magnitudes of both demand and supply shocks, it would just result in a larger but still flat ellipsis, like in the top panel. Hence, identification through heteroskedasticity requires the *relative* structural variances to change magnitude.

Theory Consider first the simple model

$$p_t = \beta q_t + \varepsilon_t, \quad q_t = \alpha p_t + \eta_t,$$

where ε_t and η_t are the structural uncorrelated errors. The covariance matrix of the reduced form is

$$\Omega = \frac{1}{(1-\alpha\beta)^2} \begin{pmatrix} \beta^2\sigma_\eta^2 + \sigma_\varepsilon^2 & \beta\sigma_\eta^2 + \alpha\sigma_\varepsilon^2 \\ & \sigma_\eta^2 + \alpha^2\sigma_\varepsilon^2 \end{pmatrix}.$$

This gives three equations with four unknowns. Suppose, however, that there are S different regimes $(\sigma_{\eta,s}^2, \sigma_{\varepsilon,s}^2)$, $s = 1, \ldots, S$, but the coefficients (α, β) stay constant across regimes. We can consistently estimate S reduced-form covariance matrices Ω_s from the data, giving us $3S$ equations with $2 + 2S$ unknowns. Hence, if $S \geq 2$, it should be possible to solve for the parameters of interest. However, a rank condition needs to be satisfied, i.e., that the equations are independent. Proposition 1 shows that the rank condition holds if the reduced-form covariance matrices are not proportional, i.e., identification requires relative variances to change across regimes. Note also that (α, β) are only identified up to the transformation $(\alpha, \beta) \to (1/\beta, 1/\alpha)$, as is evident from the structural equations.

The model is generalized to allow for K common shocks and N endogenous variables:

$$Ax_t = \Gamma z_t + \varepsilon_t,$$

where x_t and ε_t are N-dimensional vectors, z_t is a K-dimensional vector of common shocks, A is $N \times N$ and Γ is $N \times K$. The common and idiosyncratic shocks are all mutually uncorrelated, also serially. Normalizations must be imposed on A and Γ. Proposition 2 gives the order conditions for identification. The estimation of the model is done by minimum distance, using the SN^2 restrictions

$$A\Omega_s A' = \Gamma\Omega_{z,s}\Gamma' + \Omega_{\varepsilon,s}, \quad s = 1, \ldots, S.$$

The first stage requires estimation of Ω_s.

Section IV shows that the estimates of (α, β) (but not the structural shock variances) in the simple two-equation model with no common shocks remain consistent under two types of misspecification. First, suppose that the correct number of regimes has been identified as two, but the precise time windows of the two regimes have been misspecified. Then the estimated covariance matrices Ω_{r1} and Ω_{r2} are convex combinations of the true ones:

$$\Omega_{r1} = \lambda_{r1}\Omega_1 + (1-\lambda_{r1})\Omega_2, \quad \Omega_{r2} = (1-\lambda_{r2})\Omega_1 + \lambda_{r2}\Omega_2.$$

Here $\lambda_{r1} = \lambda_{r2} = 1$ corresponds to correct classification. The equations that are used to estimate (α, β) are then

$$\lambda_{r1}\Omega_1 + (1-\lambda_{r1})\Omega_2 = \frac{1}{(1-\alpha\beta)^2} \begin{pmatrix} \beta^2\sigma_{\eta,1}^2 + \sigma_{\varepsilon,1}^2 & \beta\sigma_{\eta,1}^2 + \alpha\sigma_{\varepsilon,1}^2 \\ & \sigma_{\eta,1}^2 + \alpha^2\sigma_{\varepsilon,1}^2 \end{pmatrix},$$

$$(1-\lambda_{r2})\Omega_1 + \lambda_{r2}\Omega_2 = \frac{1}{(1-\alpha\beta)^2}\begin{pmatrix} \beta^2\sigma^2_{\eta,2}+\sigma^2_{\varepsilon,2} & \beta\sigma^2_{\eta,2}+\alpha\sigma^2_{\varepsilon,2} \\ & \sigma^2_{\eta,2}+\alpha^2\sigma^2_{\varepsilon,2} \end{pmatrix}.$$

Let $(\tilde{\sigma}^2_{\eta,s}, \tilde{\sigma}^2_{\varepsilon,s})$ be such that

$$\sigma^2_{\eta,1} = \lambda_{r1}\tilde{\sigma}^2_{\eta,1} + (1-\lambda_{r1})\tilde{\sigma}^2_{\eta,2}, \quad \sigma^2_{\eta,2} = (1-\lambda_{r2})\tilde{\sigma}^2_{\eta,1} + \lambda_{r2}\tilde{\sigma}^2_{\eta,2},$$

and similarly for ε. The two matrix equations above are therefore convex combinations (with weights λ_{rs}) of the *correct* matrix equations, except that the structural variances are replaced with their "tilde" counterparts. It follows that the estimates of (α, β) based on the erroneous windows will be consistent. The only thing that can go wrong is that the rank condition may no longer hold. For example, if $\lambda_{r1} = 1 - \lambda_{r2}$, there is no heteroskedasticity in the misspecified covariance matrices Ω_{r1} and Ω_{r2}.

The other type of misspecification considered is where the number of regimes has been underspecified (if it has been overspecified, the rank condition can't hold). For example, suppose there are actually S^* regimes, but $\hat{S}$ of them have mistakenly been lumped together as one regime, and the remaining $S^* - \hat{S}$ as another regime. In this case, a similar argument to the above shows that the equations determining (α, β) are just linear combinations of the correct equations (again with modified values for the structural shock variances), although the number of equations has been reduced. If the order reduction is not so large that the system becomes unidentified, the estimates of (α, β) based on the misspecified equation system will remain consistent.

Empirics The method is illustrated using data on daily sovereign debt yields from Argentina, Brazil and Mexico. Rigobon is interested in the contemporaneous effects of these yields on each other. However, exclusion, sign or long-run restrictions are dubious in this case. Instead, he exploits the presence of many emerging markets crises during the 1990s and early 2000s, which should be a source of heteroskedasticity. The time period is split into several periods, some tranquil and some corresponding to different crises. The model is

$$Ax_t = c + \phi(L)x_t + \phi\, US_t + \Phi(L)\, US_t + \varepsilon_t + \Gamma z_t,$$

where x_t contains the three sovereign yields, and US_t is the U.S. yield. The reduced form is obtained by premultiplying by A^{-1}. The reduced form residuals ν_t then satisfy

$$A\nu_t = \varepsilon_t + \Gamma z_t,$$

which is of the form studied in the theory sections. The reduced-form residuals are estimated by first running a VAR on the entire sample that removes the influence of lags and U.S. yields. Having isolated the residuals, the MD procedure is used to estimate A and Γ, using various subsets of the crisis periods. Standard errors are computed using a residual bootstrap.

The Mexican yield is significant in the Argentinian equation, but otherwise most coefficients are insignificant. However, the common shocks are very significant. The coefficients seem to change significantly from the first half to the second half of the sample.

6.4 King, Plosser, Stock and Watson (1991): "Stochastic Trends and Economic Fluctuations"

Summary The paper tests whether the U.S. business cycle may be thought of as primarily driven by technology shocks, as suggested by RBC models. In such models, permanent TFP innovations

impart permanent effects on output, consumption and investment, while the ratios between these three variables remain constant (i.e., they are cointegrated in logs). There is therefore one stochastic trend in the system. Only this permanent shock can have long-run effects on the variables; the remaining shocks in the system must be transitory. Starting with the Wold representation of the first difference of the three variable, it is shown that these restrictions identify the three structural shocks, provided it is also assumed that the balanced-growth innovation is orthogonal to the other two innovations. Impulse responses suggest that the balanced-growth innovation does generate the behavior predicted from a TFP shock, and it accounts for most of the business cycle variation. These conclusions, however, are overturned when three nominal variables are added to the system: the balanced-growth shock accounts for less than half the forecast error variance for output and consumption and even less for investment. Furthermore, most of the explanatory power arises from the 1960s. Instead, an important driver of the business cycle seems to be real interest rate innovations. Inflation shocks are surprisingly impotent. The estimated balanced-growth innovations are fairly robustly correlated with outside-the-model estimates of the Solow residual.

Theory Let X_t be an n-dimensional vector of I(1) time series. The reduced-form Wold representation of the first difference is

$$\Delta X_t = \mu + C(L)\varepsilon_t, \quad C_0 = I,$$

where the one-step-ahead forecast errors ε_t are serially uncorrelated with covariance matrix Σ_ε. Consider also a structural model of the form

$$\Delta X_t = \mu + \Gamma(L)\eta_t,$$

where η_t has covariance matrix Σ_η. It follows that $\varepsilon_t = \Gamma_0\eta_t$ and $C(L) = \Gamma(L)\Gamma_0^{-1}$. Possible identification strategies are to (1) impose that certain blocks of $\Gamma(L)$ are zero, so that some variables are exogenous, (2) impose cross-equation restrictions implied by a fully-specified model, or (3) impose restrictions on Σ_η and the matrix of structural impact multipliers Γ_0.

King et al. impose two restrictions. First, the two cointegration restrictions in the above-mentioned three-variable system implies that there is only one permanent innovation, labeled the balanced-growth innovation. This implies that

$$\Gamma(1) = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

where an arbitrary normalization has been imposed (note that the two balanced-growth cointegration restrictions imply that all elements in the first column must be the same). The second restriction is that the balanced-growth innovation is uncorrelated with the two transitory innovations. These restrictions are enough to identify Γ_0 and thus $\Gamma(L) = C(L)\Gamma_0$.

More generally, if there are k shocks with permanent effects, η_1, the long-run structural multiplier can be written $\Gamma(1) = (A, 0)$, where A is $n \times k$. It is natural to suppose that η_1 is uncorrelated with η_2, the $n - k$ shocks with transitory effect. The long-run effects of the innovations on the variables in X_t are $A\eta_1$. King et al. impose that A is lower triangular, which implies that it can be written $A = \tilde{A}\Pi$ for a known $n \times k$ matrix $\tilde{A}$ and a $k \times k$ lower triangular matrix Π. The idea is that $\tilde{A}$ can be chosen so that each shock is connected to a familiar economic mechanism, motivated by the cointegrating relations.

For example, for the six-variable system in output, consumption, investment, real balances, nominal interest rate and inflation, the cointegrating relations suggested by the data are $c - y = \phi_1(R - \Delta p)$ (stable C/Y ratio), $i - y = \phi_2(R - \Delta p)$ (stable I/Y ratio) and $m - p = \beta_y - \beta_R R$ (stable money demand). The structure adopted for A is

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & \phi_1 \\ 1 & 0 & \phi_2 \\ \beta_y & -\beta_R & -\beta_R \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ \pi_{21} & 1 & 0 \\ \pi_{31} & \pi_{32} & 1 \end{pmatrix}.$$

The interpretation is the following for the first shock. It is a balanced-growth shock, since it leads to a unit long-run increase in Y, C and I, as well as a β_y increase in money balances through the money demand equation. Similarly, the second shock can be interpreted as an inflation shock, as it has no permanent effect on C, Y and I but unit long-run effect on the nominal interest rate and inflation, etc. The coefficients in Π ensure that the permanent innovations are mutually uncorrelated and a causal ordering of the shocks has been assumed.

The reduced form of the system is obtained by estimating a VECM and then inverting the AR polynomial to obtain the MA representation. The Appendix shows how to derive an estimate of Γ_0 from the estimated reduced form, given the choice of $\tilde{A}$. Standard errors are computed by bootstrapping.

Finally, the authors suggest that their modeling method may be used to split economic time series into permanent and cyclical components, since their model implies that the long-run forecast of the variables is just an easily computable linear combination of the permanent structural shocks.

6.5 Uhlig (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure"

Summary Uhlig proposes identifying monetary policy shocks in an SVAR by restricting the sign of the impulse responses (out to some horizon) of prices, nonborrowed reserves and the Fed funds rate. The response of output, which is the object of interest, is left unrestricted. A framework for conducting Bayesian inference is presented to get around the inconvenience of a non-singleton identified set. The prior puts zero mass on parameters that imply a violation of the IRF sign restrictions. In the plots, Uhlig reports the median IR along with 68% credible bands (all pointwise). The empirical conclusions arising from the agnostic identification procedure are that the response of output is ambiguous, with monetary policy shocks accounting for very little of the GDP forecast variance, and not much more for other variables.

Theory The reduced-form model is $B(L)Y_t = u_t$, where $E[u_t u_t'] = \Sigma$, where Y_t is m-dimensional. The structural errors v_t satisfy $u_t = Av_t$, so $\Sigma = AA'$. The interest is purely in a monetary policy shock, i.e., in only one of the columns of A, call it a. Let $\tilde{A}\tilde{A}' = \Sigma$ be the lower triangular Cholesky decomposition. Then there exists a vector α of unit length such that $a = \tilde{A}\alpha$. If $r_i(k) \in \mathbb{R}^m$ denotes the impulse responses at horizon k to the i-th Cholesky shock, then the corresponding impulse responses associated with a are $r_a(k) = \sum_i \alpha_i r_i(k)$. The fraction of the variance of the

forecast revision $E_t[Y_{j,t+k}] - E_{t-1}[Y_{j,t+k}]$ that is attributable to the monetary policy shock is

$$\phi_{a,j,k} = \frac{[r_{a,j}(k)]^2}{\sum_i [r_{i,j}(k)]^2}.$$

Given VAR coefficients B, the reduced-form error covariance Σ and a horizon K, the set $\mathcal{A}(B, \Sigma, K)$ contains all a that lead to IRFs satisfying the sign restrictions. This identified set may be numerically traced out by simulating a bunch of *alpha*-vectors, multiplying them by $\tilde{A}(\Sigma)$ to obtain a candidate a and checking whether $a \in \mathcal{A}$.

Uhlig proposes two methods for inference.

- *Pure-sign-restriction approach.* This is a fully Bayesian approach and is Uhlig's preferred method. As in standard Bayesian VAR analysis, the prior on (B, Σ) is Normal-Wishart. The conditional prior on α is uniform on the portion of the unit sphere in $\mathbb{R}^m$ on which $\tilde{A}(\Sigma)\alpha \in \mathcal{A}(B, \Sigma, K)$. This prior is convenient because it leaves the procedure invariant to the choice of decomposition $\tilde{A}$. Furthermore, the posterior is just an indicator function times the standard Normal-Wishart posterior for (B, Σ). Draws from the joint posterior may be obtained by an acceptance algorithm, where first (B, Σ) are drawn from their marginal posterior, then α is drawn uniformly on the entire unit sphere, but the full draw (B, Σ, α) is only retained if $\tilde{A}(\Sigma)\alpha \in \mathcal{A}(B, \Sigma, K)$. Medians and credible bands for IRFs are calculated directly from the posterior draws.

- *Penalty-function approach.* This approach leaves the reduced-form estimation as is and then finds the unique point in the identified set that minimizes a penalty function. The reduced-form parameters are estimated in a Bayesian way, again with a Normal-Wishart prior (and thus posterior) on (B, Σ). Given a draw from the posterior, the associated a vector is taken to be the point in $\mathcal{A}(B, \Sigma, K)$ that minimizes a penalty function that rewards correct signs and heavily penalizes wrong signs. Medians and credible bands are computed from the resulting draws of (B, Σ, a).

Empirics Uhlig uses (sometimes interpolated) monthly data for six time series in log levels. The sign restriction identifies a contractionary monetary policy shock as one that, in the first 5 periods, does not raise prices or nonborrowed reserves but does raise the Fed funds rate. Figure 6 shows the benchmark results, plotted as the posterior median and 68% credible bands (one standard deviation). Figure 5 are results obtained from a conventional point-identified Cholesky approach. The latter IRFs exhibit the "price puzzle" (Sims, 1992): the GDP deflator increases somewhat after a contractionary MP shock before declining. The sign restriction results avoids the price puzzle by construction; however, instead the response of real GDP is ambiguous and most likely close to zero at all horizons. The GDP deflator falls very sluggishly. A variance decomposition (calculated on each of the posterior draws) indicates that MP shocks account for 5–10% of the forecast error variance in real GDP, and only a bit more for the other variables (most explanatory power for prices).

The conventional approach typically restricts the instantaneous response of output to be zero. If this restriction is added to the sign restrictions, the GDP response does seem closer to the conventional wisdom. Uhlig criticizes the literature for having relied on this one critical assumption to generate the desired results.

6.6 Moon, Schorfheide, Granziera and Lee (2011): "Inference for VARs Identified with Sign Restrictions"

Summary IRFs identified by sign restrictions are only partially identified. Hence, inference from Bayesian and frequentist methods differ in large samples. Intuitively, the Bayesian posterior is concentrated on the estimated identified set, so a 90% credible region, say, will lie strictly within it. A frequentist confidence set, however, extends outside of the boundaries of the estimated identified set, since it must achieve a minimal asymptotic coverage rate. Thus, frequentist confidence sets for sign restricted IRFs tend to be substantially larger than Bayesian credible sets. The authors show how to conduct asymptotically valid frequentist inference under the present kind of partial identification. Confidence regions are based on a minimum-distance objective function. Three different types of valid confidence regions are proposed, differing in their degree of conservativeness and computational burden.

Example Consider the VAR with lag order 0: $y_t = u_t$, where y_t is two-dimensional (inflation, output growth) and $u_t \sim \mathcal{N}(0, \Sigma_u)$. The structural shock is ε_t, where the first component is a demand shock and the other a supply shock. Let Σ_{tr} be the lower triangular Cholesky decomposition of Σ, then $y_t = \Sigma_{tr}\Omega_\varepsilon\varepsilon_t$, where Ω_ε is orthogonal and $\varepsilon_t \sim \mathcal{N}(0, I)$. Interest is in the effects of the structural demand shock, i.e., the first column of Ω_ε, call it q. The object of interest θ is the inflation response. Write

$$\phi = (\phi_1, \phi_2, \phi_3)' = (\Sigma_{11}^{tr}, \Sigma_{21}^{tr}, \Sigma_{22}^{tr})', \quad q = (q_1, q_2)' = (\cos\varphi, \sin\varphi)'.$$

The sign restrictions are that the demand shock produces non-negative responses in inflation and output growth, i.e.,

$$\theta \equiv q_1\phi_1 \geq 0, \quad q_1\phi_2 + q_2\phi_3 \geq 0.$$

Take the reduced form parameter ϕ as given. The inequalities and unit length requirement on q impose restrictions on q_1 that can be translated into restrictions on $\theta = q_1\phi_1$. The identified set $\Theta(\phi)$ is the collection of θ values consistent with the restrictions, given ϕ.

Bayesian inference begins by specifying a prior on (ϕ, q), which may be factored as $p(\phi, q) = p(\phi)p(q|\phi)$. Typically, researchers specify a uniform prior for φ, so that $q(\varphi)$ is uniformly distributed on the unit sphere. The prior is truncated so that the sign restrictions are satisfied, i.e., $p(\varphi|\phi) \propto \mathbf{1}_{\{(\cos\varphi)\phi_1 \in \Theta(\phi)\}}$. A change of variables to $\theta = \phi_1 \cos\varphi$ shows that the prior distribution of θ on the identified set is *not* uniform. Conditional on ϕ, the parameter θ does not enter the likelihood. The posterior on θ may be written

$$p(\theta|\phi) = \int p(\theta|\phi)p(\phi|Y)d\phi,$$

where $p(\phi|Y)$ is just a standard VAR posterior on the reduced-form parameters. As the sample size increases, this latter posterior concentrates around the ML estimate $\hat{\phi}$, so $p(\theta|\phi) \approx p(\theta|\hat{\phi})$ in large samples (shown rigorously in Moon and Schorfheide, 2009). But this posterior is of course concentrated on the estimated identified set $\theta(\hat{\phi})$. Any Bayesian credible set for θ must therefore lie strictly within $\theta(\hat{\phi})$.

Consider now frequentist inference. As in Chernozhukov, Hong and Tamer (2007), confidence regions can be constructed by considering an objective function

$$Q(\theta; \phi, W) = \min_{\mu \geq 0, \varphi} \left\| \begin{pmatrix} (\cos\varphi)\phi_1 - \theta \\ (\cos\varphi)\phi_2 + (\sin\varphi)\phi_3 - \mu \end{pmatrix} \right\|_W^2,$$

where W is a p.d. weight matrix and $\|A\|_W^2 = \text{tr}(WA'A)$. It holds that $\theta \in \Theta(\varphi)$ if and only if both $Q(\theta; \phi, W) = 0$ and $\theta \geq 0$. A sample analog of the objective function is obtained by inserting the ML estimator $\hat{\phi}$ and a data-dependent weight matrix. A confidence set is then defined as

$$CS^\theta = \{\theta \colon \theta \geq 0, Q(\theta; \hat{\phi}, \hat{W}) \leq c\}.$$

The critical value c must ensure that $\inf_{\theta \in \Theta(\phi)} \text{Prob}(\theta \in CS^\theta) \geq 1 - \tau$ for large T, where τ is the significance level. Because the frequentist confidence region must cover the entire identified set with high probability, we have the inclusions

$$CR^\theta \subset \theta(\hat{\phi}) \subset CS^\theta,$$

where CR^θ is any Bayesian credible region.

General theory Moon et al. set up a general VAR(p) subject to sign restrictions on certain (functions of) impulse responses. The notation is heavy since they must operate with a lot of selection vectors and keep track of the ranks of these. As in the simple example, frequentist inference is based on a MD objective function $Q(\theta; \phi, W)$, where ϕ are reduced-form parameters and θ are the IRFs of interest. A high-level assumption of asymptotic normality is imposed on $\hat{\phi}$. Two main approaches to constructing confidence regions are developed.

- *Profile objective function approach.* The first approach is completely analogous to the approach in the simple example. The objective function minimizes over both q and μ, where the latter is the slackness in the sign restrictions. The confidence region is defined similarly to in the simple example. The critical value is a quantile from a χ^2 plus truncated χ^2 distribution, and so is easy to calculate. The critical value does not depend on θ. Theorem 1 shows that the confidence region is asymptotically valid. The proof rewrites the objective function in terms of $\sqrt{T}(\hat{\phi} - \phi)$, and conservative choices for q and θ are chosen in the minimization inside the objective function to bound it by some nuisance-parameter-free quantity. A series of inequalities then leads to a quantity that is asymptotically distributed χ^2 plus another pivotal term.

- *Projection approach with moment selction.* The second approach instead defines the objective function G as only minimizing over the slack μ but not q. A joint confidence region for θ and q is then

  $$CS_{(2)}^{\theta,q} = \{\theta, q \colon \|q\| = 1, \theta \geq 0, G(\theta, q; \hat{\phi}, \hat{W}) \leq c_{(2)}(q, \theta)\}.$$

  Projecting onto Θ gives the confidence region

  $$CS_{(2)}^\theta = \left\{\theta \colon \theta \geq 0, \min_{\|q\|=1} (G - c_{(2)}(q, \theta)) \leq 0\right\}.$$

  for θ. Conditional on q and for $\theta \in \Theta(\phi)$, the distribution of G does not depend on θ. Hence the critical value can be written as a function of q alone. Because $Q(\theta; \phi, W) = \min_{\|q\|=1} G(\theta, q; \phi, W)$, we see that if the critical value $c_{(2)}(q, \theta)$ is smaller than the critical value c in the profile objective function approach (for all q), then the new confidence region will be less conservative.

  To get smaller critical values, the moment selection approach of Andrews and Soares (2010) is employed. Essentially, it works by determining how much (normalized) slack there is in

each of the inequality constraints, given q (and $\hat{\phi}$). If this slack is larger than some threshold (that grows slowly with T), then the moment inequality is classified as not binding and simply dropped from the objective function; otherwise, it is retained. The resulting reduced objective function can then be bounded by the minimum over q of a pivotal quantity that only depends on q.

Critical values can then be obtained in two ways: Either the quantity can be bounded conservatively again, thus getting rid of the minimization and ending up with a standard pivotal quantity, or the critical value can simply be simulated, at the computational cost of having to do the minimization many times. The former approach is evidently more conservative, but it is less conservative than the profile objective function approach. Theorem 2 shows that both moment selection approaches are asymptotically valid.

Section 4.4 discusses implementation. First an initial guess for $\Theta(\phi)$ is computed by trying out various values on a grid. Then the boundary of this set are refined in a stepwise fashion. A preliminary guess for any of the confidence regions may be set to $\Theta(\phi)$, and the boundary can then be refined stepwise. The authors have found that the moment selection approach with simulated critical values is very computationally burdensome. The other moment selection approach is easier, but the profile objective function approach is even quicker as it doesn't rely on figuring out which inequalities are binding.

An even more conservative approach is to construct a confidence region CS^{ϕ} for ϕ based on its asymptotic distribution. Then a trivially valid confidence region for θ can be constructed as $CS_U^{\phi} = \bigcup_{\phi \in CS^{\phi}} \Theta(\phi)$.

Zero restrictions on (linear combinations of) the IRFs can be easily incorporated by a slight modification of the objective functions, without changing any of the other ingredients.

Monte Carlo simulations and an empirical application shows that the frequentist confidence regions are often much larger than Bayesian credible sets. The order of conservativeness is sometimes very meaningful, other times insignificant, depending on whether the sign restrictions are close to binding in the true GDP.

As for Bayesian inference, Moon et al. recommend that, since the prior of the IRFs conditional on the reduced form parameters doesn't get updated by the data, it would be useful to report $\Theta(\bar{\phi})$ (e.g., at the posterior mean of ϕ) so that the reader may see whether the conditional prior distribution is concentrated in a particular area of the identified set.

6.7 Blanchard and Quah (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances"

Summary Identifies demand and supply shock in an SVAR of output growth and unemployment by imposing the restriction that the demand shock does not have a long-run effect on output. The responses of output and unemployment to a demand shock are hump-shaped and mirror images of each other. The response of output to a supply shock builds up slowly and then levels off. The response of unemployment to a supply shock is initially positive, then turns mildly negative before converging to zero. Variance decompositions indicate that demand disturbances explain most of the variation in unemployment, whereas the decomposition of output forecast errors are less conclusive. However, the estimated paths of structural shocks informally suggest that most NBER recessions can be explained by demand shocks.

Theory Let $X_t = (\Delta Y_t, U_t)$. The structural representation of the bivariate system is

$$X_t = A_0 e_t + A_1 e_{t-1} + \dots,$$

where e_t is serially uncorrelated and $E[e_t e_t'] = I$. The estimable reduced-form (Wold) representation is

$$X_t = \nu_t + C_1 \nu_{t-1} + \dots,$$

where ν_t is serially uncorrelated and $E[\nu_t \nu_t'] = \Omega$.

The authors impose the restriction $\sum_{j=0}^{\infty} a_{11,j} = 0$, i.e., that the first disturbance in e_t has a zero long-run influence on (the level of) output, Y_t. This disturbance is called a demand disturbance, the other a supply disturbance.

We have $\nu_t = A_0 e_t$ and $A_j = C_j A_0$. The long-run restriction identifies A_0 and thus the entire sequence of IRs. Let S be the lower triangular Cholesky factor of Ω. Because $A_0 A_0' = \Omega$, we have $A_0 = SQ$ for an orthonormal matrix Q. The restriction on the $(1,1)$ entry of

$$\sum_j A_j = (\sum_j C_j) A_0 = (\sum_j C_j) SQ$$

restricts the left column of Q to be orthogonal to $e_1'(\sum_j)S$. Because the right column of Q is orthogonal to the first, and both columns have unit length, this uniquely identifies Q. This procedure is also the estimation procedure used by Blanchard and Quah. One-standard-deviation bands are computed by a residual bootstrap.

Blanchard and Quah argue against the interpretation of the supply shock component of X_t as a "trend," while the demand shock component is the "cycle." If prices are sticky, supply shocks have cyclical effects as well.

In the appendix the authors give an example of an economy in which there is one supply shock but two separate demand shocks. It is shown that the single demand shock estimated by the Blanchard-Quah procedure is not in general an average (of any kind) of the actual two demand shocks. Conditions are given under which their inference is approximately valid.

6.8 Galí (1999): "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?"

Summary RBC models have typically been evaluated based on their ability to fit unconditional second moments in the data. The basic RBC model predicts a high positive correlation between hours and labor productivity, which isn't found in the data. To salvage the model, researchers have added additional shocks. However, the models still have very specific predictions about the *conditional* second moments, i.e., the responses to certain types of shocks. For example, in a flexible price model a technology shock should induce positive comovement between hours and labor productivity. The prediction is the opposite in a NK model: Due to nominal rigidities in the short run, aggregate demand doesn't rise enough to meet the increased productive potential in the economy, so hours initially fall when productivity rises. The author identifies technology and non-technology shocks in a bivariate VAR in the first differences of labor productivity and labor input (hours and effort). The system is subject to the long-run restriction that only technology shocks can have a long-run effect on measured productivity. The conditional correlations and IRFs obtained from the data agree with the NK model's predictions but not the RBC model's. The results are robust to expanding the system to five variables. The identified demand shock accounts for the bulk of postwar business cycles.

Theory The set-up for the two-variable system is exactly as in Blanchard and Quah (1989), except that both variables are in first differences.

Consider instead the extended VAR with five variables, where the first one is the first difference of labor productivity. The structural representation is

$$x_t = C(L)\varepsilon_t, \quad E[\varepsilon_t \varepsilon_t'] = I.$$

Partition $\varepsilon_t = (\varepsilon_t^z, \varepsilon_t^{m'})'$, where ε_t^m is a vector of four demand shocks. Let the reduced form be

$$x_t = A(L)u_t, \quad A_0 = I, \quad E[u_t u_t'] = \Sigma.$$

The relations are $u_t = C_0\varepsilon_t$, $A_j C_0 = C_j$. The identifying restrictions are that all four demand shocks have zero long-run impact on the level of labor productivity, i.e., the first row of $\sum_j C_j$ has zeros in its last four entries. This only allows Galí to identify the technology shock. Write $C_0 = SQ$, where S is the lower triangular Cholesky factor of Σ and Q is orthonormal. Partition $Q = (q, \tilde{Q})$, where q is a 5-dimensional vector. We seek to identify q. Partition $I_5 = (e_1, E_1)$, where E_1 is 5×4. The long-run restrictions are

$$0 = e_1' \left(\sum_j C_j \right) E_1 = e_1' \left(\sum_j A_j S \right) Q E_1 = v\tilde{Q},$$

where $v = e_1'(\sum_j A_j S) \in \mathbb{R}^{1\times 5}$. If $v \neq 0$, then the dimension of its null space is 4. The above equation says that the four columns of $\tilde{Q}$ span this null space. To be orthogonal to $\tilde{Q}$, q must therefore lie in the row space of v. Since v is just a row vector, this means that q is a scalar multiple of v, and the scale is pinned down by the restriction that q has unit length. Thus, the first column of C_0 is identified.

7 Estimation and inference in linearized DSGEs

7.1 Iskrev (2010): "Local identification in DSGE models"

Summary Gives sufficient conditions for local identification of structural DSGE parameters from observed first and second moments. Analytical formulas for calculating the Jacobian are provided. It's emphasized that identification failure can be due to both lack of data (too few moment conditions) or, more commonly, intrinsic ambiguity in the model's mapping from deep parameters to observable features. A case study of the Smets and Wouters (2007) model is undertaken, and it's shown how analysis of the Jacobian can point out which parameters cause the lack of identification. Both full information and limited information estimation (IRF matching) is considered.

Theory A typical linearized DSGE has the form

$$\Gamma_0(\theta)z_t = \Gamma_1(\theta)E_t z_{t+1} + \Gamma_2(\theta)z_{t-1} + \Gamma_3(\theta)u_t, \quad Eu_t = 0,\ E[u_t u_t'] = I,$$

where $z_t = \hat{z}_t - \hat{z}^*$ are deviations from steady state, $\hat{z}^*$ are the steady state values and $\theta \in \Theta \subset \mathbb{R}^k$ are the structural parameters. Let $\Theta' \subset \Theta$ be the determinate region. In this region, the reduced form of the model may be written

$$z_t = A(\theta)z_{t-1} + B(\theta)u_t.$$

Let $\Omega = BB'$. Let $\tau = (\tau_z', \tau_A', \tau_\Omega')'$ denote the non-constant elements of $\hat{z}^*$, A and Ω, i.e., those that depend on the structural parameters. We typically observe only a subset of the z_t series, so we have a measurement equation

$$x_t = s(\theta) + Cz_t, \quad s(\theta) = C\hat{z}^*(\theta), \quad C \in \mathbb{R}^{l \times m}.$$

It's straight-forward to derive the first moment and autocovariances of x_t as functions of A, C and Ω. Let m_T be the collection of unique first and second moments, using all available autocorrelations in a sample of size T.

Let $X = (x_1, \ldots, x_T)$. Identification requires that if $f(X; \tilde{\theta}) = f(X; \theta_0)$ with probability 1, then $\tilde{\theta} = \theta_0$. Local identification only requires this to hold in an open neighborhood around θ_0. Theorem 1 says that a sufficient condition for θ_0 the be globally identified is that $m_T(\tilde{\theta}) = m_T(\theta_0)$ iff. $\tilde{\theta} = \theta_0$, i.e., the mapping from population moments to θ is unique. Theorem 2 says that a sufficient condition for local identification is that the Jacobian $J(\theta) = \partial m_q / \partial \theta'$ has full column rank at θ_0 for some $q \leq T$. A necessary condition for the latter is the order condition $k \leq (T-1)l^2 + l(l+3)/2$, i.e., that the number of structural parameters doesn't exceed the dimension of $m_T(\theta)$. One example of identification failure is if a structural parameter is irrelevant for the statistical properties of the model-generated data (such as the Taylor rule coefficients, cf. Cochrane, 2007). Another example is when two variables enter the equilibrium relations in a way that makes them indistinguishable, say, as a product.

It's best to compute the Jacobian analytically. This may be done in the following way. First, use the chain rule

$$J(T) = \frac{\partial m_T}{\partial \tau'} \frac{\partial \tau}{\partial \theta'}. \tag{2}$$

The first factor is explicitly available from the formulas determining m_T as a function of (A, C, Ω). To get the second, note that $\partial \tau_z / \partial \theta$ only involves steady-state relationships, which are easily differentiable. To get $\partial \tau_A / \partial \theta$ and $\partial \tau_\Omega / \partial \theta$, we compute $\partial \text{vec}(A) / \partial \theta'$ and $\text{vech}(A) / \partial \theta'$ and remove the irrelevant rows. Note that from the reduced form, $E_t z_{t+1} = Az_t$, which can be inserted into the structural equation to yield

$$\Gamma_0 z_t = \Gamma_1 A z_t + \Gamma_2 z_{t-1} + \Gamma_3 u_t,$$

and again using the reduced form,

$$[(\Gamma_0 - \Gamma_1 A)A - \Gamma_2] z_{t-1} + [(\Gamma_0 - \Gamma_1 A)B - \Gamma_3] u_t = 0.$$

We can use the IFT to get $\partial \text{vec}(A) / \partial \theta'$ from $(\Gamma_0 - \Gamma_1 A)A - \Gamma_2$. Then $B = (\Gamma_0 - \Gamma_1 A)^{-1} \Gamma_3$, which gives us $\partial \text{vec}(B) / \partial \theta'$ and thus $\partial \text{vech}(\Omega) / \partial \theta'$.

The decomposition (2) shows that identification depends on two things. First, the structural parameters must map uniquely into the characteristics of the underlying z_t process. This is intrinsic to the economic model. Second, the we must observe enough variables in the measurement equation and the sample size must be large enough that there are enough moments to allow us to map from the observed data to the latent processes z_t.

Practically, Iskrev recommends the following procedure for checking local identification. Draw from a prior on Θ. Determine if $\theta \in \Theta'$ (the determinate region), if not, discard. Then check whether $J(q)$ has full rank at θ. q can be chosen at first as the smallest value for which the order condition holds; typically this will be enough.

A couple of extensions are pursued where estimation is based on a transformation of the first and second moments of the data. The leading example is IRFs. The IRF of the i-th variable in x_t to the j-th shock in u_t in the above model is given by the (i,j) element of $\xi^h = CA^hB$. Identification can be checked by directly computing $\partial \xi^h_{i,j}/\partial\theta'$. This particular estimator doesn't utilize the mean of x_t.

Application As an application, Iskrev carries out a case study of the Smets and Wouters (2007) DSGE with 41 structural parameters, seven shocks and seven observed series. It's demonstrated that $\partial\tau/\partial\theta'$ is rank-deficient at the prior mean. The Jacobian allows the researcher to determine which of the parameters cause the rank deficiency. In this case, it's due to the curvature of the cost function having the same effect (in a linearization) on observed price stickiness as the Calvo parameter, since higher curvature implies smaller adjustments conditional on resetting.

The dependence of the various theoretical IRFs on the structural parameters shows which IRFs we may use to draw inference about certain parameters. In the Smets-Wouters model, each IRF can only identify about half of the structural parameters (but different IRFs identify different parameters, of course).

Finally, the exercise underscores the need to use analytical derivatives. Using finite-difference approximations leads to ambiguous results.

7.2 Komunjer and Ng (2011): "Dynamic Identification of Dynamic Stochastic General Equilibrium Models"

Summary Provides conditions under which structural parameters in a DSGE are identified. Identifiability is defined relative to the spectral density matrix, which avoids dependence on the sample size as in Iskrev (2010). Furthermore, the conditions for local identification are stated directly in terms of the coefficients matrices in the state space representation so that numerical computation of the autocovariances is not necessary. The framework covers both stochastically singular models (the number of disturbances is less than or equal to the number of observable variables) and non-singular ones (more disturbances than observed variables), and measurement error as well as a priori restrictions on the parameters are easily accommodated.

Theory The parameter of interest is $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$. Let X_t^a denote the variables of the model. The log-linearized equilibrium conditions take the form $E_t\Gamma_0(\theta)X_{t+1}^a = \Gamma_1(\theta)X_t^a + \varepsilon_{zt}$. Let X_t be a subvector of X_t^a of state variables. The reduced form of the model, if unique, is then written

$$X_{t+1} = A(\theta)X_t + B(\theta)\varepsilon_{t+1},$$

and since we may not observe all latent variables, the measurement equation is

$$Y_{t+1} = C(\theta)X_t + D(\theta)\varepsilon_{t+1}.$$

Here ε_{t+1} is white noise which includes both the model disturbances $\varepsilon_{z,t+1}$ and potentially measurement error. Let $\Sigma_\varepsilon(\theta) = E[\varepsilon_t\varepsilon_t']$. The VMA($\infty$) representation for Y_t is

$$Y_t = \sum_{j=0}^{\infty} h_\varepsilon(j,\theta)\varepsilon_{t-j} = H_\varepsilon(L^{-1};\theta)\varepsilon_t, \tag{3}$$

where the *Markov parameters* are

$$h_\varepsilon(0,\theta) = \begin{cases} D(\theta), & j = 0 \\ C(\theta)A(\theta)^{j-1}B(\theta), & j \geq 1 \end{cases},$$

and the *transfer function* (the z-transform of the impulse response function) is

$$H_\varepsilon(z;\theta) = D(\theta) + C(\theta)[zI_{n_X} - A(\theta)]^{-1}B(\theta) = \sum_{j=0}^{\infty} h_\varepsilon(j,\theta)z^{-j},$$

with $n_X = \dim(X_t)$. The *spectral density matrix* of Y_t can then be written

$$\Omega_Y(z;\theta) = H_\varepsilon(z;\theta)\Sigma_\varepsilon(\theta)H_\varepsilon(z^{-1};\theta)'.$$

Two parameter vectors θ_0 and θ_1 are defined to be *observationally equivalent* if $\Omega_Y(z;\theta_0) = \Omega_Y(z;\theta_1)$ for all $z \in \mathbb{C}$. This is equivalent with the j-th autocovariance coinciding for all $j \geq 0$. θ_0 is then said to be *locally identifiable* if there exists an open neighborhood around it such that for all θ_1 in this neighborhood, θ_0 and θ_1 are observationally equivalent only if $\theta_0 = \theta_1$.

Equivalent spectral densities can arise if, for given $\Sigma_\varepsilon(\theta)$ multiple quadruples $A(\theta), \dots, D(\theta)$ imply the same transfer function $H_\varepsilon(z;\theta)$, i.e., the same impulse responses to given innovations. However, they can also arise if many pairs of $\Sigma_\varepsilon(\theta)$ and transfer functions give rise to the same spectral density, which means that innovations combine with the propagation mechanism to yield the same autocovariances.

One approach to giving conditions for identifiability would be to analyze the Jacobian of the autocovariances wrt. θ. However, the autocovariances would have to be computed numerically given $A(\theta), \dots, D(\theta)$, and the highest available autocovariance is of order T. Instead, the authors develop conditions directly in terms of $A(\theta), \dots, D(\theta)$.

In the singular case, the number n_ε of disturbances is smaller than or equal to the number of observed variables n_Y. Let $\Lambda^S(\theta)$ be a vectorization of the quadruple $A(\theta), \dots, B(\theta)$ and $\text{vech}(\Sigma_\varepsilon(\theta))$. The spectral density of Y_t depends on θ only through $\Lambda^S(\theta)$. Assumptions are made such that the system is left-invertible, i.e., the disturbances are fundamental (lie in the span of $Y_t, Y_{t-1}, Y_{t-2}, \dots$). Furthermore, the system is assumed to be *minimal*, a term from the optimal control literature. This requires that (1) for any initial state it's possible to design an input sequence that puts the system in any given desired final state, and (2) the initial state can be reconstructed by observing the evolution of the output and input sequence. Such conditions imply restrictions on $A(\theta), B(\theta), C(\theta)$. In practice, for DSGE models one is required to write the system in terms of the smallest possible vector of exogenous and endogenous variables that gets rid of redundant dynamics but still fully characterizes the properties of the model. For example, if X_t can be split into $X_{1,t}$ and $X_{2,t}$, and it's possible to express $X_{2,t+1} = \tilde{A}_2(\theta)X_{1,t} + \tilde{B}_2(\theta)\varepsilon_{t+1}$, then $X_{2,t}$ can be dropped from the state vector and the expression for Y_t can be rewritten to only include $X_{1,t}$ plus measurement error.

The minimality and left-invertibility assumptions imply that observational equivalence arises if and only if the elements of $\Lambda^S(\theta_0)$ and $\Lambda^S(\theta_1)$ are obtained from each other by a particular transformation using square matrices U and T (Proposition 1-S). Let the transformation be denoted $\delta^S(\theta, T, U)$. This implies that to check identifiability, we must check whether the equations $\delta^S(\theta_0, I_{n_X}, I_{n_\varepsilon}) = \delta^S(\theta_1, T, U)$ have a locally unique solution $(\theta_1, T, U) = (\theta_0, I_{n_X}, I_{n_\varepsilon})$. Hence, we get local identifiability if and only if the Jacobian of $\delta^S(\cdot)$ has full rank (Proposition 2-S). Note that it is not generally enough that $\partial\delta^S/\partial\theta'$ has full rank, because the assumptions don't necessarily

ensure that $\Lambda^S(\theta)$ is identifiable from the data. As the authors show on p. 2013, the nullspace of the Jacobian reveals which parameters in θ cause the lack of identification.

In the non-singular case, we have $n_\varepsilon \geq n_Y$. It's then impossible for ε_t to be fundamental, and (3) is not the Wold representation for Y_t. Instead, the system may be written in the *innovations representation*

$$\hat{X}_{t+1|t+1} = A(\theta)\hat{X}_{t|t} + K(\theta)a_{t+1},$$
$$Y_{t+1} = C(\theta)\hat{X}_{t|t} + a_{t+1},$$

where $K(\theta)$ is the steady state Kalman gain, $\hat{X}_{t|t}$ is the optimal linear predictor of X_t based on the history of Y_t, and a_{t+1} is the one-step ahead forecast error. The innovations representation exists if the discrete algebraic Ricatti equation (DARE) has a solution, which the authors give primitive conditions for.

Importantly, a_{t+1} is n_Y-dimensional and fundamental by construction. It's then possible to express the spectral density matrix of Y_t in terms of the transfer function and error covariance in the innovations representation, and the analysis of observational equivalence proceed largely as above. The rank conditions now involve the Kalman gain $K(\theta)$, which must be solved numerically by filtering.

For both the singular and non-singular cases, it's straight-forward to incorporate additional a priori restrictions on θ of the form $\varphi(\theta_0) = 0$. These could for example be steady-state restrictions involving the means of variables, or long-run restrictions. The δ^S function above is then augmented with the $\varphi(\cdot)$ restrictions, and local injectivity in θ around θ_0 of the augmented restriction function is then necessary and sufficient for local identifiability.

Finally, the authors study identification in the model in An and Schorfheide (2007).

8 Dynamic Factor Models

8.1 Onatski (2009): "Testing Hypotheses About the Number of Factors in Large Factors Models"

Summary Considers dynamic factor models with large n and T. The objective is to test the null of k_0 dynamic factors against the alternative that there are $k \in (k_0, k_1]$. The data is subjected to a discrete Fourier transform (DFT). The spectral density matrix at some frequency of interest is then obtained by smoothing over the DFTs at nearby frequencies. The test statistic is a measure of the curvature of the scree plot (associated with the smoothed spectral density matrix) for eigenvalues of order $i = k_0 + 1, \ldots, k_1$. Intuitively, if k is the true number of factors, the largest k eigenvalues should explode in size as $n, T \to \infty$, while the remaining eigenvalues that stem from the idiosyncratic errors should be determined by a Tracy-Widom-like distribution. Hence, the scree plot drops off dramatically after order k, and the curvature test statistic is consistent.

Model and test The model is a generalized dynamic factor model with k factors,

$$X_{it} = \Lambda_{i1}(L)F_{1t} + \cdots + \Lambda_{k1}(L)F_{kt}, \quad t = 1, \ldots, T,\ i = 1, \ldots, n.$$

The idiosyncratic error process is independent of the factor process. Let $\hat{X}_s = T^{-1/2}\sum_{t=1}^{T} X_t e^{-i\omega_s t}$ be the DFT of $X_t = (X_{1t}, \ldots, X_{nt})'$, where $\omega_s = 2\pi s/T$, and similarly for e_t. Let further $\hat{\Lambda}_s = \sum_{j=0}^{\infty}(\Lambda_{j,1}, \ldots, \Lambda_{j,n})' e^{-ij\omega_s}$.

Choose a set of m frequencies ω_s around some frequency of economic interest ω_0. Then calculate the smoothed periodogram estimate $(2\pi m)^{-1}\sum_{s=1}^{m}\hat{X}_s\hat{X}_s'$. Denote its i-th largest eigenvalues by γ_i. The test statistic of $H_0\colon k = k_0$ against $H_a\colon k_0 < k \leq k_1$ is

$$R = \max_{k_0<i\leq k_1}\frac{\gamma_i - \gamma_{i+1}}{\gamma_{i+1} - \gamma_{i+2}}.$$

It rejects for large critical values. Note that the statistic is a measure of the largest curvature of the scree plot for orders $i = k_0 + 1, \ldots, k_1$.

The reason why the above test statistic is sensible is as follows. Under an assumption about the rate at which the factor loadings die out with the lag j, one can write

$$\hat{X}_s = \hat{\Lambda}_0\hat{F}_s + \hat{e}_s + R_s,$$

where the remainder R_s can be bounded uniformly. As usual, the DFTs $\hat{e}_1, \ldots, \hat{e}_m$ converge in distribution to m independent normal vectors whose variance depends on the matrix spectral density of e_t. Consequently, the smoothed periodogram estimate of the matrix spectral density of e_t converges to a complex Wishart random matrix. If properly centered and scaled, its eigenvalues follow the Tracy-Widom law of type 2.

Because $\hat{X}_s$ has an approximate k-factor structure asymptotically, the first k eigenvalues will explode. For example, suppose $k = 1$ and that we are dealing with an exact factor model $X_t = F_t l_t$, where $\lambda_t \in \mathbb{R}^n$. Then $E[X_t X_t'] = E[F_t^2]\lambda_t\lambda_t'$, with eigenvalues $\gamma_1 = ||\lambda||^2 E[F_t^2] = O(n)$ and $\gamma_2 = \cdots = \gamma_n = 0$.

Going back to Onatski's model, if the true number of factors is $k = k_0$, the eigenvalues $\gamma_{k_0+1}, \ldots, \gamma_{k_1}$ will asymptotically equal the first $k_1 - k_0$ eigenvalues of the above-mentioned complex Wishart matrix associated with the idiosyncratic errors. Hence, since R gets rid of the centering and scaling, R will be pivotal and a functional of the Tracy-Widom distribution under H_0. However, under the alternative H_a, the eigenvalue γ_{k_0+1} will explode, so the test will reject w.p.a. 1.

The technical results requires T to diverge faster than m and n (and also n/m be bounded), but simulations show that the test works well even when n is large relative to T.

The test can also be used to estimate the number of factors with high probability by sequentially testing up from some a priori lower bound. The test will then estimate the true number of factors w.p.a. $1 - \alpha$, the confidence level of the test.

9 Forecast Evaluation

9.1 West (2006): "Forecast Evaluation"

Summary West surveys the literature on out-of-sample forecast evaluation, and in particular how to test whether a model significantly outperforms others. The main example is Mean Square Prediction Error (MSPE), but a general framework is introduced as well. The discussion is split up into asymptotic theory that applies to non-nested models, and asymptotic theory that applies to nested models. Non-nested models are handled in a fairly standard way, and asymptotics are based on normal approximations with a possible correction for estimation error if the estimation sample is not very large compared to the prediction sample. Nested models are more challenging because a critical long-run variance is rank deficient under the null of no superior predictive ability. Finally, the bootstrap procedure of White (2000) is discussed as a way of dealing with tests between a large

number of models, maybe even a larger number than the sample size (in which case the previous asymptotics are dubious or infeasible).

Theory Assume first that the true parameters are known, so that predictions don't rely on estimated parameters. The object of interest if Ef_t, which is a vector of moments of predictions or prediction errors, e.g., $Ef_t = Ee_{1t}^2 - Ee_{2t}^2$ if the MSPE criterion is used to compare two models. We have a sample of size P to make predictions with. Let $\bar{f}^* = P^{-1}\sum_t f_t$ be the sample mean. Then

$$\sqrt{P}(\bar{f}^* - Ef_t) \xrightarrow{d} \mathcal{N}(0, V^*),$$

where V^* is the long-run covariance matrix of f_t. The null $Ef_t = 0$ can then be tested with a standard Wald test. The relevant estimator of V^* depends on the application. If (e_{1t}, e_{2t}) are τ-step ahead forecast errors, then they are $(\tau-1)$-dependent, and V^* may be estimated with a NW estimator with $\tau - 1$ lags (this is positive semidefinite, unlike the sample analog of the population V^*). In more general cases, a non-parametric estimator may be needed.

Now suppose we have R observations to estimate the parameters β of some parametric models, and then we compute P prediction observations, so that the total sample size is $R + P$. The estimation can be done either recursively (where data from $T = 1, \ldots, R$ is used to predict $P + 1$, then $T = 1, \ldots, R + 1$ is used to predict $P + 2$, etc.), using a rolling scheme (where the estimation window is always R long) or using a fixed scheme (where only one estimate for β is produced). The prediction errors will be polluted by estimation noise, which may be non-negligible if not $R \gg P$.

Let $f_t(\beta^*)$ be a random variable whose expectation is of interest. For MSPE, we have $f_t(\beta^*) = e_{1t}^2 - e_{2t}^2 = (y_t - X_{1t}'\beta_1^*)^2 - (y_t - X_{2t}'\beta_2^*)^2$. If we set $f_t(\beta^*) = e_{1t}X_{2t}'\beta_2^* = (y_t - X_{1t}'\beta_1^*)X_{2t}'\beta_2^*$, then $Ef_t = 0$ means there is zero correlation between one model's prediction error and another's prediction, which is a type of *forecast encompassing* test (Chong and Hendry, 1986). Let $\hat{f}_{t+1} = f_{t+1}(\hat{\beta}_t)$ be the sample counterpart and $\bar{f} = P^{-1}\sum_{t=R}^{T}\hat{f}_{t+1}$. Also, let $F = \partial Ef_t(\beta^*)/\partial\beta$. Then we can often expand

$$\sqrt{P}(\bar{f} - Ef_t) = \sqrt{P}(\bar{f}^* - Ef_t) + F(P/R)^{1/2}O_p(1) + o_p(1).$$

Here the $O_p(1)$ factor stems from estimation error in the sequence of estimates $\hat{\beta}_t$. The expansion says the uncertainty about Ef_t can be split into (1) the uncertainty that would be present if β^* were known, and (2) additional uncertainty due to estimation. If $P/R \to 0$ (*asymptotic irrelevance*) or $F = 0$, the latter doesn't matter asymptotically, so inference can proceed as described above. Otherwise, estimation error contributes to the asymptotic limit distribution. West gives a general result in Section 5.

Consider a simple Mean Prediction Error example, $f_t = e_t$. Suppose the only parameter is the constant term, $y_t = \beta^* + e_t$, and that a fixed scheme is used. We are interested in testing whether $Ef_t = 0$. This single estimate is $\hat{\beta}_R = R^{-1}\sum_{s=1}^{R} y_s$. Then $\hat{e}_{t+1} = e_{t+1} - (\hat{\beta}_R - \beta^*) = e_{t+1} - R^{-1}\sum_{s=1}^{R} e_s$. Thus,

$$P^{-1/2}\sum_{t=R}^{T}\hat{e}_{t+1} = P^{-1/2}\sum_{t=R}^{T}e_{t+1} - (P/R)^{1/2}\left(R^{-1/2}\sum_{s=1}^{R}e_s\right).$$

Suppose e_t is i.i.d. with finite variance σ^2. Then the vector $(P^{-1}\sum_{s=R}^{T}e_{t+1}, R^{-1}\sum_{s=1}^{R}e_s)'$ is

asymptotically normal with variance $\sigma^2 I_2$, so

$$P^{-1/2}\sum_{t=R}^{T} e_{t+1} - (P/R)^{1/2}\left(R^{-1/2}\sum_{s=1}^{R} e_s\right) \xrightarrow{d} \mathcal{N}(0,(1+\pi)\sigma^2),$$

where $\pi = \lim_{T\to\infty} P/R$. Thus, estimation error in β^* scales up the variance by a factor $(1+\pi)$.

For the general result to hold, a full-rank condition on a long-run variance must apply. In particular, this requires that the long-run matrix V^* from above is positive definite. But this often won't be the case under the null if models are nested. For example, consider an out-of-sample test for Granger causality in which Model 1 is

$$y_t = \beta_{10} + \beta_{11} y_{t-1} + e_{1t},$$

and Model 2 is

$$y_t = \beta_{20} + \beta_{21} y_{t-1} + \beta_{22} x_{t-1} + e_{2t}.$$

Under the null $\beta_{22} = 0$ of no Granger causality, we have $e_{1t} = e_{2t}$. Consequently, expressions such as $e_{2t}^2 - e_{1t}^2$ for the MSPE are degenerate, so the usual asymptotics break down. Clark and McCracken have developed non-standard limit theory for special cases in a series of papers. For MSPE, we get that $\sqrt{P}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2) = o_p(1)$ and $P(\hat{\sigma}_1^2 - \hat{\sigma}_2^2) = O_p(1)$, unlike in the non-nested case. Simulation evidence indicates that one rejects the null hypothesis of equal predictive power far too rarely if inference is based on the standard asymptotic theory. Intuitively, in finite samples the added estimation error in the larger (alternative) model will bias upward the estimate of its MSPE relative to the MSPE of the null model, leading to too few rejections. Clark and West have suggested that one could increase power by analytically adjusting for the finite-sample difference in the MSPEs.

Finally, the author discusses West's (2000) "reality check" bootstrap procedure, which may be employed even when the number of potential models is very large. The procedure assumes asymptotic irrelevance, $P/R \xrightarrow{} 0$, and the idea is to sample with replacement from the computed prediction errors and then calculate the maximal amount by which any of the alternative models in each bootstrapped sample outperforms the benchmark. The quantiles of these bootstrapped outperformance measures may then be compared with the actual relative performance of the benchmark model in the data.