

VAR, SVAR и BVAR

Теоретические материалы

1 План

1. Векторные авторегрессии (VAR)
2. Работа с данными, необходимая для создания векторной авторегрессии
3. Структурные векторные авторегрессии (SVAR) и причины перехода к ним
4. Байесовский подход к оценке
5. Практическая часть: BEAR Toolbox

2 Векторные авторегрессии (VAR)

Статья [Sims \(1980\)](#) сделала VAR центральной альтернативой традиционному подходу в эмпирической макроэкономике. Поскольку экономике свойственно совместное движение и взаимовлияние переменных, идея многомерной регрессии выглядит естественно. VAR - это предложение аппроксимировать систему уравнений на данных в отсутствие жестких теоретических предпосылок.

ECONOMETRICA

VOLUME 48

JANUARY, 1980

NUMBER 1

MACROECONOMICS AND REALITY¹

BY CHRISTOPHER A. SIMS

Existing strategies for econometric analysis related to macroeconomics are subject to a number of serious objections, some recently formulated, some old. These objections are summarized in this paper, and it is argued that taken together they make it unlikely that macroeconomic models are in fact over identified, as the existing statistical theory usually assumes. The implications of this conclusion are explored, and an example of econometric work in a non-standard style, taking account of the objections to the standard style, is presented.

Проблема «невероятной идентификации» («incredible identification»), о которой говорил Симс, значит, что модель формально идентифицирована, но экономическая правдоподобность ограничений, обеспечивающих эту идентификацию, сомнительна. Например, некоторые переменные могли объявляться экзогенными, потому что так удобнее, а не правдоподобнее.

Симс предлагает сначала оценить достаточно гибкую приведенную форму, не навязывая ей структурные предпосылки («a priori economic knowledge»), и только после этого обсуждать, какие ограничения нужны для причинной интерпретации и насколько мы в них уверены.

Простейшая VAR-модель

Рассмотрим VAR-модель с двумя переменными и одним лагом. Каждая переменная зависит от собственного лага и от лага другой переменной:

$$\begin{aligned}y_{1t} &= \alpha_{11}y_{1,t-1} + \alpha_{12}y_{2,t-1} + u_{1t}, \\y_{2t} &= \alpha_{21}y_{1,t-1} + \alpha_{22}y_{2,t-1} + u_{2t}.\end{aligned}$$

В матричной форме систему можно записать как:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}.$$

Обозначив

$$y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}, \quad A = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}, \quad u_t = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix},$$

получаем компактную запись VAR(1):

$$y_t = Ay_{t-1} + u_t.$$

NB! Число в скобках означает порядок (p) модели, то есть то, сколько в ней лагов.

Если развернуть лагированные значения, получим:

$$y_t = Ay_{t-1} + u_t = A^2y_{t-2} + Au_{t-1} + u_t = \dots = A^h y_{t-h} + \sum_{j=0}^{h-1} A^j u_{t-j}.$$

Если VAR является устойчивой, то есть все собственные значения матрицы A по модулю меньше единицы ($|\lambda_i(A)| < 1$), то

$$A^h \rightarrow 0 \quad \text{при} \quad h \rightarrow \infty,$$

и влияние начального состояния (y_0) затухает. Тогда VAR(1) допускает бесконечное векторное МА-представление (разложение Вольда):

$$y_t = \sum_{j=0}^{\infty} A^j u_{t-j}.$$

Иными словами, текущее значение y_t можно представить как сумму текущих и прошлых инноваций системы, передаваемых во времени через матрицы A^j .

Инструменты, доступные исследователю по итогам оценки VAR:

1. Динамические отклики
2. Прогнозы: безусловные и условные

2.1 Динамические отклики

Рассмотрим для простоты VAR(1):

$$y_t = Ay_{t-1} + u_t.$$

Предположим, что в периоде t происходит разовая инновация δ , а после этого новых инноваций не возникает. Без инновации значение переменных в периоде t составило бы

$$y_t^{(0)} = Ay_{t-1}.$$

При появлении инновации δ получаем

$$y_t^{(1)} = Ay_{t-1} + \delta.$$

Следовательно, непосредственная реакция системы равна

$$y_t^{(1)} - y_t^{(0)} = \delta.$$

В следующем периоде инновация уже отсутствует, однако ее влияние сохраняется через динамику VAR:

$$y_{t+1}^{(1)} - y_{t+1}^{(0)} = A\delta.$$

Через два периода реакция составляет $A^2\delta$, а в общем случае через h периодов:

$$IRF(h) = A^h\delta.$$

Таким образом, импульсный отклик показывает, как единичная инновация, возникшая в одном периоде, передается по системе и влияет на значения переменных в последующие периоды.

Для VAR(p) аналогичное представление записывается через коэффициенты векторного МА-представления:

$$y_t = \mu + \sum_{j=0}^{\infty} \Psi_j u_{t-j}, \quad \Psi_0 = I.$$

Тогда реакция через h периодов на инновацию δ равна

$$IRF(h) = \Psi_h\delta.$$

Важно! Инновации reduced-form VAR в общем случае коррелированы между собой. Поэтому реакция на отдельную reduced-form инновацию не имеет структурной экономической интерпретации.

Здесь мы разобрали неортогонализированные отклики на инновацию в одном уравнении при нулевых инновациях в других уравнениях. В практической работе с неструктурными VAR-моделями нам может понадобиться более естественный вариант: им могут послужить обобщенные импульсные отклики (GIRF) по Pesaran, Shin (1998). Они не обязывают занулять инновации в других уравнениях и показывают статистическую связь между неожиданным изменением одной переменной и второй переменной.

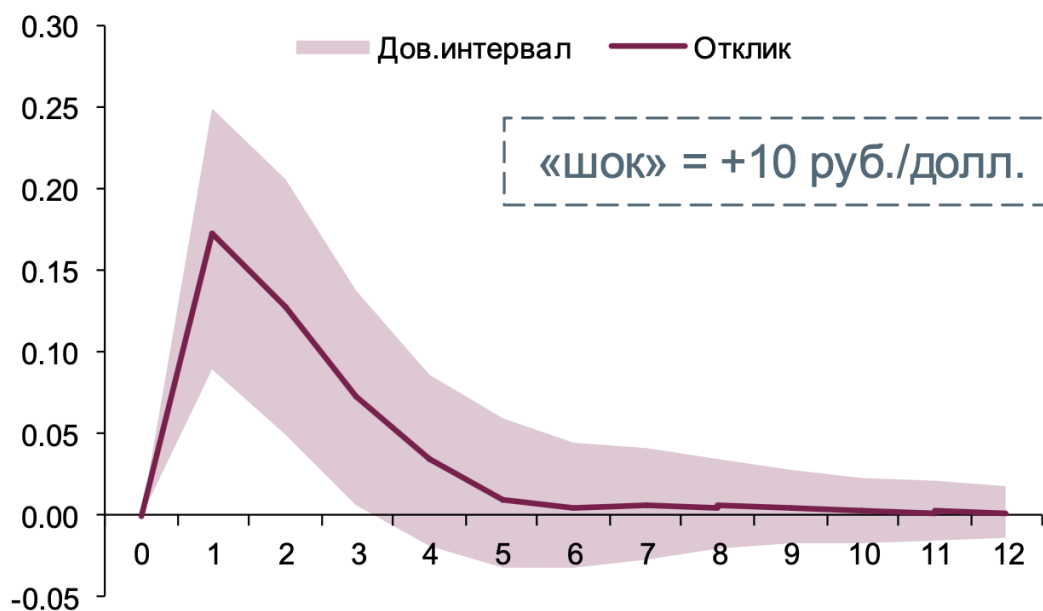
NB! Что стоит знать про GIRF. Обобщенный отклик на шок в j -м уравнении равен

$$GIRF_j(h) = \frac{1}{\sqrt{\sigma_{jj}}} \Psi_h \Sigma_u e_j,$$

то есть остальные инновации не равны нулю, а «подставляются» в среднем по их фактической корреляции с u_{jt} .

- (1) GIRF не зависит от порядка переменных - в этом его главное практическое удобство;
- (2) GIRF не ортогональны между собой: по ним нельзя строить разложение дисперсии и нельзя говорить «шок ДКП». Это статистическая, а не причинная конструкция.

Отклик инфляции на «шок» курса



2.2 Прогнозы: безусловные и условные

Рассмотрим VAR(1):

$$y_t = c + Ay_{t-1} + u_t.$$

Пусть данные наблюдаются до периода T . Поскольку будущие инновации неизвестны, их условное математическое ожидание равно нулю:

$$E_T(u_{T+h}) = 0, \quad h > 0.$$

Поэтому одношаговый прогноз имеет вид

$$\hat{y}_{T+1|T} = c + Ay_T.$$

После получения прогноза на первый период он используется для построения прогноза на следующий:

$$\hat{y}_{T+2|T} = c + A\hat{y}_{T+1|T}.$$

И так далее. Это **безусловный прогноз**: после момента T исследователь не задает никаких дополнительных будущих траекторий переменных. Прогноз определяется только оцененной динамикой VAR и информацией, доступной в момент T .

Условный прогноз строится в том случае, если исследователь дополнительно задает будущую траекторию одной или нескольких переменных. Например, можно зафиксировать будущую траекторию процентной ставки:

$$i_{T+1} = \bar{i}_{T+1}, \quad i_{T+2} = \bar{i}_{T+2}, \quad \dots, \quad i_{T+H} = \bar{i}_{T+H}.$$

Тогда остальные переменные системы прогнозируются при условии заданной траектории ставки. Технически это означает, что будущие инновации $u_{T+1}, \dots, u_{T+H}$ уже не могут все иметь нулевые значения: модель подбирает такую последовательность будущих инноваций, которая согласует динамику VAR с наложенными условиями.

Итоги раздела: в общем случае VAR порядка p записывается как

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \dots + A_p y_{t-p} + u_t,$$

где y_t - вектор из n эндогенных переменных, c - вектор констант, $A_1, \dots, A_p$ - матрицы коэффициентов размерности $n \times n$, а u_t - вектор инноваций приведенной формы.

Ковариационная матрица инноваций

$$E(u_t u_t') = \Sigma_u$$

в общем случае **не является диагональной**: инновации разных уравнений могут быть коррелированы внутри одного периода. Например, если первая переменная - инфляция, а вторая - ставка, то u_{2t} - это не шок ДКП. Это просто неожиданная с точки зрения прошлой информации составляющая ставки, которая может быть **одновременно коррелированной** с неожиданным движением инфляции. В этом состоит основное ограничение неструктурных VAR-моделей.

3 Работа с данными, необходимая для создания векторной авторегрессии

Перед оцениванием VAR необходимо привести временные ряды к виду, который позволяет корректно описывать их совместную динамику. Для макроэкономических данных особенно важны две задачи:

1. устранение сезонности;
2. приведение к стационарному виду.

3.1 Устранение сезонности

Многие макроэкономические показатели имеют выраженную сезонную структуру: объемы производства, бюджетные расходы или цены на отдельные товары могут систематически изменяться в определенные месяцы или кварталы года, и эти изменения не будут объясняться никакими фундаментальными экономическими факторами - только календарными эффектами.

Такую сезонность необходимо устранить до оценки VAR. В противном случае модель может интерпретировать регулярные сезонные колебания как экономическую динамику и переносить их на другие переменные системы. Для сезонной корректировки обычно используются специализированные процедуры, например X-13ARIMA-SEATS или TRAMO/SEATS.



Как это [сделать](#)? ☺

Дополнительный способ - через [веб-интерфейс](#).

3.2 Стационарность

Стационарный временной ряд колеблется вокруг некоторого устойчивого уровня и не меняет со временем основные свойства своей динамики. В частности, для слабо стационарного процесса выполняются условия:

$$E(y_t) = \mu, \quad Var(y_t) = \sigma^2,$$

а ковариация между значениями ряда зависит от расстояния между ними, но не от конкретного момента времени:

$$Cov(y_t, y_{t-h}) = \gamma_h.$$

То есть среднее и дисперсия ряда не должны систематически изменяться во времени, а характер его зависимости от собственного прошлого должен оставаться устойчивым. Рассмотрим простой процесс:

$$y_t = \rho y_{t-1} + u_t, \quad |\rho| < 1.$$

Если в момент t возникает положительная инновация u_t , она повышает значение y_t . Однако в следующих периодах ее влияние постепенно уменьшается:

$$u_t, \quad \rho u_t, \quad \rho^2 u_t, \quad \rho^3 u_t, \quad \dots$$

Именно такое свойство особенно важно для анализа динамических откликов в устойчивой VAR: последствия временной инновации постепенно затухают. Принципиально другой случай возникает для случайного блуждания:

$$y_t = y_{t-1} + u_t.$$

Последовательно подставляя предыдущие значения, получаем

$$y_t = y_0 + \sum_{j=1}^t u_j.$$

Следовательно, каждая инновация навсегда остается в уровне ряда. Такой процесс содержит **единичный корень**.

Что произойдет, если включить нестационарный ряд в обычную VAR? Если переменные обладают единичными корнями, стандартная стационарная VAR в уровнях уже не обладает теми свойствами, на которые мы обычно опираемся при ее анализе. Незатухание инноваций ломает импульсные отклики. Например, для случайного блуждания реакция на единичную инновацию будет иметь вид

$$IRF(h) = 1$$

для любого горизонта h . У устойчивого процесса импульсный отклик постепенно стремится к нулю.

Перейдем к свойству **устойчивости модели**. Для VAR(1)

$$y_t = A y_{t-1} + u_t$$

динамика системы определяется собственными значениями матрицы A . VAR является устойчивой, если

$$|\lambda_i(A)| < 1$$

где λ_i - собственное значение матрицы A . В VAR более высокого порядка аналогичное условие проверяется для собственных значений companion matrix (это способ переписать VAR с несколькими лагами как VAR первого порядка, увеличив размер вектора переменных, чтобы собрать в одной матрице все коэффициенты при лагах). Можно выделить три случая:

$$|\lambda| < 1 \quad \Rightarrow \quad \text{влияние инновации затухает;}$$

$|\lambda| = 1 \quad \Rightarrow \quad$ в системе присутствует единичный корень;

$|\lambda| > 1 \quad \Rightarrow \quad$ траектория является взрывной.

Взрывной характер траектории означает, что импульсные отклики будут не затухать, а наоборот сильно удаляться от изначального значения. Это противоречит логике анализа циклических колебаний, после которых система возвращается обратно.

Как привести ряд к стационарному виду? Способ преобразования зависит от причины нестационарности. Если переменная имеет стохастический тренд и является интегрированной первого порядка, помогут первые разности:

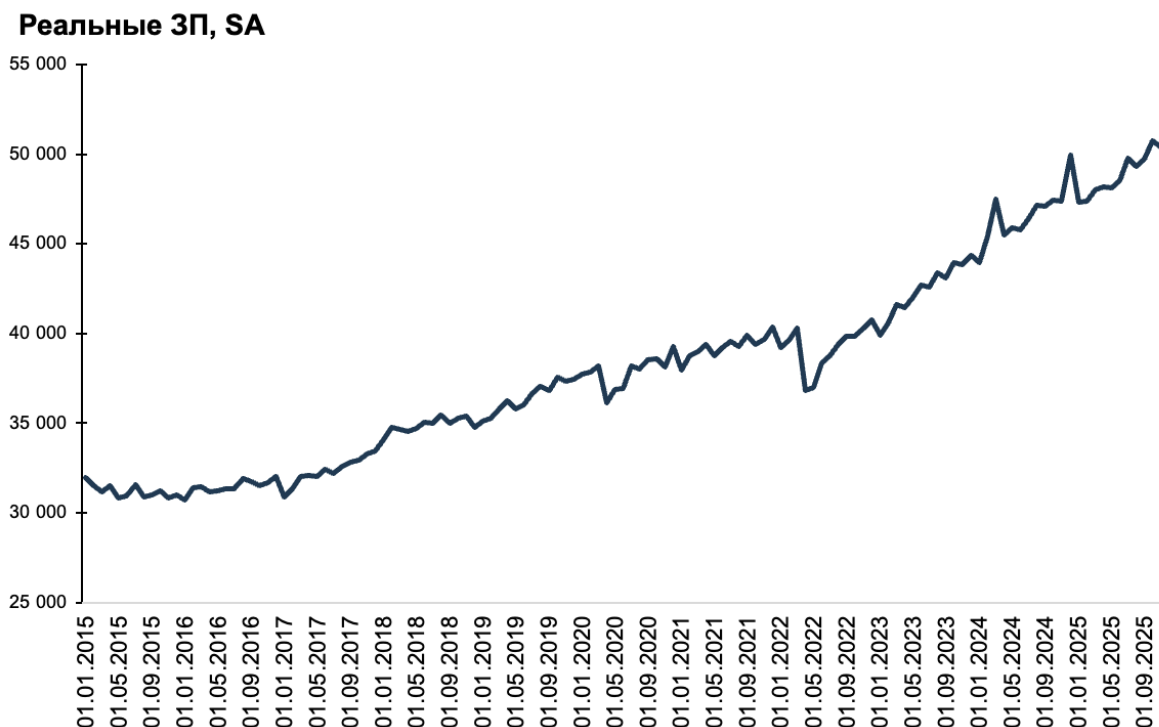
$$\Delta y_t = y_t - y_{t-1}.$$

Для положительных переменных удобно использовать разности логарифмов, так как для небольших изменений эта величина приблизительно соответствует темпу прироста переменной. Это полезно для таких переменных, как, например, ВВП:

$$\Delta \ln y_t = \ln y_t - \ln y_{t-1},$$

$$100\Delta \ln y_t \approx \text{темп роста в процентах.}$$

Но механически не следует дифференцировать все переменные. Нестационарность может быть также обусловлена наличием детерминированного тренда. Тогда его необходимо оценить и использовать отклонение переменной от него.



Как определить, стационарен ли ряд? Надежнее всего - с помощью формальных тестов:

- ADF-тест (Augmented Dickey-Fuller), H_0 : ряд имеет единичный корень;
- KPSS-тест, H_0 : ряд стационарен.



Как это [сделать](#)? ☺

NB! Шпаргалка.

- Логарифмы уровней с высокой персистентностью (реальный ВВП, реальные ЗП) - как правило, лог-разности.
- Ставки, темпы прироста в % г./г., доли и балансы к ВВП - как правило, без преобразований.

3.3 Выбор лагового порядка модели и проверка на автокорреляцию в остатках

Порядок p выбирают по информационным критериям. На практике BIC часто предлагает компактную модель, AIC - более масштабную.

После оценки VAR необходимо проверить, осталась ли в ошибках модели систематическая временная зависимость. Если она есть, инновация u_t не представляет собой новую информацию, значит, ее можно было предсказать на основе прошлого. В остатках сохранилась предсказуемая динамика, которую VAR не смогла описать: это сигнал о возможной мисспецификации модели, а импульсные отклики в этом случае смещены.

Как это [сделать](#)? ☺

4 Структурные векторные авторегрессии (SVAR) и причины перехода к ним

Экономическая, причинная интерпретация инноваций требует дополнительных идентифицирующих ограничений. Именно это позволяет сделать переход к SVAR-моделям: мы получаем возможность определить, какие **структурные шоки** стоят за наблюдаемыми инновациями.

Соотношение между инновациями reduced-form VAR и структурными шоками:

$$u_t = B_0^{-1} \varepsilon_t$$

можно получить из структурной формы (до этого мы рассматривали приведенную форму, reduced form):

$$B_0 y_t = d + B_1 y_{t-1} + \dots + B_p y_{t-p} + \varepsilon_t.$$

Матрица B_0 описывает внутрипериодные связи между текущими значениями переменных. Для системы из двух переменных после нормировки диагональных элементов можно записать:

$$B_0 = \begin{pmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{pmatrix}.$$

Поэтому структурная система в развернутом виде записывается как:

$$\begin{cases} y_{1t} = b_{12} y_{2t} + d_1 + b_{11,1} y_{1,t-1} + b_{12,1} y_{2,t-1} + \dots + b_{11,p} y_{1,t-p} + b_{12,p} y_{2,t-p} + \varepsilon_{1t}, \\ y_{2t} = b_{21} y_{1t} + d_2 + b_{21,1} y_{1,t-1} + b_{22,1} y_{2,t-1} + \dots + b_{21,p} y_{1,t-p} + b_{22,p} y_{2,t-p} + \varepsilon_{2t}. \end{cases}$$

Таким образом, в структурной форме текущие значения переменных могут определяться одновременно: y_{1t} может непосредственно воздействовать на y_{2t} в том же периоде, и наоборот. Именно эта одновременность в общем случае не позволяет оценивать структурные уравнения по отдельности обычным МНК ($Cov(y_{2t}, \varepsilon_{1t}) \neq 0$).

Именно поэтому нам необходимо перейти к приведенной форме и оценить ее, а затем восстанавливать структурные связи. Домножив структурную форму слева на B_0^{-1} , получаем:

$$y_t = B_0^{-1} d + B_0^{-1} B_1 y_{t-1} + \dots + B_0^{-1} B_p y_{t-p} + B_0^{-1} \varepsilon_t.$$

Сопоставляя это выражение с reduced-form VAR, получаем

$$c = B_0^{-1} d, \quad A_i = B_0^{-1} B_i, \quad \boxed{u_t = B_0^{-1} \varepsilon_t}$$

Матрицу B_0^{-1} называют *impact matrix*, поскольку она определяет непосредственную реакцию переменных на структурные шоки в момент их возникновения.

Проблема идентификации: в приведенной форме мы оцениваем Σ_u , но она не дает информации, какое именно разложение этой матрицы соответствует экономическим шокам. Для того, чтобы получить B_0^{-1} , необходимо наложить на нее ограничения.

NB! Сколько нужно ограничений. Структурные шоки нормируют так, чтобы

$$E(\varepsilon_t \varepsilon_t') = I_n$$

- единичная дисперсия и отсутствие корреляции между шоками. Тогда

$$\Sigma_u = E(u_t u_t') = B_0^{-1} E(\varepsilon_t \varepsilon_t') (B_0^{-1})' = B_0^{-1} (B_0^{-1})'.$$

Матрица B_0^{-1} содержит n^2 неизвестных, а симметричная Σ_u дает только

$$\frac{n(n+1)}{2}$$

уравнений. Значит, необходимо дополнительно наложить как минимум

$$\frac{n(n-1)}{2}$$

независимых ограничений - ровно столько, сколько свободных параметров у ортогональной матрицы поворота размера $n \times n$. Ответственность за разумность этих ограничений лежит на исследователе.

4.1 Краткосрочные ограничения и разложение Холецкого

Один из наиболее простых способов идентификации структурных шоков - наложить ограничения на их мгновенное воздействие на переменные. Такие ограничения называются **краткосрочными ограничениями** или **short-run restrictions**. Разложение Холецкого - это специальный рекурсивный случай. Пусть

$$\Sigma_u = PP',$$

где P - нижнетреугольная матрица Холецкого с положительными элементами на диагонали. При рекурсивной идентификации предполагается

$$B_0^{-1} = P.$$

Для системы из трех переменных это означает

$$B_0^{-1} = \begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

Поскольку $u_t = B_0^{-1}\varepsilon_t$, получаем

$$\begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix} = \begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{pmatrix}.$$

Для идентификации Холецкого принципиально важен **порядок переменных**. Переменная, поставленная первой, считается наиболее «медленной»: внутри периода она реагирует только на собственный структурный шок и не реагирует на шоки переменных, стоящих после нее. Последняя переменная, напротив, может реагировать внутри периода на все шоки.

Пример: структурные шоки на мировом рынке нефти по работе [Kilian \(2009\)](#). Рассматривается система

$$y_t = \begin{pmatrix} \Delta prod_t \\ rea_t \\ rpo_t \end{pmatrix},$$

где $\Delta prod_t$ - изменение мировой добычи нефти, rea_t - мировая реальная экономическая активность, а rpo_t - реальная цена нефти. Структурные шоки интерпретируются как

$$\varepsilon_t = \begin{pmatrix} \varepsilon_t^{OS} \\ \varepsilon_t^{AD} \\ \varepsilon_t^{OD} \end{pmatrix},$$

где OS - шок предложения нефти, AD - шок мирового спроса на промышленные товары, а OD - специфический шок спроса на нефть. Рекурсивная структура имеет вид:

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpo} \end{pmatrix} = \underbrace{\begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix}}_{B_0^{-1}} \begin{pmatrix} \varepsilon_t^{OS} \\ \varepsilon_t^{AD} \\ \varepsilon_t^{OD} \end{pmatrix}.$$

Нули помогают нам выделить интерпретируемые шоки: мировая добыча нефти не может существенно отреагировать на изменение спроса внутри одного месяца, так как изменение объемов добычи требует времени; а специфический шок спроса на нефть может сразу изменить ее цену, но не успевает внутри того же месяца изменить мировую экономическую активность.

Как это [сделать](#)? ☺

4.2 Знаковые ограничения

Рекурсивная идентификация требует точных нулевых ограничений: исследователь должен утверждать, что определенный структурный шок вообще не воздействует на некоторую переменную внутри периода. Но во многих задачах нам нужно не ограничиваться занулением реакций или сделать так, чтобы больше шоков влияло на переменные, и тогда очень удобным способом может быть определение знаков реакций.

Пусть reduced-form VAR имеет векторное МА-представление

$$y_t = \mu + \sum_{h=0}^{\infty} \Psi_h u_{t-h}.$$

Выберем некоторую матрицу P , удовлетворяющую

$$\Sigma_u = PP'.$$

P является только одним из бесконечного множества возможных разложений Σ_u . Для любой ортогональной матрицы Q ,

$$QQ' = I,$$

матрица

$$\tilde{B}_0^{-1}(Q) = PQ$$

также является допустимой impact matrix, поскольку

$$(PQ)(PQ)' = PQQ'P' = PP' = \Sigma_u.$$

Для каждого такого кандидата структурные импульсные отклики равны

$$\Theta_h(Q) = \Psi_h PQ.$$

Таким образом, задача состоит не в том, чтобы заранее выбрать одно разложение, а в том, чтобы перебрать множество допустимых структурных разложений и оставить только те из них, импульсные отклики которых соответствуют экономическим ограничениям.

Алгоритм можно представить следующим образом:

1. оценить reduced-form VAR и получить Σ_u ;
2. найти одну матрицу P , такую что $\Sigma_u = PP'$;
3. сгенерировать случайную ортогональную матрицу Q - стандартно через QR-разложение матрицы со стандартными нормальными элементами;
4. построить кандидата $\tilde{B}_0^{-1} = PQ$;
5. рассчитать структурные импульсные отклики;
6. проверить, выполняются ли заданные знаковые ограничения;
7. сохранить реализацию, если все ограничения выполнены, и отбросить ее в противном случае;
8. повторить процедуру большое число раз.

Пример: шоки совокупного спроса и предложения. Рассмотрим положительный шок совокупного спроса и положительный шок совокупного предложения. В классической схеме AD-AS ожидаются следующие реакции:

	Положительный шок AD	Положительный шок AS
Выпуск	+	+
Цены / инфляция	+	—

Знаковые ограничения можно накладывать только в момент возникновения шока или требовать их выполнения в течение нескольких последующих периодов. Для оценки SVAR со знаковыми ограничениями можно использовать пакет VARsignR в R или [пакет A. Cesa-Bianchi](#) в MATLAB.

NB! Знаковые ограничения дают множественную (set) идентификацию. Отобранные Q описывают не одну модель, а целое множество моделей, одинаково согласованных и с данными, и с ограничениями. Это принципиально меняет то, как читаются инструменты:

- Итоговые траектории смешивают две неопределенности: оценки VAR и идентификации. Это не доверительный интервал в привычном смысле, и с ростом выборки он может не снижаться.
- Точечная медиана откликов, как правило, не соответствует ни одной допустимой модели.

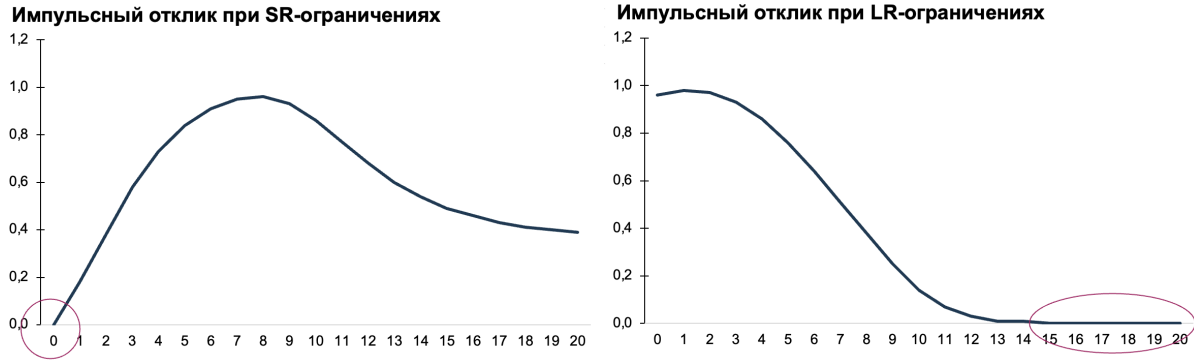
Проблема смещения знаков: предположим, что для трех переменных и трех структурных шоков мы хотим получить следующую матрицу знаков непосредственных откликов:

$$\text{sign}(\Theta_0) = \begin{pmatrix} + & - & + \\ + & + & - \\ + & + & - \end{pmatrix}.$$

Третий столбец является точной противоположностью второго. Чтобы шоки различались, их знаи не должны получаться друг из друга простой сменой знака: нужна хотя бы одна переменная, по которой ограничения различаются.

4.3 Долгосрочные ограничения

Краткосрочные ограничения используют информацию о том, какие переменные могут или не могут отреагировать на шок непосредственно. Иногда экономическая теория, наоборот, дает более убедительные выводы о долгосрочных, а не о краткосрочных эффектах. Например, теория может допускать, что номинальный шок временно влияет на реальный выпуск, но требовать, чтобы уровень выпуска в конечном счете вернулся на прежнюю траекторию.



Reduced-form VAR имеет МА-представление

$$y_t = \mu + \sum_{h=0}^{\infty} \Psi_h u_{t-h}.$$

Поскольку

$$u_t = B_0^{-1} \varepsilon_t,$$

получаем структурное МА-представление

$$y_t = \mu + \sum_{h=0}^{\infty} \Psi_h B_0^{-1} \varepsilon_{t-h}.$$

Матрица структурных импульсных откликов на горизонте h равна

$$\Theta_h = \Psi_h B_0^{-1}.$$

Совокупный долгосрочный эффект структурного шока определяется матрицей

$$\Theta(1) = \sum_{h=0}^{\infty} \Theta_h = \left(\sum_{h=0}^{\infty} \Psi_h \right) B_0^{-1}.$$

Для устойчивой VAR(p)

$$\sum_{h=0}^{\infty} \Psi_h = (I - A_1 - \dots - A_p)^{-1},$$

поэтому

$$\Theta(1) = (I - A_1 - \dots - A_p)^{-1} B_0^{-1}.$$

Long-run restriction представляет собой ограничение на отдельный элемент этой матрицы. Например,

$$[\Theta(1)]_{ij} = 0$$

означает, что структурный шок j не имеет долгосрочного накопленного эффекта на переменную i . Пример: долгосрочная нейтральность денег. Рассмотрим простую двумерную систему

$$y_t = \begin{pmatrix} \Delta \ln Y_t \\ \Delta \ln M_t \end{pmatrix},$$

где Y_t - реальный выпуск, а M_t - денежный агрегат. Предположим, что экономика подвержена двум структурным шокам:

$$\varepsilon_t = \begin{pmatrix} \varepsilon_t^R \\ \varepsilon_t^M \end{pmatrix},$$

где ε_t^R - шок реальной активности, ε_t^M - денежный шок. Денежный шок может влиять на реальный выпуск в краткосрочном периоде, однако не должен изменять его уровень навсегда. Поэтому на long-run impact matrix накладывается ограничение

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

где второй столбец соответствует денежному шоку. После денежного шока темп роста выпуска может сначала повыситься или понизиться, однако сумма этих изменений в долгосрочном периоде равна нулю. Поэтому уровень реального выпуска в конечном итоге возвращается на траекторию, на которой он находился бы без денежного шока. Можно сказать, что long-run identification позволяет разделить устойчивые и переходные источники изменения реальных переменных даже тогда, когда в краткосрочном периоде оба шока способны воздействовать на выпуск.

Классическим примером long-run identification является подход [Blanchard, Quah \(1989\)](#). В статье рассматривается система

$$X_t = \begin{pmatrix} \Delta \ln Y_t \\ U_t \end{pmatrix},$$

где Y_t - реальный выпуск, а U_t - уровень безработицы. Предполагается существование двух структурных шоков: шока спроса и шока предложения. Ключевое идентифицирующее ограничение состоит в том, что шок спроса не оказывает постоянного воздействия на уровень реального выпуска, тогда как шок предложения может иметь такой эффект:

$$\sum_{h=0}^{\infty} IRF_{\Delta \ln Y, \varepsilon^{AD}}(h) = 0.$$

Выбор такой формы ограничений требует осторожности: долгосрочный мультипликатор $\Theta(1)$ зависит от суммы большого числа динамических эффектов, то есть от $(I - A_1 - \dots - A_p)^{-1}$ - величины, которая при персистентных рядах могут оцениваться неточно.

4.4 Новые инструменты: разложение дисперсии и историческая декомпозиция

После идентификации структурных шоков SVAR позволяет не только строить импульсные отклики и прогнозы. **Разложение дисперсии ошибки прогноза (FEVD)** отвечает на вопрос «какая доля непредсказуемых колебаний переменной объясняется данным шоком». Ошибка прогноза на h шагов вперед равна

$$y_{t+h} - \hat{y}_{t+h|t} = \sum_{s=0}^{h-1} \Theta_s \varepsilon_{t+h-s},$$

и, поскольку структурные шоки ортогональны, ее дисперсия аддитивно раскладывается по шокам:

$$FEVD_{ij}(h) = \frac{\sum_{s=0}^{h-1} \theta_{ij,s}^2}{\sum_{k=1}^n \sum_{s=0}^{h-1} \theta_{ik,s}^2}.$$

Историческая декомпозиция отвечает на вопрос «что двигало переменную в конкретные даты». Оценив шоки как $\hat{\varepsilon}_t = B_0 \hat{u}_t$, фактическое значение переменной можно разложить на базовую траекторию (детерминированные члены и вклад начальных условий) и вклады отдельных шоков:

$$y_{i,t} = \text{детерминированная часть и начальные условия} + \underbrace{\sum_{j=1}^n \sum_{h=0}^{t-1} \theta_{ij,h} \hat{\varepsilon}_{j,t-h}}_{HD_{ij,t}},$$

где $\theta_{ij,h}$ - реакция переменной i на шок j через h периодов, а $\hat{\varepsilon}_{j,t-h}$ - фактическая оцененная реализация этого шока в прошлом.

5 Байесовский подход к оценке

Одним из основных недостатков обычных VAR является быстрое увеличение числа оцениваемых параметров при добавлении переменных и лагов. Для $\text{VAR}(p)$ с n эндогенными переменными только число коэффициентов при лагированных эндогенных переменных составляет $n^2 p$. Если добавить константы, необходимо оценить уже $n^2 p + n$ параметров, не считая экзогенных переменных и параметров ковариационной матрицы ошибок.

При ограниченной длине макроэкономических временных рядов это приводит к риску переобучения: модель начинает слишком точно описывать конкретную выборку, но плохо работает за ее пределами. Это называется проклятием размерности.

Идея Bayesian VAR состоит в том, чтобы не исключать большое число коэффициентов из модели полностью, а **стягивать** (shrink) их к экономически разумным априорным значениям. Байесовский подход к оценке помогает сохранить богатую информационную систему через некоторое ограничение «свободы» ее параметров.

Bańbura, Giannone, Reichlin (2008) сравнивают системы, содержащие от нескольких основных макроэкономических показателей до 131 ряда. Основной принцип работы состоит в том, что по мере увеличения размерности VAR априорные ограничения должны становиться более жесткими, чтобы контролировать переобучение при расширении информационного множества. Преимущество BVAR не безусловно: для небольшой системы классическая VAR прогнозирует не хуже байесовской модели, однако с ростом размерности unrestricted OLS-VAR быстро теряет точность. Критически важным это может быть для прогнозов.

Основная идея байесовской оценки: пусть reduced-form VAR записана как

$$y_t = c + A_1 y_{t-1} + \dots + A_p y_{t-p} + u_t, \quad u_t \sim N(0, \Sigma).$$

В матричной форме систему можно представить как

$$Y = XB + U,$$

где B содержит все коэффициенты VAR.

Классическая оценка использует только информацию, содержащуюся в выборке. В байесовском подходе к вероятности наблюдения данных добавляется априорная информация о параметрах. Пусть параметры VAR обозначены через (B, Σ) , а наблюдаемые данные - через Y . По теореме Байеса:

$$p(B, \Sigma | Y) = \frac{p(Y | B, \Sigma) p(B, \Sigma)}{p(Y)}$$

где $p(B, \Sigma | Y)$ - апостериорное распределение параметров, $p(Y | B, \Sigma)$ - likelihood, то есть правдоподобие наблюдаемых данных при заданных значениях параметров, $p(B, \Sigma)$ - априорное распределение параметров, а $p(Y)$ - предельная вероятность данных (marginal likelihood).

Так как знаменатель не зависит от конкретных значений B и Σ и необходим для того, чтобы апостериорное распределение интегрировалось в единицу, часто используется более короткая запись

$$p(B, \Sigma | Y) \propto p(Y | B, \Sigma) p(B, \Sigma).$$

Априорная информация задается тремя связанными между собой элементами:

- **априорные средние** коэффициентов - к чему будут стягиваться оценки;
- **априорные дисперсии** коэффициентов - насколько сильно они будут стягиваться;
- **гиперпараметры** - то, что задает дисперсии, помогают нам не прописывать их для большого количества коэффициентов вручную.

Чем слабее prior, тем большую роль играют данные и тем ближе байесовские оценки становятся к классическим. Роль prior зависит не только от его жесткости, но и от информативности данных. Если likelihood относительно некоторого коэффициента является очень широкой, то данные слабо определяют этот коэффициент, и постериор сильнее зависит от прайора. Если же likelihood концентрирована в узкой области, данные содержат сильный сигнал и постериор может существенно отклоняться от априорного среднего.

В отличие от классической оценки, результатом является не одно значение каждого параметра, а его апостериорное распределение, представленное множеством реализаций. Для практических расчетов из постериора генерируется большое число реализаций:

$$(B^{(s)}, \Sigma^{(s)}) \sim p(B, \Sigma | Y), \quad s = 1, \dots, M.$$

Для каждого posterior draw можно рассчитать прогноз, импульсный отклик или другую интересующую характеристику модели. Полученное множество реализаций используется для приближения соответствующего апостериорного распределения.

5.1 Minnesota prior

Априорное распределение Миннесоты, или Litterman prior, основано на предположении, что **собственная история переменной обычно содержит больше информации о ее будущем, чем лаги других переменных**. Для коэффициента при лаге ℓ переменной j в уравнении переменной i задается некоторый prior mean. Для высокоперсистентного ряда можно задать $\delta_i = 1$, что соответствует random-walk prior; если переменная стационарна - $0 < \delta_i < 1$, а для рядов, уже взятых в разностях, δ_i обычно берут близким к нулю.

Априорная дисперсия коэффициентов при этом уменьшается по мере удаления лага и обычно меньше для лагов других переменных. Априорные дисперсии задаются с помощью гиперпараметров:

$$Var(a_{ij,\ell}) \propto \begin{cases} \frac{\lambda_1^2}{\ell^{2\lambda_3}}, & i = j, \\ \frac{\lambda_1^2 \lambda_2^2}{\ell^{2\lambda_3}} \frac{\sigma_i^2}{\sigma_j^2}, & i \neq j. \end{cases}$$

Здесь σ_i^2 и σ_j^2 - оценки дисперсий остатков отдельных одномерных AR-моделей для соответствующих переменных; их отношение нужно только для того, чтобы уравнивать масштабы переменных, измеренных в разных единицах.

Основные гиперпараметры Minnesota prior:

- λ_1 - общая жесткость прайора, определяет общую силу shrinkage. Чем меньше λ_1 , тем сильнее коэффициенты стягиваются к априорному среднему; чем больше λ_1 , тем большую роль играют данные.
- $\lambda_2 \in (0, 1)$ - cross-variable shrinkage. Определяет, насколько сильнее стягиваются коэффициенты при лагах других переменных по сравнению с собственными лагами. При $\lambda_2 \rightarrow 0$ система распадается на независимые AR-уравнения.
- λ_3 - затухание лагов. Определяет, насколько быстро уменьшается априорная дисперсия при увеличении номера лага. Чем больше λ_3 , тем сильнее «штрафуются» дальние лаги.
- λ_4 - определяет априорную неопределенность относительно констант и других экзогенных детерминированных членов. Большое значение λ_4 соответствует слабому prior на эти коэффициенты.

Более подробно о том, как связаны прайоры и гиперпараметры, можно почитать в работе [Демешева, Малаховской \(2015\)](#).

5.2 Normal-Wishart prior

В классическом Minnesota prior ковариационная матрица ошибок Σ рассматривается как известная или предварительно оцененная (диагональная, по главной диагонали - оценки дисперсий остатков отдельных одномерных AR-моделей). Для анализа это ограничение не всегда удобно.

$$\beta \sim N(\beta_0, V_0), \quad \Sigma = \bar{\Sigma} = \begin{pmatrix} \bar{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & \bar{\sigma}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{\sigma}_n^2 \end{pmatrix},$$

где $\bar{\sigma}_i^2$ - дисперсия остатков одномерной AR-модели для переменной i , а V_0 диагональна.

Normal-Wishart prior рассматривает неизвестными одновременно и коэффициенты VAR, и ковариационную матрицу ошибок. Для коэффициентов используется нормальный prior, а для ковариационной матрицы - inverse-Wishart prior. Σ становится полноценным неизвестным параметром с ненулевыми внедиагональными элементами:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{nn} \end{pmatrix}, \quad \Sigma \sim \text{IW}(S_0, \alpha_0), \quad \beta \mid \Sigma \sim N(\beta_0, \Sigma \otimes \Omega_0).$$

Важное преимущество этого prior - **сопряженность**: после объединения prior и likelihood апостериорное распределение остается в том же семействе, что и априорное. Поэтому posterior выписывается аналитически. Семплирование сводится к двум шагам: сначала $\Sigma^{(s)}$ из inverse-Wishart, затем $\beta^{(s)}$ из нормального при этой $\Sigma^{(s)}$.

В сопряженном Normal-Wishart prior априорная неопределенность коэффициентов B связана с Σ через кронекерову структуру $\Sigma \otimes \Omega_0$. Это делает вычисления простыми, но ограничивает гибкость прайора, так как одна и та же Ω_0 используется для всех уравнений.

Independent Normal-Wishart prior разрывает связь между β и Σ :

$$\beta \sim N(\beta_0, V_0), \quad \Sigma \sim \text{IW}(S_0, \alpha_0), \quad \beta \perp \Sigma,$$

где V_0 - произвольная положительно определённая матрица, не обязанная иметь вид $\Sigma \otimes \Omega_0$. Теперь λ_2 задаётся свободно, что позволяет модели быть более гибкой. Но за это мы платим отсутствием сопряженности, то есть постериор больше аналитически не выписывается, и необходимо прибегать к механизмам семплирования.

Основные shrinkage-гиперпараметры имеют ту же экономическую интерпретацию, что и в Minnesota prior. Разве что в стандартном conjugate Normal-Wishart prior отдельный параметр cross-variable shrinkage отсутствует: соответствующий λ_2 фактически фиксирован на единице из-за кронекеровой структуры.

5.3 Mean-adjusted BVAR и steady-state prior

Часто в обычный VAR мы включаем константу,

$$y_t = c + A_1 y_{t-1} + \dots + A_p y_{t-p} + u_t,$$

однако сама константа c не имеет интерпретации как долгосрочного уровня переменных. Долгосрочное среднее, к которому сходится безусловный прогноз такой модели, равно

$$\mu = (I - A_1 - \dots - A_p)^{-1} c$$

- нелинейной функции всех коэффициентов системы. На персистентных рядах эта величина может оцениваться неточно, и прямого рычага влияния на нее у исследователя нет. В mean-adjusted VAR модель параметризуется относительно долгосрочного среднего напрямую (Villani, 2009):

$$A(L)(y_t - \mu) = u_t, \quad A(L) = I - A_1L - \dots - A_pL^p.$$

Если VAR является устойчивой, то

$$E(y_t) = \mu.$$

Поэтому исследователь может непосредственно задать prior относительно долгосрочного значения каждой переменной. μ_{0i} будет нашим априорным представлением о долгосрочном уровне переменной, а s_{μ_i} - степенью уверенности в этом представлении. Чем она меньше, тем сильнее долгосрочная траектория модели привязана к заданному steady state. Большая априорная дисперсия означает, что долгосрочный уровень в большей степени определяется данными.

Такой прайор полезен, когда исследователь обладает содержательной информацией о долгосрочных значениях переменных, например о долгосрочной инфляции, равновесном уровне процентной ставки или темпе роста потенциального выпуска. Однако mean-adjusted prior требует осторожности: если долгосрочный режим экономики сам является нестабильным ошибочный слишком строгий прайор на steady state способен исказить долгосрочные прогнозы.

Итоги раздела: после оценки reduced-form BVAR с использованием любого из описанного прайора к каждой реализации из постериора могут быть применены те же способы идентификации, которые рассматривались ранее, и на ее основе можно строить инструменты.

6 Практическая часть

Оцениваем BVAR и идентифицируем четыре структурных шока знаковыми ограничениями:

Таблица 1: Знаковые ограничения для идентификации структурных шоков

	ε_{AS} Шок предложения	$\varepsilon_{PrivateAD}$ Шок частного спроса	ε_{MP} Ужесточение ДКП	ε_{FP} Бюджетное расширение
ИПЦ (p)	—	+	—	+
ВВП (Y)	+	+	—	+
Ставка ДКП (i)		+	+	+
Бюджетный баланс (B)		+		—

Для [контрфактического упражнения](#): краткосрочная нейтральная ставка – это процентная ставка, которая уже в текущем квартале позволяет достигнуть цели по инфляции (1% квартал к кварталу, 4% квартал к кварталу в годовом выражении).

Переходим в MATLAB! ☺