Lecture 15 — Bootstrap Methods, GMM, and the Transition to Filtering

Chapter 6 into Chapter 7: resampling, moment conditions, and filter-based signal extraction

Jiajing Sun

School of Economics and Management, University of Chinese Academy of Sciences

Econometrics and Time Series Methods
Spring 2026

Why Lecture 15 matters

Lecture 14 expanded the denominator toolbox. Lecture 15 now asks three follow-up questions:

1. Can we approximate the sampling distribution by **resampling** rather than by analytic asymptotics?
2. Can we estimate structural parameters through **moment conditions** rather than likelihoods?
3. How does robust inference connect to **filtering**, prewhitening, and signal extraction?

Course role

This lecture closes the advanced-inference block and opens the filtering block.

$$\text{resample dependence} \Longrightarrow \text{estimate with moments} \Longrightarrow \text{separate signal from noise}$$

is the conceptual arc of the lecture.

Where Lecture 15 fits in the course

- Lecture 13: spectral intuition and classical HAC inference.
- Lecture 14: fixed-b and self-normalization.
- **Lecture 15:** bootstrap, GMM, and the bridge to filtering.
- Lecture 16: deterministic filtering in time and frequency domains.
- Lecture 17: state-space models and the Kalman filter.

Transition logic

Once we accept that dependence matters, we need broader tools: resampling for finite samples, moment-based estimation for structural models, and filtering for separating signal from noise.

Why these three topics belong together

Each topic is really about the same object: the sampling behavior of dependent sums. Bootstrap resamples them, GMM estimates parameters from them, and filtering reshapes them into signal and noise components.

Learning goals

By the end of the lecture, students should be able to:

1. explain why IID bootstrap fails for dependent time series;
2. distinguish parametric, wild, block-bootstrap, and subsampling ideas;
3. discuss bootstrap design in regression settings with time-series regressors;
4. define GMM from population moment restrictions;
5. explain why the optimal GMM weight is the inverse long-run variance of the moments;
6. interpret the J-test and the C-CAPM application;
7. connect robust inference to filtering in time and frequency domains.

Three-hour plan

Hour 1

Bootstrap methods for dependent data and regression settings.

Hour 2

GMM with time-series moment conditions; C-CAPM illustration.

Hour 3

Filtering in time and frequency domains; common filter types.

From analytic corrections to resampling

So far we have handled dependence by changing asymptotics or changing the denominator.

- HAC estimates a long-run variance analytically.
- Fixed-b changes the asymptotic experiment.
- Self-normalization builds a pivotal statistic from the sample path.

$$\widehat{\mathcal{L}}_T \approx \mathcal{L}\left(\sqrt{T}(\widehat{\theta} - \theta_0)\right) \quad \text{by resampling rather than by closed-form asymptotics.}$$

Bootstrap adds a new philosophy

Approximate the finite-sample distribution by repeatedly re-estimating the statistic on pseudo-samples that mimic the relevant dependence structure.

Why is it called bootstrap?

The term *bootstrap* was introduced by Bradley Efron in 1979 and comes from the old expression *"to pull oneself up by one's own bootstraps."*

- The phrase originally had a self-referential, almost impossible flavor.
- In statistics, that metaphor becomes useful: the sample helps us approximate its own sampling uncertainty.
- In the IID case, the sample induces the empirical distribution

$$\widehat{F}_T(x) = \frac{1}{T}\sum_{t=1}^{T} 1\{y_t \le x\}, \qquad y_t^* \overset{\text{IID}}{\sim} \widehat{F}_T.$$

- For time series, the same philosophy remains, but $\widehat{F}_T$ alone is not enough; the resampling device must also preserve dependence.

Historical meaning in one sentence

Bootstrap inference uses the data themselves, together with a carefully designed resampling rule, to approximate the distribution of the statistic.

Historical note based on Efron (1979) and standard resampling notes on the origin of the term.

Why "jackknife"?

Name and mechanism

Quenouille developed the leave-one-out bias-reduction idea, and Tukey later named it the *jackknife* to emphasize a rough-and-ready general-purpose statistical tool.

$$\widehat{\theta}_{(-t)} = s(\widehat{F}_{T,(-t)}), \quad \widehat{\theta}_{(\cdot)} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\theta}_{(-t)}, \quad \widehat{\mathrm{Bias}}_{\mathrm{jack}} = (T-1)\big(\widehat{\theta}_{(\cdot)} - \widehat{\theta}\big).$$

Bootstrap versus jackknife

- **Jackknife:** deterministic delete-one or delete-block recalculation.
- **Bootstrap:** stochastic resampling with replacement or dependence-preserving pseudo-samples.
- **Jackknife:** often strongest for bias correction and standard errors of smooth statistics.
- **Bootstrap:** usually more flexible for confidence intervals, p-values, and fuller distributional approximation.

Connection

Efron's title *"Another Look at the Jackknife"* signals continuity: bootstrap keeps the resampling spirit but is usually the richer finite-sample tool.

Why the IID bootstrap fails

The ordinary bootstrap samples observations independently with replacement from $\{y_1, \ldots, y_T\}$.

- Serial dependence is destroyed.
- Volatility clustering is destroyed.
- Dynamic regressors lose their time order.

Core principle

For time series, the resampling device must preserve enough dependence structure to mimic the distribution of the statistic of interest.

$$\mathcal{L}^*(\widehat{\theta}^* - \widehat{\theta} \mid y_1, \ldots, y_T) \not\approx \mathcal{L}(\widehat{\theta} - \theta_0)$$

if the resampling step destroys the serial dependence that drives the statistic.

Dependence-preserving principle

The right bootstrap depends on what the data and model look like.

- If a parametric dynamic model is credible, resample shocks within that model.
- If heteroskedasticity matters, use multiplier or wild perturbations.
- If dependence is weak but nontrivial, resample blocks or use subsampling.

No universal winner

Bootstrap design is part of the econometric model, not an afterthought.

Question to ask first

Which features of the data generate the sampling uncertainty: serial correlation, conditional heteroskedasticity, dynamic regressors, nonlinear recursion, or all of them together?

Toy example 1: the mean of a persistent AR(1)

Suppose

$$y_t = 0.8y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \text{IID}(0, \sigma^2),$$

and the statistic of interest is the sample mean $\bar{y}$.

- A typical observed path has *runs*: $1.4, 1.1, 0.9, 0.7, -0.2, -0.4, -0.6, \ldots$
- IID resampling destroys those runs and treats each point as if it were isolated.
- Parametric AR bootstrap or a block bootstrap keeps the local persistence that drives

$$\Omega = \gamma(0) + 2\sum_{h=1}^{\infty} \gamma(h),$$

the asymptotic variance of $\sqrt{T}\bar{y}$.

Classroom message

If the statistic depends on persistence, the bootstrap world must contain persistence too.

Toy example 2: residual bootstrap with a fixed regressor

Consider the deterministic-trend regression

$$y_t = \beta_0 + \beta_1 \frac{t}{T} + u_t, \qquad t = 1, \ldots, T.$$

- The regressor path t/T is fixed by construction.
- We estimate $(\widehat{\beta}_0, \widehat{\beta}_1)$, compute residuals $\widehat{u}_t$, and resample those residuals.
- The bootstrap sample is

$$y_t^* = \widehat{\beta}_0 + \widehat{\beta}_1 \frac{t}{T} + u_t^*.$$

Why this makes sense

The inferential uncertainty comes mainly from the disturbance process, not from the regressor path.

Toy example 3: why wild bootstrap is better under heteroskedasticity

Suppose

$$y_t = \beta_0 + \beta_1 x_t + u_t, \qquad \mathrm{Var}(u_t \mid x_t) = \sigma^2 x_t^2.$$

- Small x_t tends to produce small residuals; large x_t tends to produce large residuals.
- Plain residual resampling mixes those scales together too aggressively.
- Wild bootstrap keeps the local magnitude:

$$u_t^* = z_t^* \widehat{u}_t, \qquad \mathrm{Var}(u_t^* \mid \widehat{u}_t) = \widehat{u}_t^2.$$

Classroom message

Wild bootstrap is a scale-preserving repair when heteroskedasticity, not serial dependence, is the main concern.

Toy example 4: block bootstrap for monthly inflation

Suppose y_t is monthly inflation and dependence is mostly local, with annual seasonal patterns.

- A natural first pass is a block length of $b = 12$.
- One block keeps a full year's short-run dynamics and seasonality together.
- The bootstrap sample concatenates sampled blocks:

$$Y_T^* = (B_{I_1}, \ldots, B_{I_k}), \qquad k = \left\lceil \frac{T}{12} \right\rceil .$$

- If $b = 1$, dependence is destroyed; if b is too large, we have too few effectively distinct blocks.

Classroom message

Block length is the bootstrap analogue of bandwidth choice in HAC.

Parametric bootstrap: when should we trust it?

Parametric bootstrap is attractive when:

1. the time-series model is well specified;
2. innovations are plausibly IID after fitting;
3. recursive simulation from the fitted model is straightforward.

Typical examples

AR, ARMA, ARIMA, and many likelihood-based dynamic models.

$\widehat{\theta}^{*(b)}$ is informative only if the fitted model generates a believable pseudo-world.

Main risk

If the fitted parametric law is wrong, the bootstrap can replicate the wrong dynamic world very accurately.

Parametric bootstrap algorithm

1. Fit the model and collect parameter estimates and residuals.
2. Recenter or standardize residuals if needed.
3. Draw pseudo-innovations from the fitted innovation law.
4. Simulate a bootstrap series recursively from the estimated model.
5. Re-estimate the parameter and recompute the statistic.
6. Repeat many times to approximate the sampling distribution.

$$\widehat{F}_B(c) = \frac{1}{B}\sum_{b=1}^{B} 1\{\widehat{\theta}^{*(b)} \le c\}.$$

Common summaries

From the empirical bootstrap distribution one can extract standard errors, percentile intervals, studentized intervals, or finite-sample p-values.

AR(1) bootstrap example

Suppose

$$y_t = \phi y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \text{IID}(0, \sigma^2).$$

After estimating $\widehat{\phi}$, generate

$$y_t^* = \widehat{\phi} y_{t-1}^* + \varepsilon_t^*.$$

- The pseudo-series inherits the same dependence structure as the fitted AR(1).
- The bootstrap distribution of $\widehat{\phi}^*$ approximates the sampling distribution of $\widehat{\phi}$.
- Initialization matters in short samples, so many implementations use a burn-in or start from an observed value.

$$\text{Var}(y_t) = \frac{\sigma^2}{1 - \phi^2}$$

under stationarity, which is why high persistence makes initialization and burn-in more consequential.

Residual bootstrap with fixed regressors

For

$$y_t = \beta' x_t + u_t,$$

residual bootstrap is often justified only when regressors are strongly exogenous.

$$\mathbb{E}[u_t \mid x_1, \ldots, x_T] = 0 \quad \text{is the comforting benchmark.}$$

$$\widehat{u}_t = y_t - x_t'\widehat{\beta}, \qquad \widetilde{u}_t = \widehat{u}_t - \bar{\widehat{u}}$$
$$y_t^* = x_t'\widehat{\beta} + u_t^*, \qquad u_t^* \sim \{\widetilde{u}_1, \ldots, \widetilde{u}_T\}$$
$$\widehat{\beta}^* = (X'X)^{-1}X'Y^*.$$

Warning

If x_t is dynamically related to u_t, this design can be invalid.

Why wild bootstrap exists

Parametric bootstrap typically assumes homoskedastic innovations.

- Financial and macroeconomic data often violate that assumption.
- Residual variance may change over time.
- Wild bootstrap preserves the fitted dependence structure while randomizing signs or scales.

Connection to Eicker–White logic

The multiplier changes the innovation draw, but leaves the local residual magnitude $|\widehat{\varepsilon}_t|$ in place.

Wild bootstrap construction

Given fitted residuals $\widehat{\varepsilon}_t$, define

$$\varepsilon_t^* = z_t^* \widehat{\varepsilon}_t,$$

where z_t^* has mean zero and variance one.

$$\mathbb{E}[z_t^*] = 0, \qquad \mathrm{Var}(z_t^*) = 1.$$

$$y_t^* = x_t' \widehat{\beta} + z_t^* \widehat{\varepsilon}_t, \qquad \widehat{\beta}^* = (X'X)^{-1} X' Y^*.$$

- Common choices: Rademacher multipliers $z_t^* \in \{-1, 1\}$ with probability $1/2$ each, or Mammen multipliers.
- The pseudo-residual inherits the local scale of $\widehat{\varepsilon}_t$.
- This is especially useful under conditional heteroskedasticity.

When wild bootstrap helps

- Volatility clustering is present.
- Residuals are clearly heteroskedastic.
- The parametric dependence model is reasonable, but homoskedastic Gaussian shocks are not.

Interpretation

Wild bootstrap modifies the innovation law without throwing away the fitted dynamic skeleton of the model.

$$\mathrm{Var}(\varepsilon_t^* \mid \widehat{\varepsilon}_t) = \widehat{\varepsilon}_t^2$$

keeps the local residual scale visible in the pseudo-sample.

Block bootstrap intuition

If we do not want to trust a complete parametric model, we can resample *blocks* of consecutive observations.

- Within each block, local dependence is preserved.
- Across blocks, the pseudo-series becomes an approximate patchwork of observed dependence patterns.
- The key tuning parameter is the block length.

$(y_t, \ldots, y_{t+b-1})$ acts as one dependence-preserving resampling unit.

What block length controls

Dependence at horizons shorter than b is preserved mechanically within each block; dependence at longer horizons is only approximated through the concatenation of many sampled blocks.

Moving block bootstrap algorithm

$$B_i = (y_i, \ldots, y_{i+b-1}), \qquad i = 1, \ldots, N_b, \qquad N_b = T - b + 1.$$

$$I_1, \ldots, I_k \overset{\text{iid}}{\sim} \text{Unif}\{1, \ldots, N_b\}, \qquad k = \left\lceil \frac{T}{b} \right\rceil.$$

$$Y_T^* = (B_{I_1}, \ldots, B_{I_k}) \quad \text{trimmed to length } T.$$

$$\widehat{F}_B(x) = \frac{1}{B} \sum_{m=1}^{B} 1\left\{ \sqrt{T}\left(\widehat{\theta}^{*(m)} - \widehat{\theta}\right) \le x \right\}$$

is the empirical bootstrap law of the root.

Circular and stationary bootstrap

Two common refinements are:

- **Circular bootstrap:** wrap the series around so end-of-sample observations can join beginning-of-sample observations.
- **Stationary bootstrap:** use random block lengths so block boundaries are less mechanically tied to one chosen length.

$$B_i^{\circ} = (y_i, \ldots, y_{i+b-1}), \qquad \text{indices understood mod } T.$$

$$\mathbb{P}(L = \ell) = p(1-p)^{\ell-1}, \qquad \ell = 1, 2, \ldots, \qquad \mathbb{E}[L] = \frac{1}{p} = b.$$

Why refinements matter

They reduce edge effects and can provide smoother finite-sample behavior.

Subsampling

Subsampling avoids bootstrap resampling altogether.

- Choose a window length b.
- Compute the statistic on all or many overlapping subsamples of length b.
- Use the empirical distribution of those subsample statistics as an approximation.

$$\widehat{\theta}_{t,b} \equiv \widehat{\theta}(y_t, \ldots, y_{t+b-1}), \qquad t = 1, \ldots, T-b+1.$$

$$R_{t,b} = \sqrt{b}\,(\widehat{\theta}_{t,b} - \widehat{\theta}_T), \qquad L_{T,b}(x) = \frac{1}{T-b+1} \sum_{t=1}^{T-b+1} 1\{R_{t,b} \le x\}.$$

Advantage

Subsampling needs weaker assumptions and avoids pseudo-random recursive generation.

Trade-off

It usually uses a smaller window $b < T$, so the approximation is more robust but can be noisier than a well-designed bootstrap.

Choosing the block or window length

- Too short: important dependence is broken.
- Too long: effective number of independent resampled units becomes too small.
- Just as bandwidth matters for HAC, block length matters for bootstrap and subsampling.

$$b \to \infty, \qquad \frac{b}{T} \to 0$$

is the block-bootstrap analogue of the familiar small-b HAC regime.

$$\text{effective number of resampled units} \approx \frac{T}{b}.$$

Shared lesson

Resampling does not eliminate tuning choices. It changes which tuning choice is central.

Bootstrap for time-series regression

Regression bootstrap design depends on the data-generating process.

- Strongly exogenous regressors: fixed-regressor residual bootstrap may be appropriate.
- Dynamic regressors: resample (y_t, x_t) blocks or simulate the full system.
- Serially dependent errors: wild or block methods are often more appropriate than IID residual resampling.

$$\text{fixed } x_t: \quad y_t^* = x_t'\widehat{\beta} + \varepsilon_t^*$$

$$\text{joint dependence:} \quad (y_t^*, x_t^*) \text{ resampled jointly, often in blocks}$$

$$\widehat{\beta}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*.$$

Decision rule

Ask whether the uncertainty comes mainly from the regression error, from the regressor process, or from their joint dynamics. Match the bootstrap to that source.

What bootstrap can estimate

Bootstrap methods approximate the sampling distribution of an estimator and can therefore deliver standard errors, confidence intervals, p-values, and bias corrections from the same resampling scheme.

Common interval types

Percentile, basic, bias-corrected, and studentized intervals all use the same resampling engine but summarize it differently.

Practical appeal

Bootstrap is most useful when analytic variance formulas are messy or unreliable in moderate samples.

$$\widehat{p}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} 1\left\{ |T^{*(b)}| \geq |T_{\text{obs}}| \right\}$$

is the bootstrap analogue of a p-value.

$$CI_{\text{perc}} = \left[\widehat{q}_{\alpha/2}(\widehat{\theta}^*),\ \widehat{q}_{1-\alpha/2}(\widehat{\theta}^*)\right],$$

$$CI_{\text{basic}} = \left[2\widehat{\theta} - \widehat{q}_{1-\alpha/2}(\widehat{\theta}^*),\ 2\widehat{\theta} - \widehat{q}_{\alpha/2}(\widehat{\theta}^*)\right],$$

$$T^{*(b)} = \frac{\widehat{\theta}^{*(b)} - \widehat{\theta}}{\widehat{\text{se}}^{*(b)}} \quad \text{for studentized inference.}$$

Strengths and limits of bootstrap

Strengths

- flexible;
- finite-sample oriented;
- useful for complicated estimators.

Limits

- depends on design choices;
- computationally heavier;
- can fail under model misspecification.

Bottom line

Bootstrap is most convincing when the pseudo-data generator reflects the real source of dependence rather than merely looking computationally convenient.

Healthy workflow

When possible, compare two plausible bootstrap designs. Agreement across them is often more reassuring than a single beautifully computed bootstrap p-value.

Transition from bootstrap to GMM

Bootstrap approximates a sampling distribution. GMM starts from economic theory.

New question

What if the model implies valid population moment conditions, even though a full likelihood is difficult or undesirable to specify?

- Then we can estimate the parameter by making sample moments close to zero.
- Dependence still matters, because the covariance of the sample moments is again a long-run variance problem.

$$g(\theta_0) = \mathbb{E}[h(\theta_0, w_t)] = 0$$

replaces a full likelihood as the core identifying object.

Bridge between the two topics

Bootstrap asks how to approximate the distribution of a statistic once we know what to estimate. GMM asks what to estimate once theory gives us moments instead of a full density.

Method of moments is already familiar

Suppose a sample $y_1, \ldots, y_T$ comes from a population with unknown mean μ and variance σ^2.

$$\mathbb{E}(y_t - \mu) = 0, \qquad \mathbb{E}\left[(y_t - \mu)^2 - \sigma^2\right] = 0.$$

The sample analogues are

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \mu) = 0, \qquad \frac{1}{T}\sum_{t=1}^{T}\left[(y_t - \mu)^2 - \sigma^2\right] = 0.$$

Solving gives

$$\widehat{\mu} = \bar{y}, \qquad \widehat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2.$$

Why this matters

We have been using method of moments all along. The familiar case is exactly identified: the number of moment conditions equals the number of unknown parameters.

Why generalized method of moments?

Let

$$g_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} h(\theta, w_t) \in \mathbb{R}^q, \qquad \theta \in \mathbb{R}^p.$$

If $q > p$, one parameter vector cannot generally force every sample moment to equal zero exactly.

- **IV and regression:** $\mathbb{E}[z_t(y_t - x_t'\beta)] = 0$ with many instruments z_t.
- **Asset pricing:** Euler-equation restrictions for several assets and several instruments.
- **Time-series fitting:** match mean, variance, and several autocovariances with only a few structural parameters.

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} g_T(\theta)' W_T g_T(\theta).$$

Core motivation

GMM is useful because it generalizes ordinary method of moments: it lets theory give us many valid restrictions and lets the data weight noisy or correlated moments sensibly.

Moment conditions and identification

Suppose theory implies

$$\mathbb{E}[h(\theta_0, w_t)] = 0.$$
$$h(\theta, w_t) \in \mathbb{R}^q, \qquad \theta \in \mathbb{R}^p.$$

- $h(\theta, w_t)$ is a vector of moment functions.
- The true parameter θ_0 makes those moments vanish in population.
- Identification requires that no other parameter value does the same thing.

$$\operatorname{rank}(G) = p, \qquad G = \mathbb{E}\left[\frac{\partial h(\theta_0, w_t)}{\partial \theta'}\right],$$

is the standard local-identification condition in the differentiable case.

Exactly identified versus overidentified

- If $q = p$, the system is exactly identified.
- If $q > p$, the system is overidentified.
- Overidentification creates a goodness-of-fit question: can one parameter vector make all sample moments approximately zero at once?

This is where the J-test enters

It evaluates whether the extra moment restrictions are jointly compatible with the data.

Economic reading

Exact identification uses moments only to pin down parameters. Overidentification asks a stronger question: does the model fit several theoretically implied relationships at once?

$$q = p \Rightarrow g_T(\widehat{\theta}) \approx 0 \quad \text{by construction,} \qquad q > p \Rightarrow g_T(\widehat{\theta}) \neq 0 \text{ in general.}$$

Sample moment vector

The sample analogue is

$$g_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} h(\theta, w_t).$$

- If $\theta = \theta_0$, then $g_T(\theta)$ should be close to zero in large samples.
- Sampling noise and dependence mean it will not be exactly zero.
- GMM chooses the parameter that makes $g_T(\theta)$ as small as possible in a weighted quadratic sense.

Geometric picture

GMM projects the sample-moment vector as close to the origin as possible, but in a metric determined by the weight matrix.

$$\sqrt{T}\, g_T(\theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} h(\theta_0, w_t)$$

is the object whose CLT drives the entire method.

The GMM criterion function

The estimator is

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} g_T(\theta)' W_T g_T(\theta),$$

where W_T is a positive-definite weight matrix.

Analogy with OLS

OLS minimizes squared residuals. GMM minimizes a weighted sum of squared sample moments.

$$Q_T(\theta) = g_T(\theta)' W_T g_T(\theta)$$

is the sample objective function that the optimizer sees.

Metric interpretation

Changing W_T changes the notion of distance from the origin in moment space. A noisy moment direction should count less than a precise one.

Why weighting matters

Different moments need not be equally informative.

- Some moments are noisy.
- Some moments are highly correlated with others.
- Efficient weighting discounts unstable moments and accounts for cross-moment covariance.

Key point

The weight matrix is not an arbitrary numerical trick. It encodes the second-order structure of the moment process.

$$\operatorname{Var}\left(\sqrt{T}\, g_T(\theta_0)\right) = \Omega.$$

Practical benchmark

The identity matrix $W = I$ is fine for a first pass, but efficient GMM needs a weight matrix that reflects Ω, not a numerically convenient placeholder.

Long-run variance of the moments

Under dependence,

$$\sqrt{T}\, g_T(\theta_0) \Longrightarrow N(0, \Omega),$$

where

$$\Omega = \sum_{j=-\infty}^{\infty} \Gamma(j), \qquad \Gamma(j) = \mathrm{Cov}(h_t, h_{t-j}).$$

- This is another long-run variance problem.
- The GMM chapter therefore inherits the entire HAC discussion from earlier lectures.

$$\widehat{\Omega}_{\mathrm{HAC}} = \sum_{j=-(T-1)}^{T-1} k(j/m)\widehat{\Gamma}(j)$$

is the workhorse estimator in practice.

Why the optimal weight is Ω^{-1}

The efficient GMM weighting matrix is

$$W_{\text{opt}} = \Omega^{-1}.$$

- High-variance moments receive less weight.
- Correlated moments are de-duplicated appropriately.
- Efficient GMM is impossible without a credible estimate of the long-run covariance of the moments.

$$\text{asymptotic variance of } \widehat{\theta}_{\text{eff}} = (G'\Omega^{-1}G)^{-1}.$$

Efficiency intuition

Moments that are highly persistent or highly collinear get down-weighted because they contribute little new information once their long-run covariance is accounted for.

Two-step GMM

$$\widehat{\theta}^{(1)} = \arg\min_{\theta} \; g_T(\theta)' I g_T(\theta)$$

$$\widehat{\Omega}\left(\widehat{\theta}^{(1)}\right) = \sum_{j=-(T-1)}^{T-1} k(j/m)\,\widehat{\Gamma}_j\left(\widehat{\theta}^{(1)}\right)$$

$$\widehat{\theta}^{(2)} = \arg\min_{\theta} \; g_T(\theta)' \widehat{\Omega}\left(\widehat{\theta}^{(1)}\right)^{-1} g_T(\theta).$$

Interpretation

The first step gets consistency. The second step buys efficiency.

Operational reality

In most software, the second step is where all the earlier HAC choices return through the estimate of Ω.

Why the first step still matters

If the first step is numerically poor, the HAC estimate of Ω can also be unstable because it is evaluated at a poor preliminary parameter value.

Continuously updated GMM

Instead of separating estimation into two steps, one can update the weight matrix during optimization.

- This is called continuously updated GMM.
- It can improve some finite-sample properties.
- It is also more numerically delicate.

Pedagogical focus

For this course, two-step GMM is the central workhorse.

$$\widehat{\theta}_{\mathrm{CUE}} = \arg\min_{\theta} \; g_T(\theta)' \widehat{\Omega}(\theta)^{-1} g_T(\theta).$$

Why CUE can be harder

Every candidate parameter now changes both the moments and the metric used to judge them, so the objective surface can become much more irregular.

Asymptotic distribution of GMM

Let

$$G = \mathbb{E}\left[\frac{\partial h(\theta_0, w_t)}{\partial \theta'}\right].$$

Then

$$\sqrt{T}(\widehat{\theta} - \theta_0) \Longrightarrow N\left(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}\right).$$

- Under the optimal weight $W = \Omega^{-1}$, the expression simplifies.
- The entire covariance formula depends on the long-run covariance of the moments.

$$\sqrt{T}(\widehat{\theta} - \theta_0) \Longrightarrow N\left(0, (G'\Omega^{-1}G)^{-1}\right)$$

under efficient weighting.

J-test logic

When there are more moments than parameters, define

$$J_T = T\, g_T(\widehat{\theta})' \widehat{W} g_T(\widehat{\theta}).$$

- Under correct specification, J_T is asymptotically chi-squared with $q - p$ degrees of freedom.
- Large J_T means the moments cannot all be jointly matched well.

$$p\text{-value} = \mathbb{P}\left(\chi^2_{q-p} \geq J_T\right)$$

is the familiar large-sample reporting device.

Reading a J-test carefully

- A rejection often signals model misspecification.
- It can also reflect poor instruments or weak finite-sample performance.
- A non-rejection does not prove the model is true.

Correct reading

The J-test evaluates the joint compatibility of the chosen theory, the chosen instruments, and the chosen weighting scheme with the data.

Empirical attitude

Use the J-test as a diagnostic for tension between theory and data, not as a stamp of model truth.

Finite-sample caution

Because the J-test inherits HAC estimation and instrument choice, a dramatic rejection is informative, but a borderline non-rejection should not be over-read as strong support for the model.

Key references for GMM and the C-CAPM

- **Course text:** Chapter 6, "Generalized Method of Moments" and "Application to the C-CAPM."
- **Hansen (1982):** large-sample properties of GMM estimators.
- **Hansen and Singleton (1982):** direct estimation and testing from stochastic Euler equations.
- **Hansen and Singleton (1983):** restricted time-series implications for consumption and returns.
- **Hansen and Jagannathan (1997):** pricing-error diagnostics and distance-based model comparison.

Reading strategy

Use the textbook for notation and workflow, then use the classic papers to see how GMM entered empirical asset pricing.

Why the C-CAPM is a good teaching illustration

The consumption-based CAPM is useful pedagogically because:

- the theory directly delivers Euler-equation moments;
- the estimator is nonlinear but still manageable;
- the J-test has a clear interpretation;
- empirical difficulties connect to major themes in finance, such as the equity premium puzzle.

Pedagogical bonus

It shows students how nonlinear asset-pricing theory becomes a concrete econometric moment system.

$$g_T(\beta, \gamma) = \frac{1}{T} \sum_{t=1}^{T} \left[1 - \beta(1 + R_{i,t+1}) \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \right] x_t$$

makes the theory-to-estimation mapping explicit.

Euler equation intuition

With CRRA utility, the stochastic discount factor is

$$M_{t+1}(\theta) = \beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma}.$$

The Euler equation requires

$$\mathbb{E}_t[M_{t+1}(\theta)(1 + R_{i,t+1}) - 1] = 0.$$

Economic meaning

Assets with low payoffs in bad consumption states must offer high average returns.

$$\text{bad states} \Rightarrow M_{t+1} \text{ high} \Rightarrow \text{valuable payoffs}$$

is the intuition behind the pricing restriction.

$$\mathbb{E}(R_{i,t+1} - R_{f,t+1}) = -\frac{\mathrm{Cov}(M_{t+1}, R_{i,t+1})}{\mathbb{E}(M_{t+1})}$$

is the covariance-pricing reading of the same Euler equation.

Unconditional moments with instruments

Multiply the Euler equation by instruments $x_t \subseteq \mathcal{F}_t$:

$$\mathbb{E}\left[\left\{1-\beta(1+R_{i,t+1})\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\right\}x_t\right]=0.$$

- These are GMM moment conditions.
- The instruments translate conditional restrictions into unconditional ones.

$$h_t(\theta)=\left[1-\beta(1+R_{i,t+1})\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\right]x_t.$$

$$u_{i,t+1}(\theta)=1-M_{t+1}(\theta)(1+R_{i,t+1}),\qquad h_t(\theta)=u_{t+1}(\theta)\otimes x_t$$

if we stack assets and instruments into one moment vector. Here $\otimes$ is the tensor-product notation for "all asset-instrument products."

Why instruments are needed

They let one conditional restriction generate multiple unconditional moments, which is what makes GMM empirically testable and overidentified in practice.

What does the tensor product mean here?

In this finite-dimensional setting, the tensor product is just the Kronecker-product way to list all pairwise products between asset pricing errors and instruments.

$$u_{t+1}(\theta) = \begin{bmatrix} u_{1,t+1}(\theta) \\ \vdots \\ u_{N,t+1}(\theta) \end{bmatrix}, \qquad x_t = \begin{bmatrix} x_{1,t} \\ \vdots \\ x_{M,t} \end{bmatrix} \quad \Longrightarrow \quad u_{t+1}(\theta) \otimes x_t = \begin{bmatrix} u_{1,t+1}x_{1,t} \\ \vdots \\ u_{1,t+1}x_{M,t} \\ u_{2,t+1}x_{1,t} \\ \vdots \\ u_{N,t+1}x_{M,t} \end{bmatrix}.$$

$$\text{If } u_{t+1} = (u_1, u_2)' \text{ and } x_t = (1, z_t)', \quad u_{t+1} \otimes x_t = (u_1, u_1 z_t, u_2, u_2 z_t)'.$$

Why this notation is useful

It shows immediately that N assets and M instruments produce NM moment conditions. Nothing mysterious is happening: we are simply stacking every pricing error multiplied by every instrument.

Instrument choice and economic meaning

Good instruments should be:

- known at time t;
- relevant for variation in the Euler equation error;
- not so numerous that they destabilize finite-sample estimation.

Practical warning

Too many weak instruments can make GMM look sophisticated while reducing reliability.

Typical examples

Constants, lagged returns, lagged consumption growth, and other variables in the investor's information set.

Connection to modern practice

This is the same basic concern as in IV settings: instrument proliferation can overfit the sample moments and make diagnostics less informative.

GMM workflow for the C-CAPM

1. construct consumption growth and asset return data;
2. define the moment function in (β, γ);
3. choose instruments x_t;
4. run first-step and second-step GMM;
5. inspect coefficient estimates, standard errors, and the J-test.

Checklist

At the end, ask three separate questions: are the parameters plausible, do the standard errors look stable, and do the moments pass the overidentification diagnostic?

$$g_T(\beta, \gamma) = \frac{1}{T} \sum_{t=1}^{T} h_t(\beta, \gamma), \qquad \widehat{W} = \widehat{\Omega}^{-1}$$

is the practical object passed to the optimizer.

What to compare across specifications

Change instruments, weighting details, or sample endpoints one at a time. If the economic conclusion only survives one fragile implementation, the theory has not really won the empirical argument.

Hansen and Singleton (1982): why it was a breakthrough

The 1982 paper showed that nonlinear rational expectations models can be estimated directly from stochastic Euler equations:

$$\mathbb{E}[h_t(\theta_0) \mid \mathcal{F}_t] = 0 \quad \Rightarrow \quad \mathbb{E}[h_t(\theta_0) x_t] = 0.$$

For the C-CAPM,

$$h_t(\theta) = 1 - \beta(1 + R_{i,t+1}) \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma}.$$

- Preference parameters become estimable instead of purely calibrated.
- Instruments turn conditional restrictions into an over-identified system.
- The model can be tested, not merely matched informally.

Big idea

Estimate the objective-function parameters of economic agents without solving the full stochastic equilibrium explicitly.

Hansen and Singleton (1983): restricted time-series implications

Write the log stochastic discount factor as

$$m_{t+1} = \log M_{t+1} = \log \beta - \gamma \Delta c_{t+1}, \qquad \Delta c_{t+1} = \log C_{t+1} - \log C_t.$$

Under additional distributional restrictions, the Euler equation implies a restricted representation linking consumption growth and asset returns.

- Preference parameters govern how strongly returns co-move with consumption growth.
- The model generates testable restrictions on means, variances, and covariances.
- Postwar data can then be used to estimate preference parameters and test the implied restrictions.

Interpretation

The 1983 paper moved the literature from "here is an Euler equation" to "here is a full time-series system that returns and consumption should satisfy jointly."

Hansen and Singleton: empirical design in practice

Their empirical implementations used:

- consumption proxied by nondurable and services expenditure;
- test assets such as market indices and industry portfolios;
- instruments such as constants, lagged consumption growth, and lagged returns;
- two-step reweighting based on the covariance structure of the moment process.

$$u_{i,t+1}(\theta) = 1 - M_{t+1}(\theta)(1 + R_{i,t+1}), \qquad h_t(\theta) = u_{t+1}(\theta) \otimes x_t.$$

$$J_T = T\, g_T(\widehat{\theta})' \widehat{W} g_T(\widehat{\theta})$$

then asks whether the whole moment system is jointly credible.

What the Hansen-Singleton evidence taught the literature

The early GMM estimates often delivered

$$\widehat{\beta} \approx 0.99$$

and, in some specifications, economically plausible values of γ.

- Conditioning information mattered a great deal.
- Plausible parameter estimates did not guarantee that all moments fit well.
- Over-identifying restrictions were often fragile or rejected.

Historical lesson

GMM made structural asset-pricing estimation feasible and, at the same time, exposed the empirical limits of the standard C-CAPM.

Legacy

This is one of the roads leading to later work on Euler equation errors, the equity premium puzzle, and Hansen–Jagannathan distance diagnostics.

Reading the output: the equity premium puzzle

Empirically, standard C-CAPM estimates often struggle:

- plausible γ values frequently fail to explain observed excess returns;
- the J-test often rejects the overidentifying restrictions;
- this points to model misspecification, not merely computational inconvenience.

Big picture

GMM is powerful because it lets data challenge theory through moment restrictions.

Interpretation

When the model needs implausibly high risk aversion or still fails the J-test, that is evidence against the model, not against the GMM method.

Euler equation errors and the HJ distance

Once a candidate stochastic discount factor leaves pricing errors,

$$g(\theta) = \mathbb{E}[M_{t+1}(\theta)\mathrm{R}_{t+1} - 1_N] \neq 0,$$

we want to measure how serious that misspecification is.
Hansen and Jagannathan's diagnostic is

$$d_{HJ}(\theta) = \sqrt{g(\theta)' U^{-1} g(\theta)}, \qquad U = \mathbb{E}\left[\mathrm{R}_{t+1}\mathrm{R}_{t+1}'\right].$$

Its sample version is

$$\widehat{d}_{HJ} = \min_{\theta} \sqrt{g_T(\theta)' G_T^{-1} g_T(\theta)},$$

where $g_T(\theta) = T^{-1}\sum_{t=1}^{T}[M_{t+1}(\theta)\mathrm{R}_{t+1} - 1_N]$ and $G_T = T^{-1}\sum_{t=1}^{T}\mathrm{R}_{t+1}\mathrm{R}_{t+1}'$.

Why people use it

If $d_{HJ} = 0$, the model prices the test assets exactly. If it is positive, the model has an irreducible pricing error. Hansen and Jagannathan emphasized that this kind of comparison does *not* reward a model merely for making its SDF proxy more volatile.

How should we read the HJ distance?

The same object has an economically intuitive portfolio representation:

$$d_{HJ}^2(\theta) = \max_{a \neq 0} \frac{(a' g(\theta))^2}{a' U a}.$$

- It looks for the **most mispriced portfolio** among linear combinations of the test assets.
- The denominator $a' U a$ rescales by the portfolio's second moment, so the diagnostic is not driven mechanically by raw volatility.
- Smaller is better; zero means exact pricing on the asset span under study.

Relation to the J-test

The J-test asks whether exact validity can be statistically rejected under a chosen weighting matrix. The HJ distance asks how large the economically meaningful misspecification is, even when we already know the model is only approximate.

Bootstrap and GMM together

These topics are not separate islands.

- Bootstrap can improve finite-sample inference for GMM estimators.
- GMM itself still needs robust covariance estimation for the moments.
- Both topics remind us that dependence enters through the sampling behavior of sums and moments.

Practical use

In finite samples, a block bootstrap can be used to refine inference for GMM coefficients or the J-test when asymptotic approximations feel fragile.

Design principle

The bootstrap must resample the dependence structure in the moment conditions, not just the raw observations in a way that destroys the instruments or timing structure.

Why filtering comes next

Filtering might look like a new topic, but the bridge is natural.

- Prewhitening is already a filter.
- HAC kernels have frequency-domain interpretations.
- Many inferential problems improve after separating trend, cycle, and noise.

$$(1 - \phi L) y_t$$

is already a simple one-sided filter, so the move from prewhitening to filtering is smaller than it first appears.

Transition

Filtering is another way of asking how to isolate the relevant signal in dependent data.

Conceptual bridge

Inference asks for the right uncertainty about low-frequency movements. Filtering asks for the right transformation that exposes those movements in the first place.

Linear filters in the time domain

A linear filter transforms y_t into

$$x_t = \sum_{j=-\infty}^{\infty} a_j y_{t-j}.$$

$$x_t = A(L)y_t, \qquad A(L) = \sum_{j=-\infty}^{\infty} a_j L^j.$$

- The coefficients a_j determine which local patterns are kept or attenuated.
- A moving average smooths.
- Differencing removes low-frequency trend components.

$$\sum_j a_j = 1$$

is typical for smoothing filters that preserve a constant signal.

Frequency-domain view and transfer function

The same filter can be described in the frequency domain by its transfer function

$$A(\lambda) = \sum_{j=-\infty}^{\infty} a_j e^{-ij\lambda}.$$

- $|A(\lambda)|$ tells us which frequencies are passed or suppressed.
- Low-pass filters preserve slow movements.
- High-pass filters preserve rapid movements.
- Band-pass filters isolate intermediate cycles.

$$\text{gain} = |A(\lambda)|, \qquad \text{phase} = \arg A(\lambda).$$

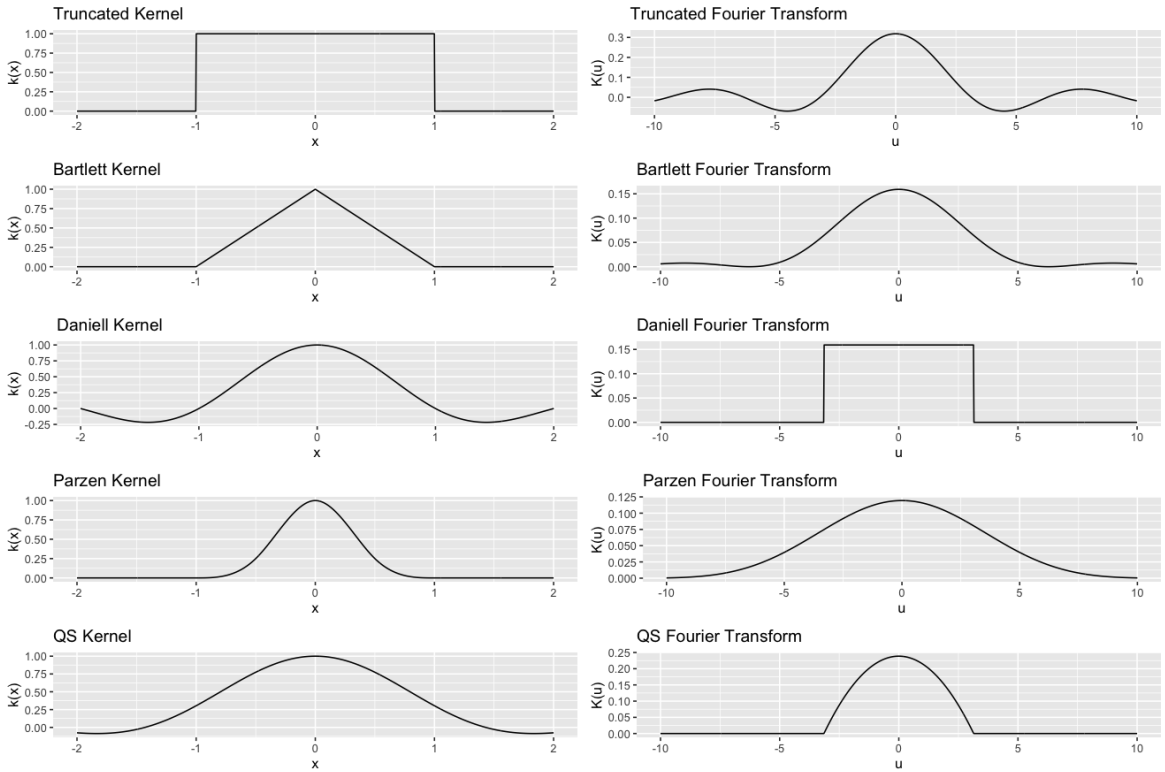
$$f_x(\lambda) = |A(\lambda)|^2 f_y(\lambda)$$

links the output spectrum directly to the input spectrum.

Interpretation by frequency

If $A(0)$ is large, very low frequencies survive. If $|A(\lambda)|$ is near zero for small λ, the filter suppresses trend-like movements.

Kernels and Fourier transforms



How this connects back to HAC

The same kernel ideas that appeared in long-run variance estimation also appear as frequency-response objects. This is why smoothing, spectral estimation, and filtering are so tightly connected.

Common filter types

- **Low-pass filters:** retain trend-like low frequencies.
- **High-pass filters:** remove smooth low-frequency components.
- **Band-pass filters:** isolate business-cycle frequencies.
- **One-sided filters:** use only current and past data.
- **Two-sided filters:** use past and future data and therefore are mainly offline tools.

Economic translation

Trend extraction, seasonal adjustment, business-cycle isolation, and real-time nowcasting are all filtering problems in disguise.

Decision question

Before choosing a filter, ask which frequencies are signal for the research question and which frequencies are nuisance.

Notes-style interpretation

No filter is uniformly best. Trend, business-cycle, and nowcasting problems preserve different frequencies, so they usually need different filters.

Simple moving average filter

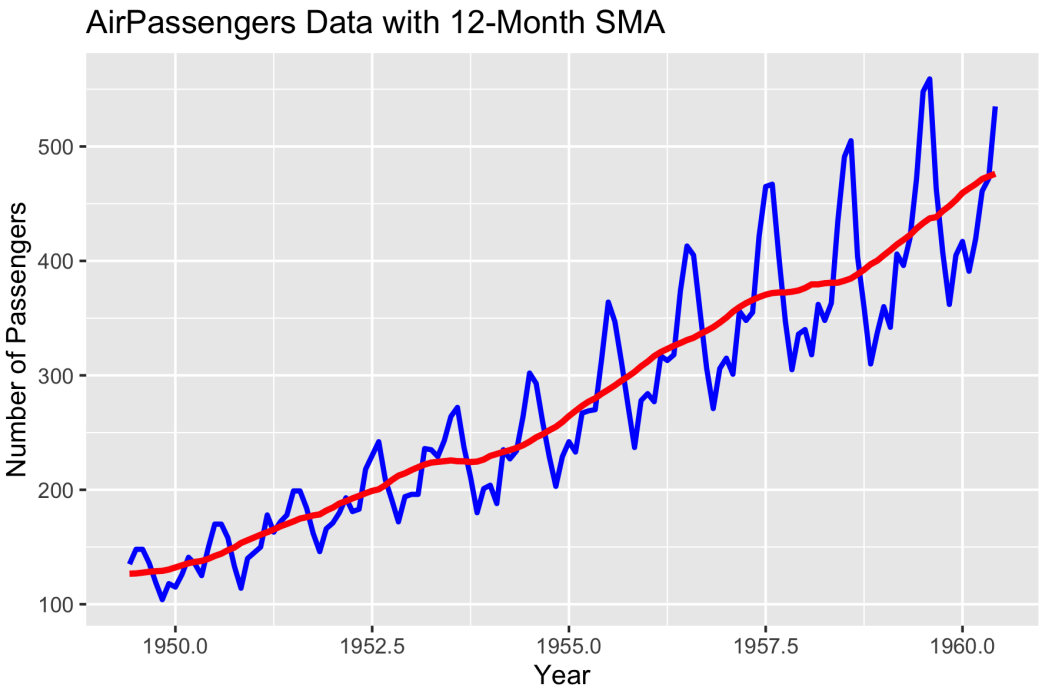
$$X_t = \frac{1}{2r+1} \sum_{j=-r}^{r} Y_{t-j}$$

$$B_{\mathrm{SMA}}(\lambda) = \frac{1}{2r+1} \frac{\sin((2r+1)\lambda/2)}{\sin(\lambda/2)}$$

$$B_{\mathrm{SMA}}(0) = 1, \qquad \text{phase} = 0.$$

Interpretation

This is an approximate low-pass filter: it damps high-frequency movement and leaves slower trend-like fluctuations more visible.



Differencing as a high-pass filter

$$\Delta Y_t = (1 - L) Y_t$$

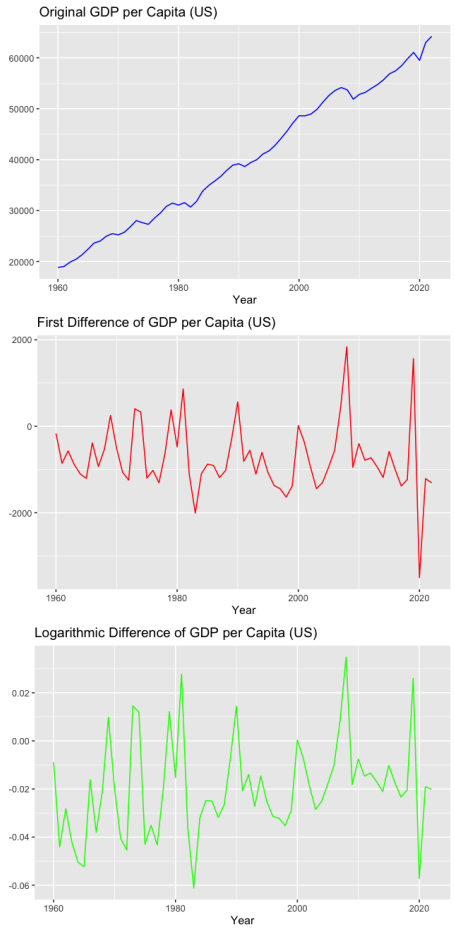
$$B_\Delta(\lambda) = 1 - e^{-i\lambda}$$

$$|B_\Delta(\lambda)|^2 = 2(1 - \cos\lambda)$$

$$|B_\Delta(0)|^2 = 0, \qquad |B_\Delta(\pi)|^2 = 4.$$

Interpretation

Differencing removes very low frequencies, so it naturally emphasizes short-run change rather than persistent level movement.



Baxter-King style business-cycle filter

$$\overline{\lambda} = \frac{2\pi}{P_l}, \qquad \underline{\lambda} = \frac{2\pi}{P_u}$$

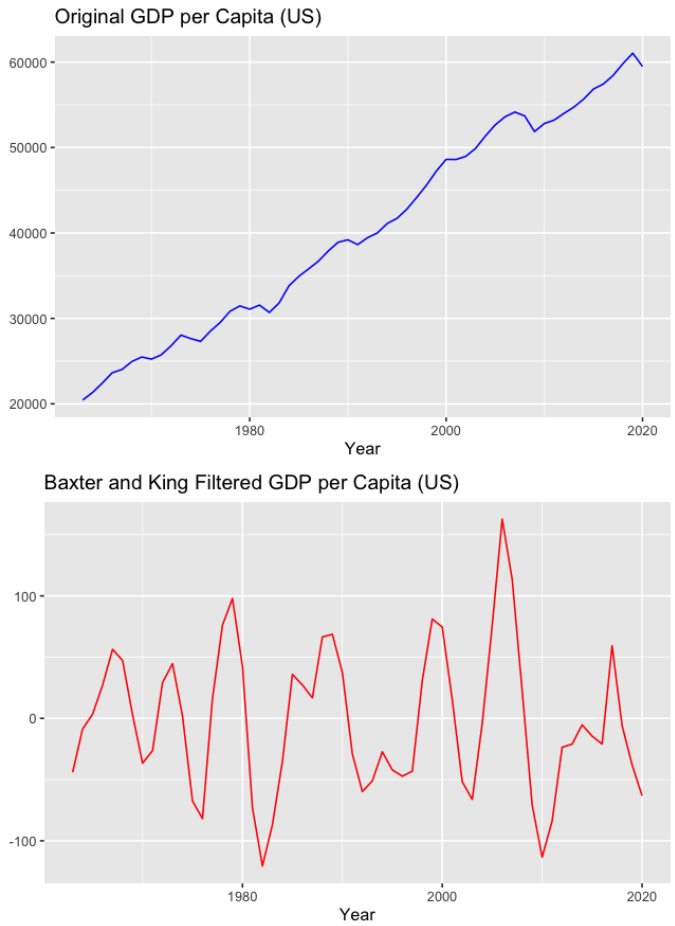
$$B_{\text{ideal}}(\lambda) = 1\{\underline{\lambda} \le |\lambda| \le \overline{\lambda}\}$$

$$\phi_0 = \frac{\overline{\lambda} - \underline{\lambda}}{\pi}, \qquad \phi_k = \frac{\sin(k\overline{\lambda}) - \sin(k\underline{\lambda})}{k\pi}$$

for $k = \pm 1, \ldots, \pm K$.

Interpretation

BK is a finite symmetric approximation to an ideal band-pass filter, so it keeps medium-horizon cycles and sacrifices K observations at each endpoint.



One-sided versus two-sided filters

One-sided

- usable in real time;
- causal;
- often noisier.

Two-sided

- smoother;
- uses future information;
- unsuitable for real-time forecasting.

Interpretive rule

If the empirical question is historical decomposition, two-sided filters are attractive. If the question is real-time policy use, causality usually forces a one-sided filter.

$$\text{one-sided: } x_t = \sum_{j=0}^{\infty} a_j y_{t-j}, \qquad \text{two-sided: } x_t = \sum_{j=-k}^{k} a_j y_{t-j}.$$

Endpoint problems and interpretation

- Two-sided filters behave poorly near the beginning and end of the sample.
- Filtered output depends on the choice of smoothing parameters and cut-off frequencies.
- Filtering can clarify structure, but it can also distort timing and amplitude if used mechanically.

Warning

The smoother the extracted signal looks, the more important it is to ask what timing information was sacrificed to obtain that smoothness.

Forward link

Lecture 16 turns filtering into a probabilistic problem by treating the signal as a latent state in a state-space model.

Lecture 15 takeaways

- Bootstrap extends robust inference by approximating sampling distributions directly.
- GMM estimates structural parameters through moment restrictions, but still depends on long-run covariance estimation.
- Filtering is the natural next step once we care about separating signal and noise in dependent data.

Big picture

Lecture 15 links three complementary ideas:

$$\text{resample the data} \Longrightarrow \text{estimate structural moments} \Longrightarrow \text{separate signal from noise.}$$

Unifying lesson

Dependence never disappears. It only changes form: as a resampling design problem, as a long-run covariance problem inside GMM, and as a frequency-selection problem inside filtering.

Preview of Lecture 16

Lecture 16 focuses on:

1. filtering in the time and frequency domains with fuller mathematical detail;
2. common deterministic filters, gain, phase, and endpoint issues;
3. the bridge from deterministic signal extraction to state-space modeling in Lecture 17.