Low-Frequency Analysis of Economic Time Series*

Ulrich K. Müller and Mark W. Watson
Department of Economics, Princeton University
Princeton, NJ, 08544

This Draft: October 29, 2019

*Draft chapter for *Handbook of Econometrics*, Volume 7, edited by S. Durlauf, L.P. Hansen, J.J. Heckman, and R. Matzkin. Müller acknowledges financial support from the National Science Foundation grant SES-1919336. An online appendix for this chapter is available at http://www.princeton.edu/~mwatson/.

1 Introduction

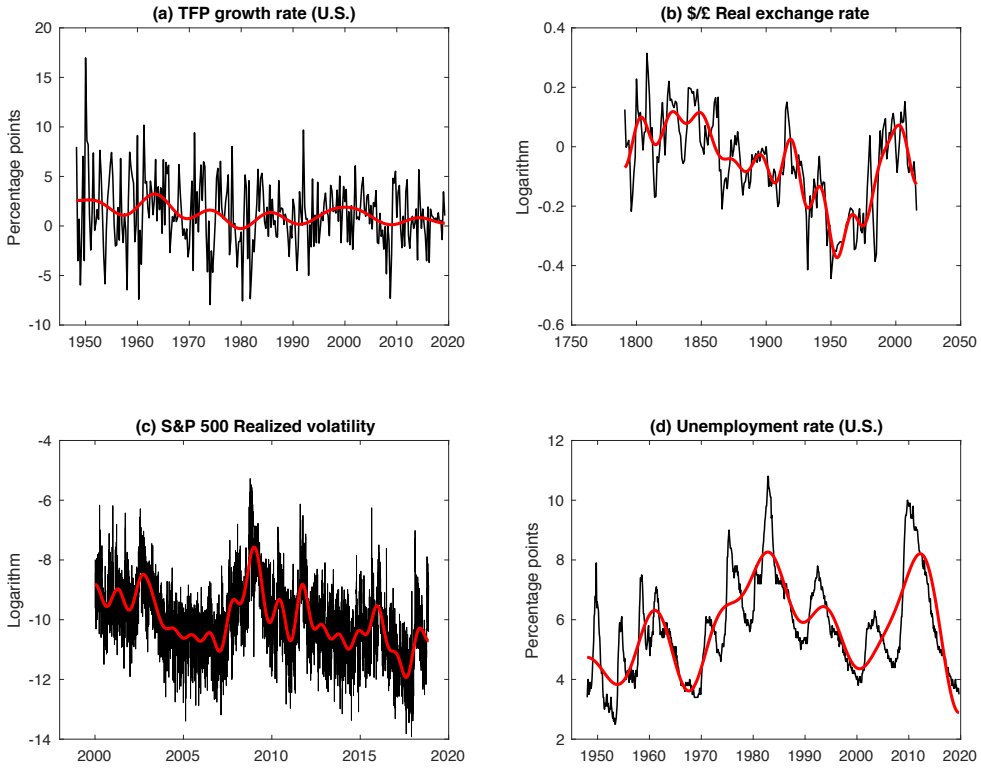
This chapter discusses econometric methods for studying *low-frequency* variation and covariation in economic time series. We use the term low-frequency for dynamics over time spans that are a non-negligible fraction of the sample period. For example, when studying 70 years of post-WWII quarterly data, decadal variation is low-frequency, and when studying a decade of daily return data, yearly variation is low-frequency. Much of this chapter is organized around a set of empirical exercises that feature questions about low-frequency variability and covariability, and there is no better way to introduce the topics to be covered than to look at the data featured in these exercises.

Figure 1 plots four univariate time series, the growth rate of total factor productivity in the United States (quarterly, from 1948-2019), the \$/£ real exchange rate (annually, from 1791-2016), realized volatility for returns on the S&P composite index (daily, from 2000-2018), and the U.S. unemployment rate (monthly, from 1948-2019).[1] Each figure shows the raw data and a low-frequency trend computed using methods described in Section 3; for now just note that these trends seem to capture low-frequency movements in the series. We will study three sets of questions involving low-frequency features of these univariate time series.

The first involves the *level* of the series. For a stationary process, the level is the mean, and the first empirical exercise involves inference about the mean growth rate of TFP allowing for the familiar $I(0)$ serial correlation patterns that underlie HAC standard errors. A related exercise involves inference about the mean of a highly serially correlated series like the unemployment rate, where standard $I(0)$ inference is misleading. Empirical researchers studying productivity have asked about changes in the level of the TFP growth rate, and this leads us to consider methods for inference about discrete breaks in the mean and a 'local-level' model that allows slowly changing variation in the level.

The second set of questions involves low-frequency *persistence*. For example, a large empirical literature has asked whether real exchange rates are covariance stationary (or $I(0)$) or have a unit root (are $I(1)$), or more generally sought to measure the half-life of shocks. We take up these questions. Another empirical literature has argued that asset return volatility exhibits long-memory, in the form a fractionally integrated $I(d)$ model. We provide methods for inference about the value of d. Returning to the local-level model, we

[1] The data and sources are described in the online appendix for this chapter.



Notes: Panel (a): Units: percentage points at an annual rate; Sample frequency/sample period: quarterly, 1948:Q2-2019:Q2 ($T = 285$); Low-frequency transforms: $q = 14$, shortest period = 41 quarters
Panel (b): Units: logarithm relative to value in 1900; Sample frequency/sample period: annual, 1791-2016 ($T = 226$); Low-frequency transforms: $q = 22$, shortest period = 20 years
Panel (c): Units: logarithm; Sample frequency/sample period: trading days, 1/3/2000 – 11/14/2018 ($T = 4738$); Low-frequency transforms: $q = 37$, shortest period = 256 trading days
Panel (d): Units: percentage points; Sample frequency/sample period: monthly, 1948:M1 – 2019:M9 ($T = 861$); Low-frequency transforms: $q = 14$, shortest period = 123 months

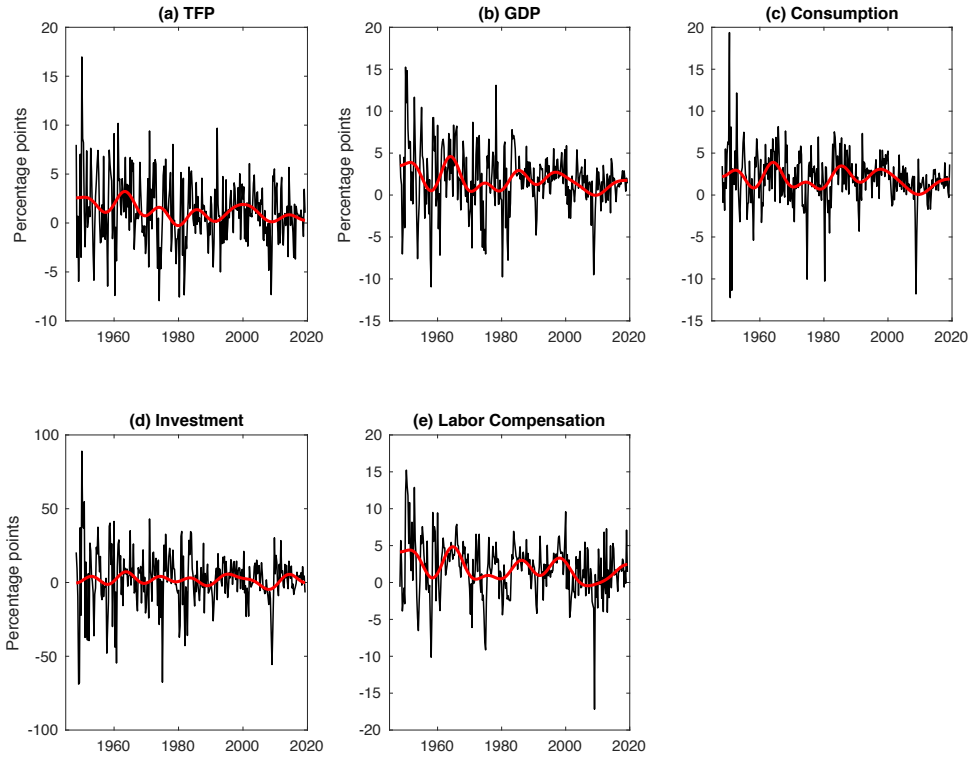
Figure 1: Four Economic Time Series

also conduct inference on the relative importance of permanent versus transitory shocks, another measure of persistence used in empirical work.

The third set of questions involves long-horizon predictions. How fast will TFP grow over the next 75 years and how certain are we about its future level? More precisely, based on a sample of T observations, we discuss Bayes methods for constructing the *predictive distribution* for the average value of a series over the next h periods, where h is of the same order as T, and for constructing analogous frequentist *prediction sets*.

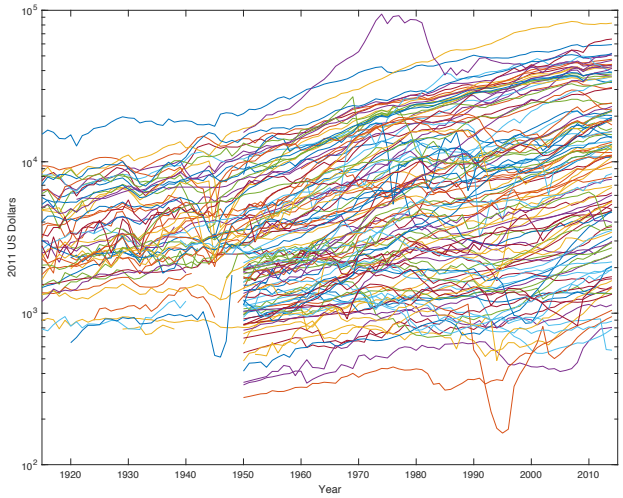
We also discuss methods to analyze the low-frequency features of multivariate times series. Figure 2 plots five times series, the growth rate of TFP together with growth rates of per-capita values of GDP, consumption, investment and labor compensation for the U.S. Neoclassical growth theory asserts a tight connection between these variables over the long run and we discuss econometric methods to evaluate these assertions. Specifically, we define low-frequency covariances, correlations, and linear regression coefficients, and illustrate inference about their values for the data in Figure 2 using a low-frequency linear factor model. Figure 3, taken from Müller et al. (2019), motivates a significantly more ambitious exercise: it plots per-capita values of real GDP for 112 countries over the last century. These data have three features not present in five-variable data set: the panel is unbalanced (data for some countries are unavailable for some years), the series are in log-levels rather than growth rates, and of course, there are many more series. Despite these complications, we discuss how the methods used to analyze the five-variable dataset can be generalized to study questions associated with low-frequency growth, development, and convergence in this 112-country dataset.

Inference about low-frequency features of a time series is inherently a small-sample statistical problem. For example, the quarterly macro data in Figure 2 span 71 years (or 285 quarters), but contain only seven decades and fewer than three 25-year periods. Thus, when interest focuses on variation over these long periods, the effective sample size is small. One way to overcome this small-sample problem is to use shorter-run or high-frequency variability in the sample to learn about low-frequency variability. For example, parametric time series models such as ARMA, VAR or VECM models use a common set of parameters to model autocovariances over both short and long horizons, and use both short-run and low-frequency variation in the sample to estimate the model parameters. While this is sensible if there are tight connections between low-frequency and higher-frequency variability, it leads



Notes: Units: percentage points at an annual rate; Sample frequency/sample period: quarterly, 1948:Q2-2019:Q2 (T = 285); Low-frequency transforms: q = 14, shortest period = 41 quarters

Figure 2: Growth rate of TFP and per-capita GDP, consumption, investment and labor compensation in the United States



Notes: This figure is taken from Müller, Stock and Watson (2019).

Figure 3: GDP per-capita for 112 countries

to potentially serious misspecification absent these connections. A more robust approach is to conduct inference about the low-frequency characteristics of a time series based solely on the low-frequency characteristics of the sample. This chapter describes inference using this latter approach. This requires small-sample statistical methods that efficiently utilize the limited low-frequency information in the sample. We discuss both Bayes and frequentist inference.

The general approach is straightforward to describe. Let x_t, $t = 1, ..., T$ denote a time series with T observations: for example, the raw data plotted in Figure 1-3. The low-frequency variation in these data is extracted using a small number of weighted averages collected in the vector $\mathbf{X}_T^0$. For example, one of these averages is the sample mean, another captures variability with period $2T$, then period T, period $2T/3$, and so forth, until the highest period of interest in included. The low-frequency trends plotted in Figures 1 and 2 are functions of $\mathbf{X}_T^0$. Low-frequency inference is based solely on $\mathbf{X}_T^0$; that is, the higher-frequency variability in the sample is treated as irrelevant and is therefore discarded. The vector $\mathbf{X}_T^0$ has two important properties. First, it contains only a few elements; that is, only a small number of low-frequency weighted averages are needed to summarize the low-frequency variability in the sample. This reflects the small-sample nature of the low-frequency inference problem. The second property of $\mathbf{X}_T^0$ is that in large-T samples, $\mathbf{X}_T^0 \overset{a}{\sim} \mathcal{N}(\mu_X, T^{-1}\mathbf{V}_X)$. This

central limit theorem result is, perhaps, not surprising for an $I(0)$ process, but it turns out to hold more generally for highly persistent covariance stationary processes, unit-root processes, and other persistent processes used to describe economic time series. Given this framework, low-frequency inference involves inference about a small sample – the elements in $\mathbf{X}_T^0$ – of jointly normally distributed random variables.

This chapter discusses low-frequency inference using parametric models. Specifically, we consider models that describe the low-frequency level, variability and covariability using a small number of parameters, say θ. One example is the $I(0)$ model, where θ includes the mean and long-run variance because these two parameters completely describe the low-frequency second moment properties of an $I(0)$ process. Other models include additional parameters to capture low-frequency persistence distinct from the $I(0)$ model. These models yield $\mathbf{X}_T^0 \overset{a}{\sim} \mathcal{N}(\mu_X(\theta), T^{-1}\mathbf{V}_X(\theta))$, with known functions linking θ to the mean and covariance matrix of $\mathbf{X}_T^0$. Inference about θ then proceeds using Bayes or frequentist methods, where the specifics depend on how θ affects μ_X and $\mathbf{V}_X$. For example, in the $I(0)$ model, the elements of $\mathbf{X}_T^0$ are independent and homoskedastic, $\mathbf{V}_X = \sigma^2\mathbf{I}$, and efficient low-frequency inference leads to familiar methods such as Student-t confidence intervals. In other models, μ_X and $\mathbf{V}_X$ are more complicated function of θ. Despite this, Bayes methods are easily implemented using modern computational methods. As a general matter, frequentist inference is relatively less straightforward, since the small sample problem often involves nuisance parameters which complicate the construction of efficient tests and confidence sets.

1.1 Reader's Guide

We emphasize at the outset that this chapter is not designed as a survey. We have strong opinions about the best methods for conducting inference about low-frequency characteristics of economic time series,[2] and the chapter focuses on these methods. That said, these methods build on the important contributions of a multitude researchers and we do our best to highlight these connections. In particular, we build on classic work on spectral analysis discussed in textbooks such as Brillinger (2001), Brockwell and Davis (1991), Hannan (1970), and Priestley (1981), and on the band-spectrum regression work in Engle (1974). The $I(0)$ analysis builds on HAC/HAR contributions in Domowitz and White (1982), Newey and

[2]See, in particular, Müller and Watson (2008, 2011, 2013, 2017, 2018).

West (1987), Andrews (1991), Kiefer et al. (2000), Jansson (2004), Kiefer and Vogelsang (2005), Sun et al. (2008) and others. The focus on low-frequency models uses insights from the unit-root literature exemplified by Dickey and Fuller (1979), Engle and Granger (1987), Johansen (1988), Stock (1994), Bierens (1997) and Phillips (1998), and the various bridges between $I(0)$ and $I(1)$ models including local-to-unity models (Cavanagh (1985), Chan and Wei (1987), Phillips (1987), and Stock (1991)), fractionally integrated models (e.g., Granger and Joyeux (1980), Geweke and Porter-Hudak (1983), and Robinson (2003)), and $I(0) + I(1)$ unobserved component models described in Harvey (1989).

Rather than providing a survey, the chapter embraces the definition of a *Handbook* as both an *instruction* and *reference* manual, and provides material for different types of readers. One group of readers is interested in how existing low-frequency methods can be applied to their specific empirical problems. Another is interested in the theory underlying the methods. A third is interested in developing inference and computational procedures that are related to, but distinct from the specific problems discussed in this chapter. Addressing the issues that might be of interest to each of these readers leads to a lot of material, only a subset of which might be of immediate interest to a particular reader. With this in mind, we provide a short reader's guide.

To begin, Sections 2 and 3 provide introductory material that should be read by everyone.

Section 2 introduces five models that parameterize a range of low-frequency variation and covariation patterns in economic time series. These include the class of $I(0)$ models that characterize the relatively short-lived dependence in standard covariance stationary ARMA models. Because of their limited dependence, a time series that follows an $I(0)$ process behaves like a white noise processes over the low frequencies studied in this chapter. Analogously, a time series whose *first differences* is $I(0)$ follows an $I(1)$ model and has the same second moment properties as a random walk over low frequencies. Three other parametric models generalize the $I(0)$ and $I(1)$ processes but in ways that lead to different forms of low-frequency persistence. The *local-to-unity* model generates covariance stationary data, but with autocorrelations, say $Cor(x_t, x_{t+j})$, that decay at rate $e^{-c(j/T)}$ where T is the sample size and $c > 0$ is the model's decay parameter. Covariance stationary *fractionally integrated* models have autocorrelations that decay more slowly, at the rate j^{2d-1}, where $d \in (-1/2, 1/2)$ is model's decay parameter. And finally, the *local-level* model generates a time series that is a linear combination of $I(0)$ and $I(1)$ processes, where a parameter

g governs the relative importance of the $I(1)$ components. In these models, low-frequency persistence depends on the value of the model-specific c, d, or g parameters. The value of this parameter is of primary interest in some applications; in others, it is important because persistence affects uncertainty about the mean of a time series, the long-horizon average its future values, or the low-frequency correlation between two times series.

Section 3 introduces $\mathbf{X}_T^0$, the low-frequency averages of the sample data that are used for low-frequency inference. One interpretation of $\mathbf{X}_T^0$ is as the OLS regression coefficients from the regression of the sample data onto low-frequency deterministic periodic functions of time – we use a constant and cosine functions. The fitted values from this regression are the low-frequency filtered versions of the data plotted in Figures 1 and 2. Importantly, only a small number of periodic functions are needed to capture the low-frequency properties of the series, so the dimension of $\mathbf{X}_T^0$ is small. Low-frequency inference is based on these these low-frequency summaries of sample data and exploits the approximate normal distribution of $\mathbf{X}_T^0$. Section 3 illustrates this by showing how $\mathbf{X}_T^0$ can be used for low-frequency inference in the $I(0)$ model.

Sections 4 and 5 provide ready-to-use Bayes and frequentist methods to answer specific low-frequency inference questions. Section 4 focuses on Bayes methods. It shows how to conduct inference about the persistence parameters in the models of Section 2, about the mean when persistence is local-to-unity, about discrete breaks in the mean of an otherwise $I(0)$ process, and about the realization of the low-frequency $I(1)$ trend in a local-level model. Section 4 also treats multivariate problems: Initially in the multivariate $I(0)$ model, and then, as an example of non-$I(0)$ inference, it develops a low-frequency factor model for the five variables plotted in Figure 2. Finally, it shows how Bayes methods can be used to construct predictive distributions for the average values of x_t over long-horizon out-of-sample periods. In each of these problems the posterior is readily computed using simple numerical methods requiring trivial amounts of modern computer time.

Section 5 takes up a subset of the same problems using frequentist methods. Specifically it considers tests of $I(1)$ persistence (that is, *unit-root tests*), tests of $I(0)$ persistence (sometimes called *stationarity tests*), construction of confidence intervals for the local-to-unity persistence parameter, and inference and prediction in the univariate and multivariate $I(0)$ models. As discussed in that section, these inference problems can be solved using standard methods (for the $I(0)$ model) or by straightforward calculations after eliminating nuisance

parameters using invariance restrictions. The section discusses a few other problems (low-frequency covariation in bivariate models and univariate long-horizon prediction intervals) where numerical methods have been developed for efficient inference in models beyond the $I(0)$ model.

Sections 6 and 7 discuss both the theory and numerical methods underlying the Bayes and frequentist methods used in Sections 4 and 5. These sections utilize the low-frequency problems introduced in Sections 4 and 5 as illustrative examples, but the principles and computational methods discussed in these sections have general applicability and are not specific to low-frequency inference. Section 6 focuses on Bayes methods, where the requisite theory is relatively straightforward and computational methods are part of a well-developed literature.

Section 7 discusses frequentist methods, with a focus on the problem of constructing powerful tests of competing hypothesis and the related problem of constructing efficient confidence intervals. The Neyman-Pearson (NP) Lemma is the bedrock for constructing powerful tests. But, because NP tests involve simple hypotheses, they cannot be directly applied in the many practical problems that involve composite hypotheses. Section 7 provides an in-depth discussion of this complication with topics that include weighted average power, least favorable distributions, invariance/equivariance restrictions, and efficiency results for confidence sets obtained by inverting these tests. Implementation of these efficient frequentist methods sometimes requires numerical approximations to least favorable distributions, and Section 7.5 outlines some associated algorithms.

Section 8 returns to the low-frequency analysis of economic times series and focuses on two issues. First, it shows how the limiting covariance matrix for $\mathbf{X}_T^0$ depends only on the spectrum of x_t in a local area near frequency zero, and extends this to pseudo-spectra for integrated processes. This analysis makes explicit the low-frequency nature of the analysis. The section's second contribution is a statement of the central theorem that serves as the basis for the normal likelihoods used throughout the chapter.

2 Five Low-Frequency Models

Five parametric models are widely used to describe low-frequency variation in economic time series. These models play a central role in the various examples in this chapter, and we

briefly review them here. Let x_t denote a univariate time series observed for $t = 1, \ldots, T$. Decompose x_t as

$$x_t = \mu + u_t \tag{1}$$

where μ is a constant and u_t is a zero-mean stochastic process. The five models differ in their assumptions about u_t.

2.1 A Digression on Scale Factors

We begin with a short digression on scale factors. Recall that when $u_t \sim i.i.d.(0, \sigma^2)$, the sample mean $\overline{u}_{1:T} = T^{-1} \sum_{t=1}^{T} u_t$ satisfies $Var(T^{1/2}\overline{u}_{1:T}) \sim O(1)$. In contrast, when u_t is a random walk with $\Delta u_t \sim i.i.d.(0, \sigma^2)$, $Var(T^{-1/2}\overline{u}_{1:T}) \sim O(1)$. Said differently, when u_t is *i.i.d.*, then the scale factor $\sqrt{T}$ stabilizes the variance of the sample average, but when u_t is a random walk, a scale of $1/\sqrt{T}$ is required. In the fractionally integrated models introduced below, the appropriate scale factors are yet different powers of T.

Keeping track of these different scale factors is a bookkeeping challenge and leads to cumbersome notation. Moreover, these factors don't ultimately affect inference in most of our applications because the methods we suggest lead to scale-invariant or equivariant inference. To avoid these complications, we apply a notational device by embedding the appropriate scale factor directly into the definition of each model to ensure that $Var[T^{1/2}\overline{u}_{1:T}] \sim O(1)$ throughout. For example, we define the random walk as a process with $\Delta u_t \sim i.i.d.\,(0, (\sigma/T)^2)$. As a formal matter, this means that the various stochastic processes for u_t depend on T, but we will suppress this in the notation to avoid clutter. The formal arguments in Section 8 incorporate this complication using the appropriate triangular arrays. With this background out of the way, we now define the models.

2.2 Five Models

A fundamental ingredient in each of the models is a covariance stationary $I(0)$ process that we denote by a_t. To be specific, as in Stock's chapter in Volume 4 of this *Handbook* (Stock (1994)), the $I(0)$ process a_t is defined as $a_t = c(\mathrm{L})\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}$ where ε_t is a stationary martingale difference process and where the moving average weights decay sufficiently rapidly so that $\sum_{j=0}^{\infty} j\,|c_j| < \infty$. The resulting so-called *long-run variance* of a_t is $\sigma^2 = \sigma_\varepsilon^2 c(1)^2 =$

$\sum_{j=-\infty}^{\infty} Cov(a_t, a_{t-j})$; the 'long-run' label here refers to the fact that σ^2 is the limiting variance of $T^{-1/2}\sum_{t=1}^{T} a_t$ and is proportional to the zero-frequency (i.e., infinite-period) spectrum of a_t. Each of the models introduced below can be described in terms of distributed lags of an $I(0)$ variable a_t, where the models differ in the decay pattern in their distributed lag weights.

In the $I(0)$ model, $u_t = a_t$.

In the $I(1)$ model, u_t is $I(1)$, that is $(1-\mathrm{L})u_t = (1/T)\, a_t$, where the term $(1/T)$ is the scale factor appropriate for this model. Recursive substitution yields $u_t = u_0 + T^{-1}\sum_{j=1}^{t} a_j$, so that $x_t = (\mu + u_0) + T^{-1}\sum_{j=1}^{t} a_j$. Because μ and u_0 have the same effect on the observations x_t for $t = 1, ..., T$, we impose $u_0 = 0$ as a normalization.

The final three models provide continuous bridges between these $I(0)$ and $I(1)$ models. The *local-to-unity* (LTU) model is the covariance-stationary model[3]

$$(1 - \rho_T \mathrm{L})u_t = (1/T)\, a_t \text{ where } \rho_T = 1 - c/T \text{ with } c > 0. \tag{2}$$

To appreciate the parameterization in (2) consider the model with $\rho_T = \rho$ and $|\rho| < 1$. In this fixed-coefficient model, $Cor(u_t, u_{t+j}) = \rho^j \to 0$ as $j \to \infty$ and u_t is an $I(0)$ process. Yet, when ρ is close to one and j is a non-negligible fraction of the sample size T, say $j = \lfloor sT \rfloor$ for $s > 0$, then $\rho^{\lfloor sT \rfloor}$ differs significantly from zero, and large-sample $I(0)$ approximations for the distribution of statistics computed from u_t are not accurate. More accurate large-sample approximations are obtained by the local-to-unity parameterization (2) where the value of c is held fixed as $T \to \infty$. In large samples, with c and $s > 0$ fixed, $(\rho_T)^{\lfloor sT \rfloor} \to e^{-sc}$. In this parameterization, large values of c produce $I(0)$ low-frequency dynamics and moderate positive values of c capture persistence patterns between $I(1)$ and $I(0)$.[4]

Another $I(0)/I(1)$ bridge is the $I(d)$ or *fractional* (FR) model with

$$(1 - \mathrm{L})^d u_t = \left(T^{-d}\right) a_t \tag{3}$$

where the parameter d is allowed to take on non-integer fractional values.[5] Fractional models

[3]Early references for the LTU model include Bobkoski (1983), Cavanagh (1985), Chan and Wei (1987), Phillips (1987), and Stock (1991).

[4]The stationarity in our definition of the LTU model means that the model itself has no well defined limit as $c \to 0$. Nevertheless, location invariant low-frequency statistics of a LTU process have a distribution that is continuous in $c \to 0$ (cf. Elliott (1999), and Section 8.1).

[5]For fractional values of d, the lag polynomial uses the binomial expansion $(1 - \mathrm{L})^d = \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)} \mathrm{L}^j$.

with $-1/2 < d < 1/2$ are covariance stationary and invertible with autocovariances γ_j that decay hyperbolically at the rate j^{2d-1} for $d \neq 0$; because of this slow decay, these processes are said to exhibit *long memory*.[6] We will consider values of d between $-1/2$ and $3/2$, where for $1/2 < d < 3/2$, $u_t = T^{-1}\sum_{j=1}^{t} v_j$ with $u_0 = 0$ and $v_t \sim I(d-1)$. It turns out that with this choice, scale and location invariance low-frequency statistics of the fractional model have a distribution that is continuous in $d \in (-1/2, 3/2)$ (see Müller and Watson (2008), and Sections 4.1.3 and 8.1 below).

The final $I(0)/I(1)$ bridge is the *local-level* (LL) model that expresses u_t as the sum of uncorrelated $I(0)$ and $I(1)$ processes

$$u_t = b_t + e_t \tag{4}$$

where $b_t \sim I(1)$ and $e_t \sim I(0)$ with $b_0 = 0$. This model is usefully parameterized as

$$b_t = (g/T)\sum_{j=1}^{t} a_j \tag{5}$$

where e_t and a_t follow uncorrelated $I(0)$ processes with common long-run variance σ^2. In this parameterization, the long-run standard deviation of Δb_t is $\sigma_{\Delta b} = \sigma g/T$. The name local-level model reflects that in (4), $\mu + b_t$ serves as the 'local-in-time level' of x_t.[7]

3 Low-Frequency Trends and Averages

This section has two primary goals: the first is to describe the low-frequency weighted averages of the data that underlie the low-frequency trends plotted in Figures 1 and 2, and the second is to introduce normal approximations for the probability distribution of these weighted averages based on central limit results formally developed in Section 8. As discussed in the introduction, these low-frequency averages are the small-dimensional data summaries that form the basis for low-frequency inference. The central limit results rationalize low-frequency inference based on Gaussian likelihoods. With these concepts in hand, the section

[6] See Baillie (1996) and Robinson (2003) for surveys of the early work on fractional long-memory models.

[7] The local-level model is an example of an unobserved-component-ARIMA model which has a long history: see Chapter 1 of Nerlove et al. (1979) for a historical survey and Harvey (1989) for the classic textbook development.

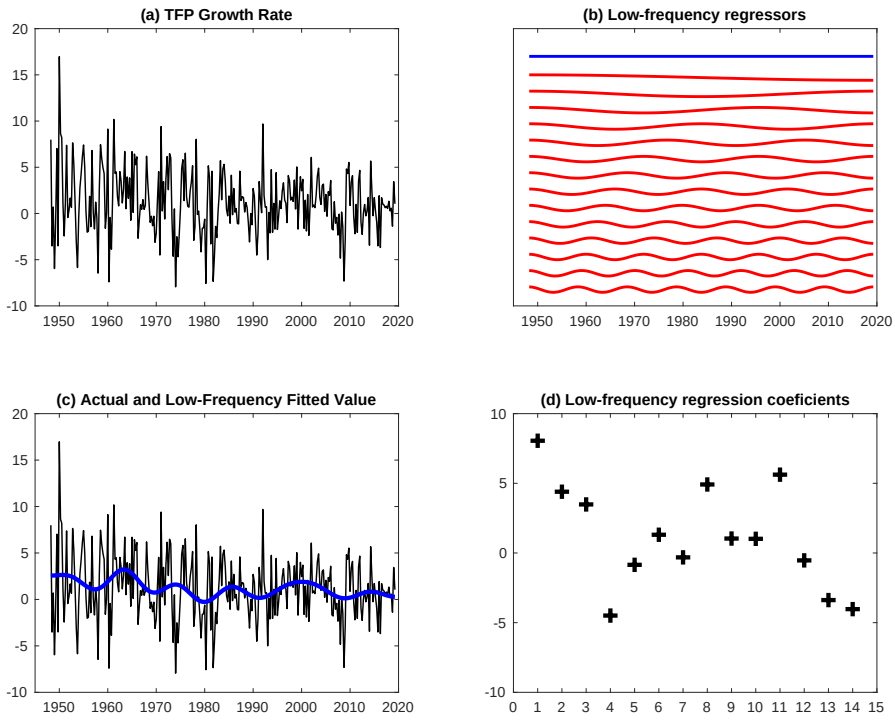


Figure 4: Constructing low-frequency transforms (trends)

then develops autocorrelation-robust inference procedures for the mean of an $I(0)$ process and discusses generalizations for $I(0)$ linear regression and IV models. Finally, the section offers some conceptual comments and practical advice on the appropriate choice of the low-frequency band used for empirical applications.

3.1 Low-Frequency Projections

Figure 4 summarizes the construction of the low-frequency trends and averages, using the data on TFP growth rates as an example. Let x_t denote the raw data; for TFP these are the growth rates plotted in panel (a). Sample data are available for $t = 1, 2, \ldots, T$; in the example, the data are available from 1948:Q2 through 2019:Q2, so that $T = 285$ quarters.

The low-frequency trends are computed as the fitted value from a regression of x_t onto the set of low-frequency functions shown in panel (b) of Figure 4. The first of these is the constant function, and the other are cosine functions. The first cosine function has a period of

$2T$, the second has period T, and the j-th has period $2T/j$. There are $q = 14$ cosine functions plotted in the figure, so the shortest period is $2T/14 \approx 41$ quarters (or slightly more than 10 years). The low-frequency fitted values are plotted in panel (c). Panel (d) shows the $q = 14$ OLS regression coefficients corresponding to each of the cosine-function regressors in panel (b). The OLS coefficients fully summarize the low-frequency fitted values.[8] In this example, there are $q+1 = 15$ regression coefficients including the constant, and low-frequency analysis of TFP growth rates is based on this small sample of 15 observations.

Some additional notation summarizes the calculations represented in Figure 4. Let $\Psi_j(s) = \sqrt{2}\cos(js\pi)$ denote a cosine function on $s \in [0,1]$ with period $2/j$ (where the factor $\sqrt{2}$ simplifies a calculation below), let $\mathbf{\Psi}(s) = [\Psi_1(s), \Psi_2(s), \ldots, \Psi_q(s)]'$ denote a vector of these functions with periods 2 through $2/q$ and let $\mathbf{\Psi}_T$ denote the $T \times q$ matrix with t-th row $\mathbf{\Psi}\left((t-1/2)/T\right)'$. The j-th column of $\mathbf{\Psi}_T$ has period $2T/j$ and is the j-th cosine function plotted in panel (b) of Figure 4. The fitted values shown in panel (c) are from the regression of $\mathbf{x}_{1:T} = (x_1, \ldots, x_T)'$ onto $\mathbf{\Psi}_T^0 = [\mathbf{l}_T, \mathbf{\Psi}_T]$ where $\mathbf{l}_T$ is a $T \times 1$ vector of ones. Denote these OLS regressions coefficients as $\mathbf{X}_T^0 = (\mathbf{\Psi}_T^{0\prime}\mathbf{\Psi}_T^0)^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{x}_{1:T}$. The specific form used for the cosine weights simplifies the analysis because the resulting columns are $\mathbf{\Psi}_T^0$ are orthogonal with

$$T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{\Psi}_T^0 = \begin{bmatrix} T^{-1}\mathbf{l}_T'\mathbf{l}_T & T^{-1}\mathbf{l}_T'\mathbf{\Psi}_T \\ T^{-1}\mathbf{\Psi}_T'\mathbf{l}_T & T^{-1}\mathbf{\Psi}_T'\mathbf{\Psi}_T \end{bmatrix} = \mathbf{I}_{q+1}. \tag{6}$$

The OLS coefficients are then

$$\mathbf{X}_T^0 = (\mathbf{\Psi}_T^{0\prime}\mathbf{\Psi}_T^0)^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{x}_{1:T} = T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{x}_{1:T}, \tag{7}$$

which can be partitioned as

$$\mathbf{X}_T^0 = \begin{bmatrix} \bar{x}_{1:T} \\ \mathbf{X}_T \end{bmatrix} \tag{8}$$

where $\bar{x}_{1:T} = T^{-1}\mathbf{l}_T'\mathbf{x}_{1:T}$ is the sample mean and

$$\mathbf{X}_T = T^{-1}\mathbf{\Psi}_T'\mathbf{x}_{1:T} \tag{9}$$

[8]The low-frequency fitted values plotted in panel (d) and in Figures 1 and 2 are nearly identical to low-pass filtered versions of x_t using, for example, the truncated ideal filter advocated in Baxter and King (1999), with frequency cutoff corresponding to $2T/q$ periods, where the only marked differences are near the endpoints. This is not surprising. Calculations shown in Müller and Watson (2008) show that the cosine transforms used here (or related Fourier transforms) produce close-to-ideal low-pass realizations.

are called *cosine transforms* of $\mathbf{x}_{1:T}$. (We append the superscript '0' to $\mathbf{X}_T^0$ because $\bar{x}_{1:T}$ can be viewed as 0-th cosine transform of $\mathbf{x}_{1:T}$.)

In this notation, the $T \times 1$ vector of low-frequency trend values plotted in panel (c) are

$$\hat{\mathbf{x}}_{1:T} = \mathbf{\Psi}_T^0 \mathbf{X}_T^0 = \mathbf{l}_T \bar{x}_{1:T} + \mathbf{\Psi}_T \mathbf{X}_T. \tag{10}$$

3.2 Large-Sample Normality of Low-Frequency Averages

The low-frequency averages introduced in the last section are normally distributed in large samples for a wide range of stochastic processes, including the models discussed in Section 2. This result is formally developed in Section 8; here we provide an overview.[9]

From (1), $\mathbf{x}_{1:T} = \mathbf{l}_T \mu + \mathbf{u}_{1:T}$ with $\mathbf{u}_{1:T} = (u_1, \ldots, u_T)'$, so the low-frequency averages are $\mathbf{X}_T^0 = \iota_{q+1}\mu + T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{u}_{1:T}$, where $\iota_{q+1} = (1\ \mathbf{0}_q')'$. Thus

$$\mathbf{X}_T^0 - \iota_{q+1}\mu = \begin{bmatrix} \bar{x}_{1:T} - \mu \\ \mathbf{X}_T \end{bmatrix} = T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{u}_{1:T} \tag{11}$$

is a weighted average of the zero-mean random variables $\mathbf{u}_{1:T}$. The large sample behavior of $\mathbf{X}_T^0$ depends on the stochastic process generating $\mathbf{u}_{1:T}$ and the weights making up $\mathbf{\Psi}_T^0$. The central limit theorem in Section 8 provides sufficient conditions on the stochastic process and weights so that (a centered and scaled version of) $\mathbf{X}_T^0$ has a limiting normal distribution. In our applications, the columns of $\mathbf{\Psi}_T^0$ are the constant term and cosine weights in panel (b) of Figure 4, and the stochastic process is one of the five model described in Section 2; these satisfy the conditions given in Section 8 and thus

$$T^{1/2}(\mathbf{X}_T^0 - \iota_{q+1}\mu) \Rightarrow \mathcal{N}\left(0, \sigma^2\mathbf{\Omega}\right). \tag{12}$$

The key implication of (12) is that $\mathbf{X}_T^0$ is approximately normally distributed in large samples

$$\mathbf{X}_T^0 = \begin{bmatrix} \bar{x}_{1:T} \\ \mathbf{X}_T \end{bmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, T^{-1}\sigma^2\mathbf{\Omega}\right). \tag{13}$$

The covariance matrix for $\mathbf{X}_T^0$ depends on both the scale and persistence of the process: the scale factor σ^2 is the long-run variance of the $I(0)$ component a_t and the matrix $\mathbf{\Omega}$ is

[9] Alternatively, the asymptotic normality may be deduced from existing functional central limit results for some models, as explained in Müller and Watson (2017).

a function of the model-specific low-frequency dynamics. In obvious notation we use $\mathbf{\Omega}^{I(0)}$, $\mathbf{\Omega}^{I(1)}$, $\mathbf{\Omega}^{LTU}(c)$, $\mathbf{\Omega}^{FR}(d)$, and $\mathbf{\Omega}^{LL}(g)$ to denote the value of $\mathbf{\Omega}$ for the various models and denote the sub-blocks of $\mathbf{\Omega}$ as $\mathbf{\Omega}_{\bar{x}\bar{x}}$, $\mathbf{\Omega}_{XX}$, and $\mathbf{\Omega}_{\bar{x}X}$. Section 8 derives an expression for $\mathbf{\Omega}$ in terms of the low-frequency properties of the spectrum for x_t. The essential ideas underlying this CLT and resulting expressions for $\mathbf{\Omega}$ can be gleaned from three examples.

In the first example, suppose that $u_t = \varepsilon_t$, a sequence of $i.i.d.(0, \sigma^2)$ random variables. In this case $\sqrt{T}\left(\mathbf{X}_T^0 - \iota_{q+1}\mu\right) = T^{-1/2}\sum_{t=1}^T \psi_t^0 \varepsilon_t$ where $\psi_t^{0\prime}$ is the t-th row of $\mathbf{\Psi}_T^0$. Thus $(\mathbf{X}_T^0 - \iota_{q+1}\mu)$ is a weighted average of $\varepsilon_{1:T} = [\varepsilon_1, \ldots, \varepsilon_T]'$ with weights given by the constant and periodic terms making up the rows of $\mathbf{\Psi}_T^0$. A central limit theorem yields the large sample normality and a direct calculation shows that $Var\left(T^{-1/2}\sum_{t=1}^T \psi_t^0 \varepsilon_t\right) = \sigma^2 T^{-1}\sum_{t=1}^T \psi_t^0 \psi_t^{0\prime} = \sigma^2 T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{\Psi}_T^0 = \sigma^2 \mathbf{I}_{q+1}$, where the first equality uses the fact that ε_t is $i.i.d.$, the second equality uses the definition of the weights ψ_t^0 and the final equality uses the properties of $\mathbf{\Psi}_T^0$ given in (6).

In the second example, suppose $u_t = \varepsilon_t + c_1\varepsilon_{t-1}$, where ε_t follows the $i.i.d.$ process from the first example, so that u_t follows a MA(1) process. Here, $\sqrt{T}\left(\mathbf{X}_T^0 - \iota_{q+1}\mu\right) = T^{-1/2}\sum_{t=1}^T \psi_t^0 u_t = T^{-1/2}\sum_{t=1}^T \psi_t^0(\varepsilon_t + c_1\varepsilon_{t-1}) = (1+c_1)T^{-1/2}\sum_{t=1}^T \psi_t^0\varepsilon_t - R_T$, where R_T is a remainder term with $R_T = T^{-1/2}c_1(\varepsilon_T - \varepsilon_0) + c_1 T^{-1/2}\sum_{t=1}^T(\psi_t^0 - \psi_{t-1}^0)\varepsilon_{t-1}$. This remainder term is $o_p(1)$: for the first term this is obvious, and for the second because the weights are sufficiently smooth so that $T^{-1}\sum_{t=1}^T(\psi_t^0 - \psi_{t-1}^0)'(\psi_t^0 - \psi_{t-1}^0) \to 0$. Thus, the second example differs from the first only through the additional scale factor $(1 + c_1)$, so that (12) holds with $\mathbf{\Omega} = \mathbf{I}_{q+1}$ and $\sigma^2 = (1+c_1)^2\sigma_\varepsilon^2$, which is the long-run variance of u_t. A similar argument applies when u_t follows a general $I(0)$ process: σ^2 is the long-run variance of u_t and $\mathbf{\Omega}^{I(0)} = \mathbf{I}_{q+1}$.

In the third example, suppose u_t is a random walk with $u_t = T^{-1}\sum_{j=1}^t \varepsilon_j$. In this case, $\sqrt{T}\left(\mathbf{X}_T^0 - \iota_{q+1}\mu\right) = T^{-1/2}\sum_{t=1}^T \psi_t^0 u_t = \sum_{t=1}^T \psi_t^a \varepsilon_t$ where $\psi_t^a = T^{-1}\sum_{j=t}^T \psi_j^0$. Again, $(\mathbf{X}_T^0 - \iota_{q+1}\mu)$ is a weighted average of $\varepsilon_{1:T}$ but now the weights are partial sums of the rows of $\mathbf{\Psi}_T^0$. Again, a direct calculation shows $Var\left(T^{-1/2}\sum_{t=1}^T \psi_t^a \varepsilon_t\right) = \sigma^2 T^{-1}\sum_{t=1}^T \psi_t^a \psi_t^{a\prime}$, so $\mathbf{\Omega}^{I(1)} = \lim_{T\to\infty} T^{-1}\sum_{t=1}^T \psi_t^a \psi_t^{a\prime}$. While not as as straightforward as in the $I(0)$ case, an explicit expression for $\mathbf{\Omega}^{I(1)}$ can be derived from this limit with a little effort, and yields

$$\mathbf{\Omega}^{I(1)} = \begin{pmatrix} \mathbf{\Omega}_{\bar{x}\bar{x}}^{I(1)} & \mathbf{\Omega}_{\bar{x}X}^{I(1)} \\ \mathbf{\Omega}_{X\bar{x}}^{I(1)} & \mathbf{\Omega}_{XX}^{I(1)} \end{pmatrix} \tag{14}$$

where $\mathbf{\Omega}^{I(1)}_{\bar{x}\bar{x}} = 1/3$, $\mathbf{\Omega}^{I(1)}_{\bar{x}X} = (-\sqrt{2}\pi^{-2}, -\sqrt{2}(2\pi)^{-2}, \ldots, -\sqrt{2}(q\pi)^{-2})$ and $\mathbf{\Omega}^{I(1)}_{XX} = diag(\pi^{-2}, (2\pi)^{-2}, \ldots, (q\pi)^{-2})$. As in the second example, this random walk result generalizes to other $I(1)$ processes with σ^2 equal to the long-run variance of the driving process.

These examples suggest a simple method for computing $\mathbf{\Omega}$ for any of the models introduced in Section 2. With $\sqrt{T}(\mathbf{X}^0_T - \iota_{q+1}\mu) = T^{-1/2}\mathbf{\Psi}^{0\prime}_T\mathbf{u}_{1:T}$, $Var(\sqrt{T}(\mathbf{X}^0_T - \iota_{q+1}\mu)) = T^{-1}\mathbf{\Psi}^{0\prime}_T\mathbf{\Lambda}_T\mathbf{\Psi}^0_T$, where $\mathbf{\Lambda}_T$ is the $T \times T$ covariance matrix for $\mathbf{u}_{1:T}$. Specifically, for each of the models with $\sigma^2 = 1$, we have

1. $I(0)$ model: $\mathbf{\Lambda}_T = \mathbf{I}_T$
2. $I(1)$ model: $\mathbf{\Lambda}_T = T^{-2}\mathbf{A}_T\mathbf{A}'_T$, where $\mathbf{A}_T$ is a lower triangular matrix of ones
3. LTU model: the ijth element of $\mathbf{\Lambda}_T(c)$ is $\Lambda_{ij,T}(c) = T^{-2}\rho_T^{|i-j|}/(1-\rho_T^2)$ with $\rho_T = 1 - c/T$
4. Stationary $I(d)$ model $(-1/2 < d < 1/2)$:[10] $\Lambda_{ij,T}(c) = Cov(u_i, u_j) = T^{-4d}\{\Gamma(k+d)\Gamma(1-2d)\}/\{\Gamma(k+1-d)\Gamma(1-d)\Gamma(d)\}$ with $k = |i-j|$; $I(d)$ model with $1/2 < d < 3/2$: $\mathbf{\Lambda}_T(d) = T^{-2}\mathbf{A}_T\mathbf{\Lambda}_T(d-1)\mathbf{A}'_T$
5. Local-Level model: $\mathbf{\Lambda}_T(g) = \mathbf{I}_T + (g/T)^2\mathbf{A}_T\mathbf{A}'_T$.

Thus

$$\mathbf{\Omega}_T = T^{-1}\mathbf{\Psi}^{0\prime}_T\mathbf{\Lambda}_T\mathbf{\Psi}^0_T \to \mathbf{\Omega}. \tag{15}$$

In many of our calculations we use this expression for $\mathbf{\Omega}_T$ with $T = 1000$ as an approximation for $\mathbf{\Omega}$.

3.3 Multivariate Models and Low-Frequency Covariation

We use the following notation for multivariate models: $\mathbf{x}_t = [x_{1,t}, \ldots, x_{n,t}]'$ is an $n \times 1$ vector-valued time series with $\mathbf{x}_t = \mu + \mathbf{u}_t$ with mean $\mu = [\mu_1, \ldots, \mu_n]'$, $\mathbf{x}_{1:T}$ is a $T \times n$ matrix with t-th row equal to $\mathbf{x}'_t$ and similarly for $\mathbf{u}_{1:T}$; $\mathbf{X}^0_T$, as defined in (7), is $(q+1) \times n$; $\hat{\mathbf{x}}_{1:T} = \mathbf{\Psi}^0_T\mathbf{X}^0_T$ is the $T \times n$ matrix of trend values.

The multivariate analog to (12) is

$$T^{1/2}\left(\mathbf{X}^0_T - \iota_{q+1}\mu'\right) \Rightarrow \mathbf{X}^0 \text{ with } vec(\mathbf{X}^0) \sim \mathcal{N}(0, \mathbf{V}) \tag{16}$$

[10] See Baillie (1996)) for this and alternative formulae.

so that

$$vec(\mathbf{X}_T^0) \overset{a}{\sim} \mathcal{N}\left(\mu \otimes \iota_{q+1}, T^{-1}\mathbf{V}\right). \tag{17}$$

The matrix $\mathbf{V}$ can be partitioned into $(q+1)\times(q+1)$ blocks $\mathbf{V}_{ij} = Cov(\mathbf{X}_i^0, \mathbf{X}_j^0)$ where $\mathbf{X}_i^0$ denotes the i-th column of $\mathbf{X}^0$ which is computed from the i-th series $x_{i,t}$. As in the univariate model, the matrix $\mathbf{V}$ depends on the scale (σ in the univariate model) and persistence. A leading case is the $I(0)$ model in which $\mathbf{V}_{ij} = \sigma_{ij}\mathbf{I}_{q+1}$, where σ_{ij} is the long-run covariance between $x_{i,t}$ and $x_{j,t}$. We discuss other multivariate models in the examples scattered throughout the chapter.

The low-frequency covariance between $x_{i,t}$ and $x_{j,t}$ is defined as the population covariance between the low-frequency trend values, $\hat{x}_{i,t}$ and $\hat{x}_{j,t}$ averaged over the length of the sample

$$\sigma_{ij}^{LF} = T^{-1}\sum_{t=1}^{T} E\left[\left(\hat{x}_{i,t} - \mu_i\right)\left(\hat{x}_{j,t} - \mu_j\right)\right]. \tag{18}$$

Collecting the low-frequency covariances into the $n\times n$ matrix $\mathbf{\Sigma}^{LF}$ with elements σ_{ij}^{LF}, some simple algebra shows the relationship between $\mathbf{\Sigma}^{LF}$ and $\mathbf{V}$ in (16):

$$\begin{aligned}
\mathbf{\Sigma}^{LF} &= T^{-1}E\left[\left(\hat{\mathbf{x}}_{1:T} - \mathbf{l}_T\mu'\right)'\left(\hat{\mathbf{x}}_{1:T} - \mathbf{l}_T\mu'\right)\right] \\
&= T^{-1}E\left[\left(\mathbf{\Psi}_T^0\left(\mathbf{X}_T^0 - \iota_{q+1}\mu'\right)\right)'\left(\mathbf{\Psi}_T^0\left(\mathbf{X}_T^0 - \iota_{q+1}\mu'\right)\right)\right] \\
&= E\left[\left(\mathbf{X}_T^0 - \iota_{q+1}\mu'\right)'\left(\mathbf{X}_T^0 - \iota_{q+1}\mu'\right)\right]
\end{aligned}$$

where the last equality uses $T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{\Psi}_T^0 = \mathbf{I}_{q+1}$. Thus, using the large-sample approximation in (17),

$$\sigma_{ij}^{LF} = T^{-1}tr(\mathbf{V}_{ij}). \tag{19}$$

Associated with the low-frequency covariance matrix are the usual correlations, linear regression coefficients, etc. For example

$$\rho_{ij}^{LF} = \frac{\sigma_{ij}^{LF}}{\sqrt{\sigma_{ii,T}^{LF}\sigma_{jj}^{LF}}}$$

is the correlation between $\hat{\mathbf{x}}_{i,t}$ and $\hat{\mathbf{x}}_{j,t}$ and

$$\beta_{1,i:j}^{LF} = \left[\mathbf{\Sigma}_{i:j,i:j}^{LF}\right]^{-1}\mathbf{\Sigma}_{i:j,1}^{LF} \tag{20}$$

shows the low-frequency coefficients from the regression of $\hat{\mathbf{x}}_{1,t}$ onto $\hat{\mathbf{x}}_{i:j,t}$. Low-frequency instrumental variable regression coefficients may be defined analogously.

3.4 An Example: Inference about μ in the $I(0)$ Model

Arguably the most important low-frequency inference problem concerns the value of the mean, μ, in the $I(0)$ model and the associated extensions to parameters in regression, instrumental variable, and GMM problems. We begin by discussing inference for scalar x_t, and extend this to multivariate models below.

In the Gaussian $I(0)$ model with σ^2 known, a classic result from Grenander and Rosenblatt (1957) shows that efficient inference about μ relies on the data through $\bar{x}_{1:T}$ with $\bar{x}_{1:T} \overset{a}{\sim} \mathcal{N}(\mu, (\sigma^2/T))$. A large literature has focused on deriving consistent estimators for σ^2, which can be used in place of σ^2; these are *heteroskedastic and autocorrelation consistent* (HAC) estimators, where 'heteroskedastic' refers to their use in regression models.[11] Inference relies on

$$\frac{T^{1/2}(\bar{x}_{1:T} - \mu)}{\hat{\sigma}} \Rightarrow \mathcal{N}(0,1) \tag{21}$$

where $\hat{\sigma}$ is a HAC estimator of σ. The result in (21) leads, for example, to $100(1-\alpha/2)\%$ confidence intervals for μ of the form $\bar{x}_{1:T} \pm z_{1-\alpha/2}\hat{\sigma}/\sqrt{T}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

Another literature has highlighted important problems with using (21) as the basis for finite-sample inference.[12] These problems arise because (21) neglects the estimation error in $\hat{\sigma}$. HAC estimators that use a small bandwidth result in a large variance for $\hat{\sigma}^2$, while estimators using a large bandwidth have large bias when the data exhibit moderate-to-severe autocorrelation. To address these problems, recent work has focused on heteroskedastic-autocorrelation-robust (HAR) inference that explicitly accounts for sampling uncertainty about σ.[13]

A class of HAR procedures utilizes (13) with $\mathbf{\Omega} = \mathbf{I}_{q+1}$ and classic inference procedures for small-sample *i.i.d.* normal samples. For example, with $s^2 = \frac{T}{q}\sum_{j=1}^{q} X_{jT}^2 = \frac{T}{q}\mathbf{X}_T'\mathbf{X}_T$, standard properties of the multivariate normal distribution (e.g., Rao (1973)) imply

$$qs^2/\sigma^2 \overset{a}{\sim} \chi_q^2 \tag{22}$$

[11] Important early references in economics include Hansen (1982), Domowitz and White (1982), Newey and West (1987), and Andrews (1991).

[12] See Section 2 of Müller (2014) for a recent survey and references.

[13] In economics, early contributions include Kiefer et al. (2000), whose methods require non-standard probability distributions and associated critical values, and Müller (2004) whose methods utilize Student-t and F distributions. Additional references are provided below. The discussion here follows Müller (2004).

and

$$\frac{\sqrt{T}\,(\bar{x}_{1:T}-\mu)}{s} \overset{a}{\sim} \text{Student-}t_q \tag{23}$$

and where the approximate distributions in both (22) and (23) become exact as $T \to \infty$. Inverting the t-statistic in (23) yields the usual $100\,(1-\alpha/2)\,\%$ confidence interval for μ

$$\bar{x}_{1:T} \pm t_{q,1-\alpha/2} s/\sqrt{T} \tag{24}$$

where $t_{q,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student-t_q distribution.

The approximation (13) also serves as the basis for large-sample Bayes inference. Standard conjugate priors for μ and σ^2 lead to standard posteriors, and in particular, the usual uninformative priors imply that (24) is a $100\,(1-\alpha/2)\,\%$ credible interval for μ.

As equation (23) makes clear, the value of q is key for HAR inference. When q is large, the Student-t_q distribution is essentially the standard normal, so that (21) is being used, while smaller q leads to more uncertainty about σ and thus wider confidence/credible intervals for μ. The choice of q faces the same bias-variance trade-offs as the bandwidth choice for HAC estimators: larger values of q mean that higher frequency variability in the data is being used to estimate the long-run variance σ^2. This induces a bias in s^2 when the data are autocorrelated. We discuss the choice of q in more detail below.

Empirical example. The TFP growth rates shown in Figure 1 have a sample mean of $\bar{x}_{1:T} = 1.21$. Using $q = 14$, the estimated long-run standard deviation is $s = 3.83$. Using (24), a 95% confidence/credible interval is $0.72 < \mu < 1.69$. Panel (a) of Table 1 shows selected quantiles of the flat-prior Bayes posterior; the table also shows results for several of the other examples discussed in the next section.

3.4.1 Multivariate and Linear Regression Extensions for the $I(0)$ Model

The $I(0)$ results outlined above are for the univariate mean, but they extend readily to the vector model. Suppose $\mathbf{x}_t$ is an $n \times 1$ vector that follows an $I(0)$ process with long-run covariance matrix $\mathbf{\Sigma}$. Then (17) holds with $\mathbf{V} = \mathbf{\Sigma} \otimes \mathbf{I}_{q+1}$. Hotelling's-$T^2$ statistic (Hotelling (1931))

$$T(\bar{\mathbf{x}}_{1:T}-\mu)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_{1:T}-\mu) \overset{a}{\sim} \frac{nq}{q+1-n} F_{n,q+1-n} \tag{25}$$

can be used for inference, where $\mathbf{S} = (T/q)\,\mathbf{X}_T'\mathbf{X}_T$, and $F_{n,q+1-n}$ is the F-distribution with n and $q+1-n$ degrees of freedom. (See Rao (1973) or, in this context, Müller (2004)).

Parameter	Posterior Mean	Posterior Quantiles				
		0.05	**0.17**	**0.50**	**0.83**	**0.95**
(a) TFP Growth Rate, $I(0)$ model. Prior: $f(\mu, \sigma) \propto 1/\sigma$						
μ	1.21	0.81	0.98	1.21	1.43	1.61
(b) \$/£ Real Exchange Rate, LTU model						
	(i) Prior: $f(\mu, \sigma) \propto 1/\sigma$, $\rho \sim U(0.50,0.999)$					
ρ	0.93	0.85	0.90	0.94	0.97	0.98
half-life	25.7	4.4	6.7	11.7	23.5	50.7
	(ii) Prior: $f(\mu, \sigma) \propto 1/\sigma$, *half-life* $\sim U(0,100)$					
ρ	0.97	0.93	0.96	0.98	0.99	0.99
half-life	42.7	9.7	15.5	36.6	74.7	92.1
(c) Daily Realized Volatility, $I(d)$ model. Prior: $f(\mu, \sigma) \propto 1/\sigma$, $d \sim$ U(-0.4,1.4)						
d	0.59	0.36	0.45	0.58	0.72	0.82
(d) TFP Growth Rate, LLM, Prior: $f(\mu, \sigma) \propto 1/\sigma$, ln(g) ~ U(ln(0.1), ln(500))						
σ	3.40	2.04	2.56	3.32	4.24	5.07
g	8.8	0.16	0.53	4.4	11.8	22.3
$\sigma_{\Delta b}$	0.067	0.002	0.007	0.051	0.118	0.193
(e) Unemployment Rate. Prior: $f(\mu, \sigma) \propto 1/\sigma$						
	(i) *LTU model*, Prior: ln(c) ~ U(ln(0.1), ln(500))					
ρ	0.90	0.57	0.78	0.96	0.99	1.00
μ	5.53	3.69	5.06	5.66	6.18	6.87
	(ii) *I(0) model*					
μ	5.74	5.10	5.38	5.74	6.11	6.38
(f) TFP growth rate, $I(0)$ with break in mean, $f(\mu_{pre}, \mu_{post}, \sigma) \propto 1/\sigma$, $r \sim$ U(0,1)						
μ	2.11	1.24	1.63	2.08	2.50	2.88
$\mu+\delta$	0.80	0.36	0.60	0.84	1.08	1.30
δ	-1.31	-2.18	-1.74	-1.28	-0.80	-0.33
(f) Bivariate TFP and GDP Growth Rates, $I(0)$ model. $f(\mu, \Sigma) \propto \lvert\Sigma\rvert^{-1}$						
ρ^{LF}	0.80	0.60	0.71	0.82	0.89	0.93
β^{LF}	1.10	0.73	0.88	1.10	1.31	1.47

Table 1: Posterior mean and selected posterior quantiles for Bayes empirical examples

Importantly, these results generalize for inference in $I(0)$ regression and GMM models. For example, in the linear regression model $y_t = \mathbf{z}_t'\beta + u_t$, $\hat{\beta} - \beta$ replaces $\bar{\mathbf{x}}_{1:T} - \mu$ in the $I(0)$ mean problem, $\mathbf{S}_{zz}^{-1}\mathbf{z}_t u_t$ replaces $\mathbf{x}_t$, where $\hat{\beta}$ is the OLS estimator and $\mathbf{S}_{zz} = T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t'$. In this case $\mathbf{\Sigma}$ is the large-sample variance of $T^{1/2}(\hat{\beta} - \beta)$, and $\mathbf{S}$ is the corresponding (HAR) estimator. See Müller (2004, 2014), Phillips (2005), Sun (2013), and Lazarus et al. (2017) for a more detailed discussion.

3.5 Choice of q and Limited-Information Low-Frequency Inference

As discussed in the introduction, and as the univariate $I(0)$ example makes clear, low-frequency inference is conducted using only the information contained in the sample mean $\bar{x}_{1:T}$ and the q low-frequency transforms in $\mathbf{X}_T$. Why, one might ask, should inference be based solely on these $q+1$ statistics, effectively discarding the rest of the data? One answer is practical: restricting the analysis to these statistics leads to tractable and robust statistical inference procedures. The key feature underlying tractability is the large-sample normal law the transformed data. The statistical analysis can thus draw on the extensive suite of methods that have been developed for finite sample (here $q+1$ dimensional) Gaussian models. Large-sample robustness shows up in a variety of forms. First, and related to the first point, the methods are robust to the distributional properties of the original data x_t, through the use of averages and normal laws. Second, by restricting the analysis to low-frequency variability in the data, modeling becomes easier, because only the low-frequency properties of the model affect inference. For example, in the $I(0)$ model, the exact form of dependence (AR(4) versus ARMA(2,2) versus MA(20), say) plays no role; all that matters is the mean and the long-run variance of the process.

This said, the low-frequency statistics used in the analysis are not sufficient statistics, so there is a loss of efficiency. For example, a special case of the $I(0)$ model is the *i.i.d.* Gaussian process, where the sufficient statistics are $\bar{x}_{1:T}$ and the sample variance. In this special case, the power loss associated with restricting the sample to the $(q+1)$ observations in $\mathbf{X}_T^0$ is easily quantified by the larger critical value in (23): An efficient test would use a t-statistic with $T-1$ degrees of freedom, while the low-frequency analysis uses a t-statistic with only q degrees of freedom. (Recall that in the TFP example $q = 14$.) Of course, in this case, the robustness considerations of the last paragraph would lead one to ask how sure one

could be that the data are *i.i.d.* Gaussian. Indeed, as shown in Müller (2011) in a general setting, it is not possible to use pretests or otherwise learn from the data that it is *i.i.d.* to obtain more efficient hypothesis tests without inducing size distortion for some process that satisfies (12); in other words, with (12) the minimal restriction that one is willing to put on the x_t process, asymptotically efficient tests simply rely only on the low-frequency transforms of the data.

There remains an important practical question: what value of q should be used in a particular application. There are two guiding principles.

First, one way to think about q is definitional: q defines *low-frequency* in the analysis. To see why, return to Figure 4 which showed the low-frequency transforms for TFP growth rates over $T = 285$ quarters from 1948:Q2 through 2019:Q2, computed using $q = 14$. The resulting low-frequency trend captures variability for periods longer than $2T/q \approx 41$ quarters (≈ 10 years).[14] If instead, $q = 8$, the analysis would capture variability $2T/8 \approx 72$ quarters (or 18 years). This suggests that a researcher interested in variability over periods of 10-years or longer should use $q = 14$; a researcher interested in variability over periods of 18-years or longer should use $q = 8$.

Second, q defines the low-frequency range over which the normal distribution in (13) provides a reliable approximation for inference. Consider the TFP example in the $I(0)$ model. The approximation in (13) has two features: the normal limit and the specific covariance matrix $\sigma^2 T^{-1}\mathbf{I}_{q+1}$. When the data are non-Gaussian, a concern is the large-q accuracy of the multivariate normal distribution, leading to misspecification of the likelihood for Bayes inference and to the results in (22), (23) and (25) that form the basis for frequentist inference in the $I(0)$ model. Moreover, as shown in Section 8, the limiting covariance matrix depends critically on the shape of (pseudo-) spectrum in a local $(1/T)$ neighborhood of frequency zero. The weighted averages in $\mathbf{X}_T$ use variability for frequencies as high as $q/(2T)$, so the limiting covariance matrix may be a poor approximation when q is large. The univariate $I(0)$ model provides a clear example. Suppose x_t follows a stationary AR(1) model with coefficient ρ and innovation variance κ^2. The long-run variance, that is, the limiting variance of $\sqrt{T}\left(\bar{x}_{1:T} - \mu\right)$, is $\sigma^2 = \kappa^2/(1-\rho)^2$. The spectrum of the process at frequency ϕ satisfies $2\pi\Upsilon(\phi) = \kappa^2/\left(1 + \rho^2 - 2\rho\cos(\phi)\right)$, so that $2\pi\Upsilon\left(\phi\right) \approx \sigma^2$ for small values of ϕ. But note

[14] See Müller and Watson (2008) for a discussion of the ability of the q cosine transforms to capture variability for frequencies lower than $q/(2T)$.

that the value of the approximation depends on the values of both ϕ and ρ. When $\rho = 0$, the data are serially uncorrelated and the approximation holds exactly for all values of ϕ. On the other hand, when $|\rho|$ is large, the approximation deteriorates quickly for larger $|\phi|$. Letting $\phi = q/2T$, suppose $T = 285$ as in the TFP example. Then, for $\rho = 0.5$, $2\pi\Upsilon(\phi)/\sigma^2$ is equal to $\{0.99, 0.96, 0.91\}$ for $q = \{7, 14, 21\}$, but when $\rho = 0.8$, $2\pi\Upsilon(\phi)/\sigma^2$ is equal to $\{0.90, 0.68, 0.49\}$ for $q = \{7, 14, 21\}$. Thus, the estimator s^2 of σ^2 in (23) is likely to be severely downward biased using $q = 14$ when $\rho = 0.8$, but exhibits little bias when $\rho = 0.5$. This suggests using a small value of q for the $I(0)$ approximation for more persistence processes. But, a small value of q reduces the degrees of freedom, resulting in tests with lower power. See Lazarus et al. (2017) and Lazarus et al. (2018) for discussion of the choice of q for $I(0)$ inference that explicitly considers the trade-off between size distortion (choosing q too large) and power loss (choosing q too small), and Dou (2019) for optimal inference procedures given an explicit upper bound for ρ.

4 Bayes Inference: Examples

As stressed in the introduction, low-frequency inference is inherently a small-sample statistical problem. For example, while the TFP example uses data from 71-year sample, variability for periods longer than a decade are summarized by the vector $\mathbf{X}_T^0$ that contains only $q+1 = 15$ elements. With small samples, Bayes and frequentist inference typically differ (although their coincidence in the $I(0)$ analysis of the mean is an interesting counterexample to this general rule). Section 6 provides a detailed overview for the Bayes methods employed in this chapter. For convenience, a few of the Bayes results used in this section's examples are summarized here.

4.1 Some Bayes Basics

Let $\mathbf{Y}$ denote vector of random variables with a probability density, say $f(\mathbf{Y}|\theta)$ (the *likelihood*) that depends on a parameter vector θ. Let $p(\theta)$ be a probability density that describes the *a priori* uncertainty about the value of θ (the *prior*). After observing $\mathbf{Y}$, the goal is the calculation of the updated probability density function $p(\theta|\mathbf{Y})$ (the *posterior*) given the observations $\mathbf{Y}$. All the low-frequency examples discussed here concern inference about the

mean and/or covariance matrix from a normal likelihood, and the Bayes analysis uses standard methods. We begin with three results (see Gelman et al. (2004)) that are useful in this context.

4.1.1 Some Specific Priors and Posteriors

Posterior for the mean: Suppose that $\mathbf{Y}|\mu \sim \mathcal{N}(\mathbf{H}\mu, \mathbf{\Sigma})$, where $\mathbf{H}$ and $\mathbf{\Sigma}$ are known and the prior is

$$\mu \sim \mathcal{N}(\mathbf{m}, \mathbf{\Lambda}). \tag{26}$$

Then $(\mathbf{Y}, \mu)$ are jointly normally distributed with $\mathbf{Y} \sim \mathcal{N}(\mathbf{Hm}, \mathbf{\Sigma} + \mathbf{H\Lambda H}')$ and $Cov(\mathbf{Y}, \mu) = H\Lambda$. The well-known conditional distribution for multivariate normals then shows that the posterior for μ is

$$\mu|\mathbf{Y} \sim \mathcal{N}\left(\mathbf{m} + \mathbf{K}\left(\mathbf{Y} - \mathbf{H}\mu\right), \mathbf{\Lambda} - \mathbf{KH\Lambda}\right) \tag{27}$$

where $\mathbf{K} = \mathbf{\Lambda H}'\left(\mathbf{H\Lambda H}' + \mathbf{\Sigma}\right)^{-1}$. Note that when $\mathbf{\Lambda} = \kappa^2\mathbf{I}$ with $\kappa^2 \to \infty$, this posterior distribution converges to

$$\mu|\mathbf{Y} \sim \mathcal{N}\left(\mathbf{H}'\mathbf{Y}, (\mathbf{H}'\mathbf{\Sigma}^{-1}\mathbf{H})^{-1}\right). \tag{28}$$

Posterior for variance: Suppose the $n_Y \times 1$ vector $\mathbf{Y}|\sigma^2 \sim \mathcal{N}\left(\mu, \sigma^2\mathbf{\Omega}\right)$, where μ and $\mathbf{\Omega}$ are known, and the prior is

$$\sigma^2 \sim \mathcal{IG}(\alpha, \beta), \tag{29}$$

the inverse Gamma distribution with parameters α and β. The posterior for σ^2 is

$$\sigma^2|\mathbf{Y} \sim \mathcal{IG}\left(\alpha + n_Y/2, \beta + \left(\mathbf{Y} - \mu\right)'\mathbf{\Omega}^{-1}\left(\mathbf{Y} - \mu\right)\right). \tag{30}$$

We also rely on a multivariate extension of this result: Suppose $\mathbf{Y}$ is an $n \times m$ matrix with $vec(\mathbf{Y}) \sim \mathcal{N}\left(0, \mathbf{\Sigma} \otimes \mathbf{I}_n\right)$ where $\mathbf{\Sigma}$ is $m \times m$, and the inverse-Wishart prior for $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} \sim \mathcal{IW}(\mathbf{\Lambda}, \nu). \tag{31}$$

The posterior for $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma}|\mathbf{Y} \sim \mathcal{IW}\left(\mathbf{\Lambda} + \mathbf{Y}'\mathbf{Y}, \nu + n\right) \tag{32}$$

and the posteriors for the associated correlations, regression coefficients, etc., follow directly.

Posterior with discrete support: Suppose $\mathbf{Y}$ has likelihood $f(\mathbf{Y}|\vartheta)$, where the parameter ϑ can take on one of n_ϑ known values $\{\vartheta_1, \ldots, \vartheta_{n_\vartheta}\}$ with prior

$$P(\vartheta = \vartheta_i) = p_i. \tag{33}$$

The posterior for ϑ is

$$P(\vartheta = \vartheta_i|\mathbf{Y}) = \frac{f(\mathbf{Y}|\vartheta_i)\, p_i}{\sum_{j=1}^{n_\vartheta} f(\mathbf{Y}|\vartheta_j)\, p_j}. \tag{34}$$

4.1.2 A Gibbs Algorithm

Many of the examples in this section have a common structure and the posterior can be formed using draws from a Gibbs MCMC algorithm. (See Section 6 for a general discussion of these methods.)

Generically, let $\mathbf{Y}$ denote an $n_Y \times 1$ vector with

$$\mathbf{Y}\,|(\mu, \sigma, \vartheta) \sim \mathcal{N}\left(\mathbf{H}(\vartheta)\mu, \sigma^2\mathbf{\Omega}(\vartheta)\right) \tag{35}$$

where ϑ has discrete support. Suppose the parameters (μ, σ, ϑ) have the normal, inverse Gamma and discrete priors given in (26), (29), and (33). Draws from the posterior $(\mu, \sigma, \vartheta)\,|\mathbf{Y}$ can be obtained from the following 3-step Gibbs algorithm:

1. Draw μ from the posterior $\mu\,|(\mathbf{Y}, \sigma, \vartheta)$ using (27) with (35) as the likelihood and (26) as the prior.

2. Draw σ^2 from the posterior $\sigma^2\,|(\mathbf{Y}, \mu, \vartheta)$ using (30) with (35) as the likelihood and (29) as the prior.

3. Draw ϑ from the posterior $\vartheta\,|(\mathbf{Y}, \mu, \sigma)$ using (34) with (35) as the likelihood and (33) as the prior.

4.1.3 Invariance and Uninformative Priors

As discussed in Sections 6 and 7, there is a close connection between certain kinds of invariance and uninformative priors for location and scale parameters. These flat prior/invariance results simplify some of the Bayes calculations in this section.

To be specific, and mimicking the notation in (13), suppose that

$$\mathbf{X}_T^0 = \begin{bmatrix} \bar{x}_{1:T} \\ \mathbf{X}_T \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, T^{-1}\sigma^2 \begin{bmatrix} \mathbf{\Omega}_{\bar{x}\bar{x}}(\vartheta) & \mathbf{\Omega}_{\bar{x}X}(\vartheta) \\ \mathbf{\Omega}_{X\bar{x}}(\vartheta) & \mathbf{\Omega}_{XX}(\vartheta) \end{bmatrix} \right) \tag{36}$$

where (μ, σ, ϑ) are unknown parameters. Suppose that the prior for μ is flat, so that the posterior is as in (28). In this case, a calculation shows that the posterior for (σ, ϑ) only depends on $\mathbf{X}_T$, and no longer on $\bar{x}_{1:T}$. That is, the posterior for (σ, ϑ) can be computed using the likelihood $\mathbf{X}_T \,|(\sigma, \vartheta) \sim \mathcal{N}(0, \sigma^2 \mathbf{\Omega}_{XX}(\vartheta))$, ignoring the data $\bar{x}_{1:T}$ and the parameter μ. Note that $\mathbf{X}_T$ is the component of $\mathbf{X}_T^0$ that remains invariant to translations of the underlying data $\{x_t\}_{t=1}^T \to \{x_t + a\}_{t=1}^T$.

As another example, suppose that

$$\mathbf{X}_T \sim \mathcal{N}\left(0, T^{-1}\sigma^2 \mathbf{\Omega}_{XX}(\vartheta)\right) \tag{37}$$

where (σ, ϑ) are unknown parameters. Suppose that the prior for σ is uninformative with $p(\sigma) \propto 1/\sigma$ (corresponding to $\alpha = \beta = 0$ in (29)). In this case, the posterior for ϑ depends on the data only through the value of $\mathbf{X}_T^s = \mathbf{X}_T / \sqrt{\mathbf{X}_T' \mathbf{X}_T}$; thus $\mathbf{X}_T$ and $a\mathbf{X}_T$ with $a > 0$ yield the same posterior for ϑ, and $\mathbf{X}_T^s$ remains invariant to scale transformations of the underlying data. The probability density for $\mathbf{X}_T^s$ is (see, for instance, Kariya (1980) or King (1980))

$$f_{\mathbf{X}_T^s}(\mathbf{x}_T^s | \vartheta) = C \left| \mathbf{\Omega}_{XX}(\vartheta) \right|^{-1/2} \left(\mathbf{x}_T^{s\prime} \mathbf{\Omega}_{XX}(\vartheta)^{-1} \mathbf{x}_T^s \right)^{-q/2} \tag{38}$$

where C is a constant. Thus, when interest is focused on ϑ, σ can be dropped from the analysis by restricting attention to $\mathbf{X}_T^s$ and using the likelihood (38).

With an uninformative prior $p(\sigma) \propto 1/\sigma$ the scale of $\mathbf{\Omega}_{XX}(\vartheta)$ in (13) is immaterial, as is easily seen by inspecting (38). It is thus without loss of generality to normalize $\mathbf{\Omega}_{XX}(\vartheta)$ to satisfy $tr\mathbf{\Omega}_{XX}(\vartheta) = q$, say. The continuity of the low-frequency implications for the fractional model over $d \in (-1/2, 3/2)$ then holds after such a normalization (which amounts to making the long-run variance σ^2 of the underlying process a_t a particular function of d). As a computational matter, this normalization is particularly useful when the elements of Ω_{XX} have very different scale, as this can lead to poor mixing in Step 3 of the Gibbs algorithm outlined above.

4.2 Inference about Low-Frequency Persistence

This section takes up three inference problems involving low-frequency persistence. The first problem concerns the parameter c in the LTU model (2); the second problem concerns d in the $I(d)$ model (3); the third involves low-frequency variability in the $I(1)$ component, b_t, in the LL model (4). Each of these problems has a common structure

$$\mathbf{X}_T^0 = \begin{bmatrix} \bar{x}_{1:T} \\ \mathbf{X}_T \end{bmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, T^{-1}\sigma^2 \mathbf{\Omega}(\vartheta) \right) \tag{39}$$

where the persistence parameter ϑ equals c for the LTU model, d for the $I(d)$ model and g for the LL model. The unknown parameters are (μ, σ, ϑ).

Draws from the posterior for (μ, σ, ϑ) can be obtained from the Gibbs algorithm from Section 4.1.2. However, the uninformative-prior considerations of the last subsection lead to two simplifications for the three exercises in this subsection. The first simplification arises because μ is not a parameter of interest in any of the exercises. Thus, with a flat prior for μ, the posterior for the remaining two parameters (σ, ϑ) can be computed using the data in $\mathbf{X}_T$. Similarly, the first two exercises focus solely on the persistence parameter ϑ (where $\vartheta = c$ in the LTU example and $\vartheta = d$ in the $I(d)$ example), so with an uninformative prior for σ, the posterior for ϑ can be computed using the scale-normalized data $\mathbf{X}_T^s$.

4.2.1 Low-Frequency Persistence in the Stationary Local-to-Unity Model

Recall that in the LTU model, x_t has low-frequency persistence parameterized by a local-to-unity AR parameter $\rho_T = 1 - c/T$ and $\mathbf{X}_T \overset{a}{\sim} \mathcal{N}\left(0, T^{-1}\sigma^2 \mathbf{\Omega}_{XX}^{LTU}(c)\right)$. Thus, with $\mathbf{X}_T^s = \mathbf{X}_T / \sqrt{\mathbf{X}_T' \mathbf{X}_T}$, from (38) the large-sample likelihood is

$$f\left(\mathbf{X}_T^s | c\right) \propto \left|\mathbf{\Omega}_{XX}^{LTU}(c)\right|^{-1/2} \left(\mathbf{X}_T^{s\prime} \mathbf{\Omega}_{XX}^{LTU}(c)^{-1} \mathbf{X}_T^s\right)^{-q/2} \tag{40}$$

With a discrete prior for c, the posterior can then be directly computed using (34).

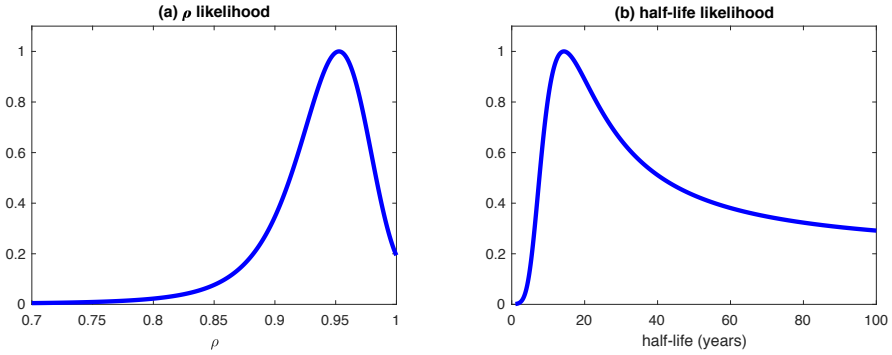
Empirical Example: An important piece of evidence underlying Rogoff's "purchasing power parity (PPP) paradox" (Rogoff (1996)) is the high degree of persistence in the real exchange rate. We apply the methods outlined in this section to learn what the data (together with priors) tell us about real exchange rate persistence in the context of the LTU model. We use the annual real \$/£ exchange rate series plotted in Figure 1; the sample period is

1791-2016 ($T = 226$ years), and the analysis uses $q = 22$ cosine transforms that summarize variability for periods longer that $2T/q \approx 20$ years. The local-to-unity persistence parameter is c, but two transformations of c are more easily interpreted. The first is $\rho = 1 - c/T$ which can be interpreted as an AR(1) parameter in a model with innovation variance σ^2. In the LTU model the jth autocorrelation is approximately $e^{-c(j/T)} \approx \rho^j$, so ρ can be used to describe long-horizon LTU persistence using familiar AR(1) low-frequency dynamics. The second parameter is a version of half-life, that is, the value of h such that $cor(x_t, x_{t+h}) = 1/2$, where in the AR(1) model, h measures the half-life of an innovation. In the LTU model, with its focus on low-frequency persistence, the half-life is defined as the value of h that solves $e^{-c(h/T)} = 1/2$. We stress that, although both h and ρ are interpretable in terms of $AR(1)$ dynamics, the LTU model does not assume that x_t follows an $AR(1)$ process; rather it assumes that, over long horizons, autocorrelations decay at the same rate as a highly persistent AR(1) model.

With this background, Figure 5 plots the likelihood (40) as a function of ρ and as a function of the half-life h. The likelihood places little weight on values of $\rho < 0.9$, or equivalently $h < 7$ years. That said, the likelihood is relatively uninformative otherwise. Thus, the prior will play an important role. To illustrate this, panel (b) of Table 1 summarizes the posterior associated with two priors. The first prior is $\rho \sim \mathcal{U}(0.5, 0.999)$, which puts less than 10% of its mass on values of $h > 13$ years, while the second is $h \sim \mathcal{U}(0, 100)$, which puts more than half its weight on values of $\rho > 0.986$; in both cases the prior is approximated with an equally spaced discrete grid of 200 points.[15] These two priors lead to substantively different posteriors. This is a generic feature of low-frequency Bayes analysis: because effective sample sizes are small, prior and sample information are both important for posterior inference. That said, in this example, both priors yield posteriors with small probability for $h < 10$ years, and in this sense provide evidence for Rogoff's PPP-puzzle.[16]

[15] Because $c = T(1-\rho) = -Tln(1/2)/h$, these different priors for ρ and h imply different priors for c.

[16] Empirical results in Müller and Dou (2018) suggest that the low-frequency AR(1) decay implied the LTU model understate the half-life for real exchange rates. Specifically, they develop a generalized LTU model that allows for ARMA$(p, p-1)$ low-frequency dynamics, and find that the $p \geq 2$ models fit better than the $p = 1$ model and imply a larger value of h.



Notes: Values are relative to the maximum.

Figure 5: Likelihood values for the local-to unity AR(1) and half-life parameters

4.2.2 Low-Frequency Persistence in the Fractionally Integrated Model

In the fractional $I(d)$ model, $\mathbf{X}_T \overset{a}{\sim} \mathcal{N}\left(0, T^{-1}\sigma^2 \mathbf{\Omega}_{XX}^{FR}(d)\right)$. With this change of $\mathbf{\Omega}$, the posterior for d can be computed in the same way as the posterior for c was computed in the LTU model.

Empirical Example: A large literature documents long-memory persistence in asset return volatility (e.g., Ding et al. (1993), Baillie (1996), Andersen and Bollerslev (1997) and Andersen et al. (2003)). We apply the methods outlined here to the logarithm of daily realized volatility plotted in panel (c) of Figure 1. This series is available for 4738 trading days from January 3, 2000 through November 14, 2018 and we use $q = 37$ to capture periods of 256 trading days (approximately one year) or longer. A flat prior, $d \sim \mathcal{U}(-0.4, 1.4)$, approximated by a discrete grid with 200 equally spaced points, yields the posterior summarized in panel (c) of Table 1. The 90% equal-tailed credible interval is $d \in (0.36, 0.82)$.

4.2.3 Low-Frequency Persistence in the Local-Level Model

In the local-level model, x_t is the sum of independent $I(0)$ and $I(1)$ processes (see (4)). Using the parameterization for the $I(1)$ component in (5), $\mathbf{X}_T \overset{a}{\sim} \mathcal{N}\left(0, T^{-1}\sigma^2 \mathbf{\Omega}_{XX}^{LL}(g)\right)$. The Gibbs algorithm outlined above can be used to obtain draws from the posterior of (σ, g).

Empirical Example. Section 3.4 presented confidence and credible intervals for the mean growth rate of TFP in the U.S. constructed using an $I(0)$ model. But an $I(0)$ model, with its constant mean, is ill suited for describing the long swings in TFP growth rates over the

past 150 years (Gordon (2016)) or the 1970 productivity slowdown (Nordhaus (1972)) in the post-WWII period. The local-level model's $I(1)$ component can be used to capture these long swings. This raises two question: first, by how much is this $I(1)$ component expected to change over, say, a decade, and second, what has been the historical evolution of this component? We answer the first question here, and tackle the second question in the next subsection.

Using the notation in (5), the $I(1)$ component is denoted by b_t, and the long-run standard deviation of the innovations is $\sigma_{\Delta b} = (g/T)\sigma$. Over a long span of h periods the standard deviation of $b_{t+h} - b_t$ is approximately equal to $\sqrt{h}\sigma_{\Delta b}$. Panel (d) of Table 1 summarizes the posterior for σ, g, and $\sigma_{\Delta b}$ using the TFP growth rate data and an uninformative prior for σ. The prior for g is informative and reflects an *a priori* belief that changes in b_t are likely to be small relative the overall variability in TFP growth rates: $\ln(g) \sim \mathcal{U}(\ln(0.1), \ln(500))$, approximated with a finite grid. The resulting posterior mean is $E[\sigma_{\Delta b}|\mathbf{X}_T] = 0.067$, which corresponds to a standard deviation of decadal ($h = 40$-quarter) changes in the level of the growth rate of TFP of $\sqrt{40} \times 0.067 = 0.42$. (To put this value in perspective, recall that TFP growth rates are measured in percentage points per year and the sample mean was 1.24 over the post-WWII period). The equal-tailed 68% credible set is $\sqrt{40}\sigma_{\Delta b} \in (0.05, 0.74)$, which implies considerable uncertainty about the variability in the level of TFP growth rates.

4.3 Inference about the Low-Frequency Level of a Time Series

Section 3 took up the problem of inference about the mean of a $I(0)$ time series. In this section we present methods for inference about the level when the stochastic process is not $I(0)$. We consider three examples. In the first, x_t is stationary (so the level is its mean), but is highly persistent as in the LTU model. In the second example we return to the $I(0)$ process, but allow the mean to have a discrete break at an unknown date. Here, interest focuses on the pre- and post-break values of the mean together with the break date. In the third example, x_t follows the local-level model (4) and interest focuses on the value of the $I(1)$ component b_t, the local-(in time)-level of the series. This is a low-frequency signal extraction problem with unknown model parameters. In each of these examples, the posterior is computed using a variant of the Gibbs algorithm outlined above.

4.3.1 Inference about the Mean in a Highly Persistent Stationary Process

This example returns to the stationary LTU model, but now interest focuses on the value of μ. The large sample likelihood is

$$\mathbf{X}_T^0 = \begin{bmatrix} \bar{x}_{1:T} \\ X_T \end{bmatrix} \overset{a}{\sim} \mathcal{N}\left(\iota_{q+1}\mu, T^{-1}\sigma^2 \mathbf{\Omega}^{LTU}(c)\right) \tag{41}$$

where $\iota_{q+1} = \begin{bmatrix} 1 & \mathbf{0}_q' \end{bmatrix}'$. Using a normal prior for μ, an inverse-Gamma prior for σ^2, and a discrete prior for c, the Gibbs algorithm using (27)-(34) yields draws from the posterior.

Empirical Example. The U.S. unemployment rate plotted in panel (d) of Figure 1 shows large low-frequency swings around a population mean that, arguably, was constant over the post-WWII period. The methods outlined above can be applied to learn about this mean. We use uninformative priors for μ and σ, but an informative prior for c that puts much of its mass on small values of c (or equivalently a prior for the AR(1) coefficient $\rho = 1 - c/T$ with most of its mass near $\rho = 1$). Specifically the prior is $\ln(c) \sim \mathcal{U}(\ln(0.1), \ln(500))$, approximated by a 200-point grid. Selected posterior quantiles are given in panel (e) of Table 1. The posterior indicates large uncertainty about the persistence parameter ρ, with a 68% credible set that ranges from 0.78 to 0.99. Comparing the LTU posterior for μ to its $I(0)$ counterpart (also shown in the table), shows that the LTU posterior is more spread out and exhibits a left skew. This skew arises from the uncertainty about ρ: large values of ρ lead to a disperse posterior for μ centered near $(x_1 + x_T)/2$; small values of ρ lead to a more concentrated posterior with $\bar{x}_{1:T}$ as its mean; and the skew arises because $(x_1 + x_T)/2 < \bar{x}_{1:T}$ in the post-WWII sample.

4.3.2 Inference about a Discrete Break in the Mean of an $I(0)$ Process

In this exercise

$$x_t = \mu_t + u_t \tag{42}$$

with

$$\mu_t = \mu + \mathbf{1}[t > rT]\delta + u_t \tag{43}$$

where μ and $\mu + \delta$ are the pre- and post-break values of μ_t and $\lfloor rT \rfloor$ is the break date, with $0 < r < 1$ the break date expressed as a fraction of the sample size.[17] The error term u_t is assumed to be $I(0)$. There are now four parameters governing the low frequency behavior of x_t, (μ, δ, σ, r). With $\mathbf{d}_T = [\mathbf{0}'_{[rT]}, \mathbf{l}'_{T-[rT]}]'$, the time series for μ_t can be written as $\mu_{1:T} = \mathbf{l}_T\mu + \mathbf{d}_T\delta$. Recalling the notation introduced in Section 3, we now have

$$\mathbf{X}_T^0 = T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{x}_{1:T} = \iota_{q+1}\mu + T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{d}_T(r)\delta + T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{u}_{1:T},$$

which yields the large-sample likelihood

$$\mathbf{X}_T^0 \overset{a}{\sim} \mathcal{N}\left(\iota_{q+1}\mu + \mathbf{v}^0(r)\delta, T^{-1}\sigma^2\mathbf{\Omega}^{I(0)}\right) \tag{44}$$

with $\mathbf{v}^0(r)$ the large-T limit of $T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{d}_T(r)$. Analogous to the discussion at the end of Section 3.2, $\mathbf{v}^0(r)$ may be conveniently approximated by computing $T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{d}_T(r)$ for a large value of T, such as $T = 1000$.

The posterior for (μ, δ, σ, r) can be formed by sequentially drawing (μ, δ) conditional on (σ, r) using (27) (or, under a flat prior on (μ, δ), using (28)), drawing σ^2 conditional on (μ, δ, r) using (30), and drawing r conditional on (μ, δ, σ) using a discrete prior and (34).

Empirical Example: We apply this model to TFP growth rates using an uninformative prior for σ and (δ, μ) and with $r \sim \mathcal{U}(0, 1)$, approximated using a discrete grid with $T - 1$ points, so each break date was equally likely. Results are summarized in Figure 6 and panel (f) of Table 1. The posterior points to the 1970s productivity slowdown as a likely break with a large fall in the value of μ_t: the 68% credible set for δ is $\delta \in (-1.74, -0.80)$.

4.3.3 Inference about the 'Local-Level' in the Local-Level Model

In this exercise x_t is assumed to follow the local-level model (4), that is $x_t = \mu + b_t + e_t$, where b_t is $I(1)$ and e_t is $I(0)$. The local-in-time level of x_t is $\mu + b_t$. The focus of this section's exercise is $\mathbf{l}_T\mu + \hat{\mathbf{b}}_{1:T}$ the realization of the low-frequency value of the local-level over the sample period. With $\mathbf{B}_T = T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{b}_{1:T}$, we have $\hat{\mathbf{b}}_{1:T} = \mathbf{\Psi}_T^0\mathbf{B}_T$, so the posterior for $\mathbf{l}_T\mu + \hat{\mathbf{b}}_{1:T}$ can be recovered from the posterior for $(\mu, \mathbf{B}_T)$.

There a variety of ways to compute this posterior; here is one. The unknown quantities are $(\mathbf{B}_T, \mu, \sigma, g)$. The joint posterior density can be factored as $p\left(\mathbf{B}_T, \mu, \sigma, g|\mathbf{X}_T^0\right) =$

[17] Early econometric papers about inference about the break date include Bai (1994), Bai and Perron (1998) amd Bai et al. (1998); see Perron (2006) for a survey and additional references.

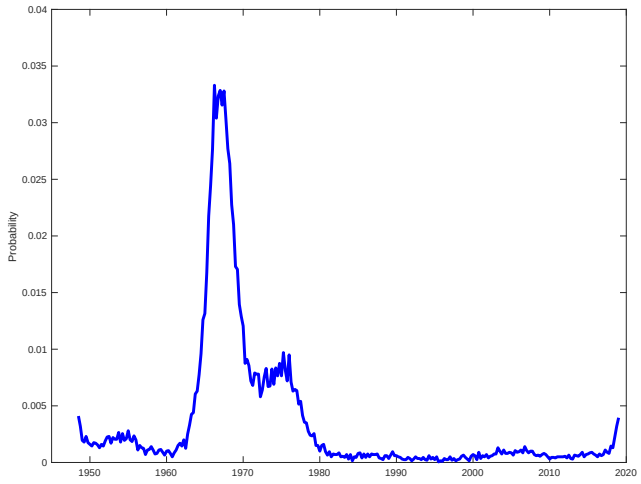


Figure 6: Break date posterior for the mean of TFP growth rates in the $I(0)$ model

$p\left(\mathbf{B}_T \left| \mathbf{X}_T^0, \mu, \sigma, g \right.\right) p\left(\mu, \sigma, g \left| \mathbf{X}_T^0 \right.\right)$. Thus, a draw from the joint posterior can be obtained by first obtaining a draw from $(\mu, \sigma, g)|\mathbf{X}_T^0$, and then conditioning on the value of (μ, σ, g) to obtain a draw from $\mathbf{B}_T \left| \mathbf{X}_T^0, \mu, \sigma, g \right.$. We discuss each of these in turn.

- $(\mu, \sigma, g) \left| \mathbf{X}_T^0 \right.$: Draws can be obtained from the 3-step Gibbs algorithm from Section 4.1.2 used earlier for the LTU model in Section 4.3.1, but with $\mathbf{\Omega}^{LL}(g)$ replacing $\mathbf{\Omega}^{LTU}(c)$ in (41).

- $\mathbf{B}_T \left| \mathbf{X}_T^0, \mu, \sigma, g \right.$: Note that $x_t - \mu - b_t = e_t \sim I(0)$, and $\mathbf{X}_T^0 = \iota_{q+1}\mu + \mathbf{B}_T + \mathbf{E}_T$, with $\mathbf{E}_T = T^{-1}\mathbf{\Psi}_T^0 \mathbf{e}_{1:T}$. Thus

$$\begin{bmatrix} \bar{x}_{1:T} - \mu \\ \mathbf{X}_T \end{bmatrix} = \mathbf{X}_T^0 - \iota_{q+1}\mu \left| (\mathbf{B}_T, \mu, \sigma, g) \right. \overset{a}{\sim} \mathcal{N}\left(\mathbf{B}_T, T^{-1}\sigma^2 \mathbf{\Omega}^{I(0)}\right). \tag{45}$$

  Furthermore, $b_t \sim I(1)$, so

$$\mathbf{B}_T \left| (\mu, \sigma, g) \right. \overset{a}{\sim} \mathcal{N}\left(0, T^{-1}\sigma^2 g^2 \mathbf{\Omega}^{I(1)}\right). \tag{46}$$

  Thus, $\mathbf{B}_T \left| \mathbf{X}_T^0, \mu, \sigma, g \right.$ follows the normal distribution in (27) using (45) as the likelihood and (46) as the prior.

Empirical Example: The last subsection found a single large break in the level of the TFP growth rate data, and the model outlined here can be used to get a more nuanced picture

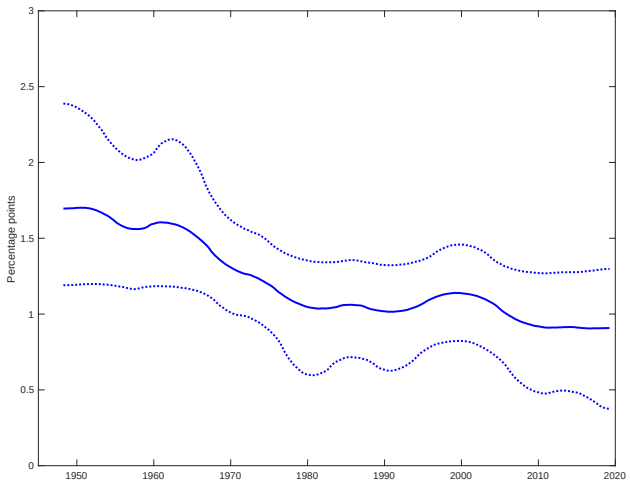


Figure 7: Low-frequency level component for the growth rate of TFP in the LL model

of the evolution of TFP. Figure 7 shows the posterior median and (pointwise) 68% credible bands for the low-frequency local-level using the same priors used in Section 4.2.3. The posterior is largely consistent with the single break model of the last section and suggests a productivity slowdown in the late 1960s and 1970s. The fall in the level is somewhat less than in the discrete break model; presumably this reflects the prior for g which puts relatively more weight on models with little time variation in b_t.

4.4 Inference about Low-Frequency Covariation

With $\mathbf{x}_t$ an $n \times 1$ vector of times series, Section 3 defined low-frequency covariation (σ_{ij}^{LF}) between $x_{i,t}$ and $x_{j,t}$ as the population covariance between the low-frequency trends, $\hat{x}_{i,t}$ and $\hat{x}_{j,t}$. And, with $\mathbf{X}_T^0$ a $q \times n$ matrix of low-frequency averages, the section showed that $\sigma_{ij}^{LF} \propto tr(\mathbf{V}_{ij})$ where $\mathbf{V}_{ij}$ is the large-sample covariance matrix of $\mathbf{X}_{i,T}^0$ and $\mathbf{X}_{j,T}^0$. This subsection discusses Bayes inference about these low-frequency covariances and related correlation and linear regression parameters.

Multivariate inference requires a multivariate model. Section 2 presented five widely-used univariate models of low-frequency variability, and these were used in the examples above. While it is relatively straightforward to extend these models for multivariate time series, there are two important challenges. The first is conceptual: A multivariate model needs to describe

the persistence of each of the univariate series in $\mathbf{x}_t$, but also any linear combination of the univariate series. For example, each series might be $I(1)$, but a particular linear combination might be $I(0)$; that is, the variables might be cointegrated (Engle and Granger (1987)). The second challenge is practical: as we have stressed, low-frequency analysis is a small-sample statistical problem, so parsimony is important. The five univariate models were extremely parsimonious, with each requiring at most three parameters, a mean (μ), a scale (σ), and a single persistence parameter (d, c, or g). Generically, multivariate models increase the number parameter by a $O(n^2)$ factor, seriously taxing the small-sample information.

The Bayes analysis in this section starts with the multivariate $I(0)$ model which can be analyzed using standard results for *i.i.d.* normal samples. We then consider a low-frequency factor model, which models the n times series in terms of k common factors and n series-specific disturbances. With k small, this model is relatively parsimonious even for large n.

4.4.1 Low-Frequency Covariance in the $I(0)$ Model

In the $I(0)$ model, $\mathbf{V} = \mathbf{\Sigma} \otimes \mathbf{I}_{q+1}$, where the $n \times n$ matrix $\mathbf{\Sigma} = [\sigma_{ij}]$ is the long-run covariance matrix for $\mathbf{x}_t$. This Kronecker structure simplifies the calculations: the low-frequency covariance of $\hat{x}_{i,t}$ and $\hat{x}_{j,t}$ is $\sigma_{ij}^{LF} = tr(\mathbf{V}_{ij}) = (q+1)\sigma_{ij}$. The task then is simply to compute the posterior of the long-run covariance matrix. With a flat prior on μ, the posterior can be computed from $\mathbf{X}_T$ data with large sample likelihood

$$vec(\mathbf{X}_T) \overset{a}{\sim} \mathcal{N}\left(0, T^{-1}\mathbf{\Sigma} \otimes \mathbf{I}_q\right). \tag{47}$$

With an inverse Wishart prior for $\mathbf{\Sigma}$, the posterior is given by (32). The posteriors for the long-run correlations (ρ_{ij}^{LF}), regression coefficients (β_{ij}^{LF}), etc., follow directly.

Empirical Example. Suppose that the TFP and per-capita GDP growth rates plotted in Figure 2 follow an $I(0)$ model. Panel (f) of Table 1 summarizes the posterior for ρ^{LF} and β^{LF}, the low-frequency regression coefficient of GDP onto TFP, where the posterior uses a flat prior for μ and an uninformative prior for $\mathbf{\Sigma}$ (corresponding to $\nu = 0$ in (31)). If the log-levels of TFP and GDP were $I(1)$ and cointegrated, $\mathbf{\Sigma}$ would be singular and $\rho^{LF} = 1$. The posterior in Table 1 shows that the correlation is large (the posterior mean is $\rho^{LF} = 0.80$), but markedly different than unity. As for β^{LF}, the one-sector neoclassical growth model provides a benchmark of $\beta^{LF} \approx 1/(1-\alpha)$, where α is labor's share of aggregate income, so

the benchmark yields $\beta^{LF} \approx 1/(1-2/3) = 1.5$. The $I(0)$ posterior suggests values of β^{LF} lower than this benchmark value.

4.4.2 A Low-Frequency Factor Model

Consider the following factor model for $\mathbf{x}_t$

$$\mathbf{x}_t = \mu + \lambda \mathbf{f}_t + \mathbf{e}_t \tag{48}$$

where $\mathbf{f}_t$ denotes the unobserved common factors, λ denotes the factor loadings, and $\mathbf{e}_t$ denotes a vector of mutually independent errors (sometimes called *uniquenesses*) that capture the residual variability in the series. The specifics of the model depend on the number of factors in $\mathbf{f}_t$ and the stochastic processes for $\mathbf{f}_t$ and $\mathbf{e}_t = [e_{1,t}, \ldots, e_{n,t}]'$. The assumptions used here are motivated by the empirical example discussed below; specifically, $\mathbf{f}_t = f_t$ is a scalar that follows the local-level model, $e_{j,t}$ follows an $I(d_j)$ model with scale parameter σ_{e_j} and $\{f_t, e_{1,t}, e_{2,t}, \ldots, e_{n,t}\}$ are independent. (Modifying these assumptions to accommodate other processes is straightforward.) The parameters are thus $(\lambda, \mu, \sigma_e, \mathbf{d}, \sigma_f, g)$ with $\sigma_e = (\sigma_{e_1}, \ldots, \sigma_{e_n})$ and $\mathbf{d} = (d_1, \ldots, d_n)$.

The posterior is easily computed using a Gibbs algorithm. To begin, $\mathbf{X}_T^0$ can decomposed as

$$\mathbf{X}_T^0 = \iota_{q+1}\mu' + \mathbf{F}_T\lambda' + \mathbf{E}_T \tag{49}$$

where $\mathbf{F}_T = T^{-1}\mathbf{\Psi}_T^{0\prime}\mathbf{f}_{1:T}$ and similarly for $\mathbf{E}_T$. The local-level model for f_t yields

$$\mathbf{F}_T \overset{a}{\sim} \mathcal{N}\left(0, T^{-1}\sigma_f^2 \mathbf{\Omega}^{LL}(g)\right). \tag{50}$$

The $I(d_j)$ model for $e_{j,t}$ implies that the jth column of $\mathbf{E}_T$ satisfies

$$\mathbf{E}_{j,T} \overset{a}{\sim} \mathcal{N}\left(0, T^{-1}\sigma_{e_j}^2 \mathbf{\Omega}^{FR}(d_j)\right). \tag{51}$$

Independence then yields

$$vec(\mathbf{X}_T^0 - \iota_{q+1}\mu') \left| \left(\mathbf{F}_T, \lambda, \mu, \sigma_e^2, d\right) \overset{a}{\sim} \mathcal{N}\left(vec\left(\mathbf{F}_T\lambda'\right), T^{-1}\mathbf{V}(\sigma_e, \mathbf{d})\right)\right. \tag{52}$$

where $\mathbf{V}(\sigma_e, \mathbf{d})$ is a block diagonal matrix with jth block given by $\sigma_{e_j}^2 \mathbf{\Omega}^{FR}(d_j)$ (from (51)).

This structure leads to a 3-step Gibbs algorithm:

1. Draw from $\mathbf{F}_T \,|(\mathbf{X}_T^0, \lambda, \mu, \sigma_e, \mathbf{d}, \sigma_f, g)$. From (27) this is normal using (52) as the likelihood and (50) as the prior. (Note: $vec(\mathbf{F}_T\lambda') = (\lambda \otimes \mathbf{I}_q)\mathbf{F}_T = (\mathbf{I}_n \otimes \mathbf{F}_T)\lambda$.)

2. Draw from $(\sigma_f, g, \mu, \sigma_e, \mathbf{d}) \,|(\mathbf{X}_T^0, \mathbf{F}_T, \lambda)$. This is done is two steps:

   (a) Draw $(\sigma_f, g) \,|\mathbf{F}_T$. This is a draw from the posterior of a univariate local-level model as described in Section 4.2.3.

   (b) Draw from $(\mu, \sigma_e, \mathbf{d}) \,|(\mathbf{X}_T^0, \mathbf{F}_T, \lambda)$. With $\mathbf{Y}_{j,T} = \mathbf{X}_{j,T}^0 - \mathbf{F}_T\lambda_j = \iota_{q+1}\mu_j + \mathbf{E}_{j,T}$, this is a draw from $(\mu_j, \sigma_{e_j}, d_j) \,|\mathbf{Y}_{j,T}$ for $j = 1, \ldots, n$, each a draw from the posterior of a univariate fractional model as described in Section 4.1.2.

3. Draw from $\lambda|(\mathbf{X}_T^0, \mathbf{F}_T, \mu, \sigma_e, \mathbf{d}, \sigma_f, g)$. Using a normal prior, $\lambda \sim \mathcal{N}(\mu_\lambda, \mathbf{\Sigma}_\lambda)$ and the likelihood (52), this is a draw from the normal posterior (27).

Empirical Example. We use a LL-factor model with a single factor and $I(d_j)$ uniquenesses to describe the low-frequency properties of the five growth rates (TFP, GDP, consumption, investment, labor compensation) plotted in Figure 2. In the factor model, the scale of f_t and λ are not separately identified and we normalize the factor loading λ_1 on TFP to unity. We posit independent priors for the remaining parameters: the prior for μ_j is flat, $\lambda_j \sim \mathcal{N}(1.5, 4)$, $\sigma_j^2 \sim \mathcal{IG}(0.1, 4)$, $\ln(g) \sim \mathcal{U}\left(\ln(0.1), \ln(500)\right)$ approximated by a discrete grid of 101 equally spaced points, $d_j \sim \mathcal{U}(-0.49, 0.49)$ approximated using 101 grid points, and the prior for σ_f is uninformative.[18] Selected posterior results are summarized in Table 2. The factor loading for GDP is larger than the low-frequency regression coefficient found in the last subsection, consistent with a measurement error interpretation of the uniquenesses in the factor model. The negative values of d_j suggest a tendency for the log-levels of the variables to revert to the local linear trend generated by the factor and (in the case of GDP and labor compensation) lead to a high low-frequency correlation between the factor and the series. The values of σ_f and g suggest that the factor is somewhat less variable and persistent than in the previously estimated univariate LL model for TFP.

A more ambitious empirical exercise is described in Müller et al. (2019) who use a version of the factor model to model the long-run evolution of per-capita GDP for the 112 countries plotted in Figure 3. The 112-country exercise includes complications not present in the

[18]The priors for σ_j keep the estimated variances away from zero avoiding the so-called 'Heywood' case.

Parameter/ Variable	Posterior Mean	Posterior Quantiles				
		0.05	**0.17**	**0.50**	**0.83**	**0.95**
(a) Factor Loadings						
TFP	1	1	1	1	1	1
GDP	2.11	1.46	1.67	2.03	2.55	3.02
Consumption	1.67	1.09	1.28	1.61	2.06	2.47
Investment	2.52	0.96	1.58	2.47	3.46	4.25
Lab. Comp.	2.58	1.74	2.01	2.49	3.14	3.73
(b) d						
TFP	0.20	-0.12	0.03	0.22	0.37	0.44
GDP	-0.25	-0.48	-0.43	-0.28	-0.07	0.08
Consumption	0.13	-0.23	-0.05	0.16	0.31	0.40
Investment	-0.20	-0.47	-0.41	-0.24	0.02	0.22
Lab. Comp.	-0.23	-0.47	-0.42	-0.26	-0.03	0.16
(c) σ_e						
TFP	1.50	0.41	0.55	1.00	2.21	4.24
GDP	1.69	0.37	0.55	1.19	2.80	4.72
Consumption	1.44	0.34	0.48	0.90	2.11	4.51
Investment	26.81	4.47	9.32	23.83	44.60	58.63
Lab. Comp.	5.58	1.07	2.11	5.01	9.00	11.99
(c) LLM Factor parameters						
σ_f	2.6	1.4	1.8	2.5	3.3	4.0
g	3.9	0.13	0.22	1.2	5.0	11.9

Table 2: Posterior mean and quantiles for selected parameters in the 5-variable LL-Factor model

five-variable model. For example, the panel data set is unbalanced, and the country-specific terms (the analogues of $e_{i,t}$ in (48)) are correlated within small groups of countries. Strategies for handling these complications are presented in Müller et al. (2019). An interesting complication arises in the prior selection for the many country-specific parameters in the model: In contrast to a model with small n, individually relatively uninformative priors can become quite informative in the aggregate when n is large. For a concrete example consider the prior $\lambda_j \sim i.i.d.\mathcal{N}(1.5, 4)$ independent prior for $j = 2, \ldots, n$ of the factor model outlined above. This implies a prior on the average $\bar{\lambda} = (n-1)^{-1} \sum_{j=2}^{n} \lambda_j$ equal to $\mathcal{N}(1.5, 4/(n-1))$. For $n = 5$, as above, this is still a fairly wide prior. But for $n = 112$, this implies a very tight prior for the average factor loading. For this reason, Müller et al. (2019) employ a hierarchal prior as discussed in Section 6.6.

4.5 Predictive Distributions for Long-Horizon Predictions

In this section we take up the problem of predicting the value of $\bar{\mathbf{x}}_{T+1:T+h}$, the average value of $\mathbf{x}_t$ from time period $T+1$ through time $T+h$, using the sample data $\mathbf{x}_{1:T}$. The goal is

to find the predictive distribution for $\bar{\mathbf{x}}_{T+1:T+h}$. We impose two constraints on the problem that make it amenable to large-sample normal approximations. First, we consider problems in which h is of the same order as T; this means that $\bar{\mathbf{x}}_{T+1:T+h}$ is approximately normally distributed. Second, we restrict the conditioning information to $\mathbf{X}_T^0$, and thus only use the low-frequency averages in the analysis.

Taken together, these constraints mean that we are interested in determining the distribution of one sample average ($\bar{\mathbf{x}}_{T+1:T+h}$) given another set of averages ($\mathbf{X}_T^0$), and Section 8 shows that

$$T^{1/2} \begin{bmatrix} vec(\mathbf{X}_T^0 - \iota_{q+1}\mu') \\ \bar{\mathbf{x}}_{T+1:T+h} - \mu \end{bmatrix} \Rightarrow \mathcal{N}\left(0, \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}\right).$$

This yields the large-sample approximation

$$\begin{bmatrix} vec(\mathbf{X}_T^0) \\ \bar{\mathbf{x}}_{1:T} \end{bmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu \otimes \iota_{q+1} \\ \mu \end{bmatrix}, T^{-1} \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}\right) \tag{53}$$

and the familiar multivariate normal conditional distribution

$$\bar{\mathbf{x}}_{T+1:T+h} \left| (\mathbf{X}_T^0, \mu, \mathbf{V}) \right. \overset{a}{\sim} \mathcal{N}\left(\mu + \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\left(vec(\mathbf{X}_T^0 - \iota_{q+1}\mu')\right), T^{-1}(\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12})\right). \tag{54}$$

If $\mu = \mu(\theta)$ and $\mathbf{V} = \mathbf{V}(\theta)$ depend on an unknown parameter θ, then the predictive distribution becomes a mixture of the normal distribution in (54), with mixing weights equal to the posterior distribution of θ given the observation $\mathbf{X}_T^0$. Thus, predictive distributions can be obtained by augmenting the Gibbs sampler for θ by an additional step that draws $\bar{\mathbf{x}}_{T+1:T+h}$ from (54) given the current value of θ.

A leading case is the univariate $I(0)$ model with $\mathbf{V}_{11} = \sigma^2\mathbf{I}_{q+1}$, $\mathbf{V}_{22} = (h/T)\sigma^2$ and $\mathbf{V}_{12} = 0$, with posterior for $\theta = (\mu, \sigma^2)$ under uninformative priors. In this case, the predictive distribution is Student-t with q degrees of freedom, location $\bar{x}_{1:T}$, and scale $\sqrt{(h^{-1} + T^{-1})\, s^2}$ with $s^2 = T\mathbf{X}_T'\mathbf{X}_T/q$, that is

$$\frac{(\bar{x}_{T+1:T+h} - \bar{x}_{1:T})}{\sqrt{(h^{-1} + T^{-1})\, s^2}} \overset{a}{\sim} \text{Student-}t_q. \tag{55}$$

See Gelman et al. (2004), page 77.

Variable	Forecast Horizon (Years)	Mean	Quantiles				
			0.05	0.17	0.50	0.83	0.95
(a) $I(0)$ model							
TFP	25	1.21	0.42	0.76	1.21	1.65	1.99
	50	1.21	0.59	0.85	1.21	1.56	1.83
GDP	25	1.80	0.74	1.20	1.80	2.41	2.87
	50	1.80	0.96	1.32	1.80	2.28	2.65
(b) LL-Factor model							
TFP	25	1.02	0.02	0.48	1.04	1.57	1.96
	50	1.05	0.08	0.56	1.07	1.54	1.92
GDP	25	1.62	0.16	0.87	1.66	2.35	2.94
	50	1.61	0.10	0.93	1.67	2.29	2.84
(c) LL-Factor model, conditional on $\bar{x}_{TFP,T+1:T+h} = 1.21$							
GDP	25	1.81	0.91	1.32	1.81	2.30	2.69
	50	1.79	1.00	1.37	1.79	2.22	2.58

Table 3: Mean and selected quantiles for predictive distributions

4.5.1 Two Examples

$I(0)$ **Model:** The first example uses the univariate $I(0)$ model to form the predictive distributions for average TFP growth over the next 25 and 50 years. TFP is measured quarterly, so these correspond to $h = 100$ and 200 quarters. The sample mean is $\bar{x}_{1:T} = 1.21$ and $s = 3.83$ with $q = 14$. Plugging these values into (55) yields the predictive distribution for $\bar{x}_{T+1:T+h}$. Selected quantiles are shown in panel (a) of Table 3. Writing $\bar{x}_{T+1:T+h} = \mu + (\bar{x}_{T+1:T+h} - \mu)$ highlights the two independent sources of uncertainty about $\bar{x}_{T+1:T+h}$ in the $I(0)$ model: the first is uncertainty about μ with $\mu|(\mathbf{X}_T^0, \sigma) \overset{a}{\sim} \mathcal{N}(\bar{x}_{1:T}, \sigma^2/T)$, and the second is uncertainty about the future with $(\bar{x}_{T+1:T+h} - \mu)|(\mathbf{X}_T^0, \sigma) \overset{a}{\sim} \mathcal{N}(0, \sigma^2/h)$. The first source of uncertainty does not depend on the forecast horizon and the second falls h increases. Thus, as evident in the table, the predictive density narrows as the forecast horizon increases from 25 to 50 years.

LL-Factor Model: The second example uses the 5-variable local-level factor model to construct joint predictive distributions for each of the five variables, again over 25-year and 50-year horizons. Panel (b) of Table 3 summarizes the marginal predictive distributions for average TFP and GDP growth rates. Comparing these predictive distributions to those from the $I(0)$ model for TFP highlights three features. First, the factor model includes persistent local-level and $I(d)$ components, and the factor model implies more uncertainty about the

values of $\bar{\mathbf{x}}_{T+1:T+h}$ than the $I(0)$ model. Second, because average growth rates decreased over the sample period, the LL factor model extrapolates this slower growth into the future, so the mean of the predictive density is lower. Finally, in the $I(0)$ model, the predictive distribution narrows as the forecast horizon increased. In the LL model, as the horizon increases, uncertainty about the average value of the $I(0)$ component falls, but uncertainty about the $I(1)$ component increases, and the predictive distributions may first narrow and then increase as the forecast horizon increases.

An advantage of multivariate model is that it produces multivariate predictive distributions that can used for conditional forecast exercises. Panel (c) of Table 3 summarizes one such exercise, and shows the predictive distribution of the growth of GDP conditional on TFP growth taking on its sample average value over the forecast period, that is, conditional on $\bar{x}_{TFP,T+1:T+h} = 1.21$.

5 Frequentist Inference: Examples

This section uses several examples to illustrate frequentist inference about low-frequency features of economic time series. As in the Bayes section, inference involves the mean and covariance matrix of a small sample of approximately normal observations, $\mathbf{X}_T^0$. Section 7 discusses the statistical foundation of frequentist inference in this context. Here we discuss the application of frequentist methods to a subset of the inference problems considered in the Section 4. We begin with frequentist analysis for the $I(0)$ model.

5.1 Frequentist Inference in the $I(0)$ Model

In the multivariate $I(0)$ model, $vec(\mathbf{X}_T^0) \overset{a}{\sim} \mathcal{N}\left(\mu \otimes \iota_{q+1}, T^{-1}\mathbf{\Sigma} \otimes \mathbf{I}_{q+1}\right)$, where $\mathbf{\Sigma}$ is the long-run covariance matrix of the $\mathbf{x}_t$ process. Section 3.4 considered both Bayes and frequentist inference about the mean. As discussed there, frequentist inference coincides with flat prior Bayes inference.

Here we focus on $\mathbf{\Sigma}$. Imposing location invariance amounts to dropping $\bar{\mathbf{x}}_{1:T}$ from the analysis, so inference relies on $\mathbf{X}_T$, where $vec(\mathbf{X}_T) \overset{a}{\sim} \mathcal{N}\left(0, T^{-1}\mathbf{\Sigma} \otimes \mathbf{I}_q\right)$. Recall from Section 3.3 that the covariance matrix of the low-frequency trends, $\hat{\mathbf{x}}_t$ was denoted by $\mathbf{\Sigma}^{LF}$, and from Section 4.4.1 that $\mathbf{\Sigma}^{LF} \propto \mathbf{\Sigma}$ in the $I(0)$ model. Thus, the low-frequency correlations

(ρ^{LF}) and regression coefficients (β^{LF}) coincide with correlations and regression coefficients implied by $\mathbf{\Sigma}$.

Letting $\mathbf{X}'_{j,T}$ denote the j-th row of $\mathbf{X}_T$, that is the j-th cosine weighted average of each of the n series in $\mathbf{x}_t$, then $\sqrt{T}\mathbf{X}_{j,T} \overset{a}{\sim} i.i.d.\mathcal{N}(0, \mathbf{\Sigma})$ for $j = 1, ..., q$. Thus, inference about $\mathbf{\Sigma}$ (or $\mathbf{\Sigma}^{LF}$) involves the covariance matrix from a multivariate sample with q $i.i.d.$ zero-mean normal observations. This is a standard problem in multivariate analysis (e.g., Anderson (1984)).

In particular, let $\mathbf{S} = (T/q)\mathbf{X}'_T\mathbf{X}_T$ denote the sample second moment matrix. Then, $q\mathbf{S} \overset{a}{\sim} \mathcal{W}(\mathbf{\Sigma}, q)$, the Wishart distribution with covariance matrix $\mathbf{\Sigma}$ and q degrees of freedom. $\mathbf{S}$ is the MLE of $\mathbf{\Sigma}$, and $\hat{\rho}_{ij} = \mathbf{S}_{ij}/\sqrt{\mathbf{S}_{ii}\mathbf{S}_{jj}}$ is the MLE for the associated correlation coefficient, ρ_{ij}. Confidence intervals for ρ_{ij} $(= \rho^{LF}_{ij})$ can be constructed from $\hat{\rho}_{ij}$ using the methods discussed in Anderson (1984), Section 4.2.2.

Finite-sample normal linear regression theory provides a basis for inference about the long-run regression coefficients. For example, consider the regression of one element of $\hat{\mathbf{x}}_t$, denoted by $\hat{y}_t$, on a subset of others elements, denoted $\hat{\mathbf{z}}_t$, where $\hat{\mathbf{z}}_t$ is $k \times 1$. Then the standard linear regression results apply: partitioning $\mathbf{X}_T$, $\mathbf{\Sigma}$ and $\mathbf{S}$ appropriately, the regression coefficients are $\beta = \mathbf{\Sigma}^{-1}_{ZZ}\mathbf{\Sigma}_{ZY}$, $\hat{\beta} = \mathbf{S}^{-1}_{ZZ}\mathbf{S}_{ZY}$ is the MLE with $\hat{\beta}|\mathbf{Z}_T \overset{a}{\sim} \mathcal{N}(\beta, \sigma^2\mathbf{S}^{-1}_{ZZ})$ where $\sigma^2 = \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YZ}\mathbf{\Sigma}^{-1}_{ZZ}\mathbf{\Sigma}_{ZY}$, $s^2 = (\mathbf{S}_{YY} - \mathbf{S}_{YZ}\mathbf{S}^{-1}_{ZZ}\mathbf{S}_{ZY})/(q-k)$ is an unbiased estimator for σ^2 with $(q-k)s^2/\sigma^2 \sim \chi^2_{q-k}$, $(\hat{\beta} - \beta)'\mathbf{S}_{ZZ}(\hat{\beta} - \beta)/s^2 \sim F_{k,q-k}$ and $\hat{\beta}_i/SE(\hat{\beta}_i) \sim$Student-$t_{q-k}$ where $SE(\hat{\beta}_i)$ is the square root of the i-th diagonal element of $s^2\mathbf{S}^{-1}_{ZZ}$.

Long-run prediction intervals are also easy to construct in the $I(0)$ model. The flat prior Bayes Student-t result (55) holds unconditionally, and therefore it can be used to construct prediction intervals that contain future realization of $\bar{\mathbf{x}}_{T+1:T+h}$ with any pre-specified probability.

Examples: Table 4 collects confidence sets for the examples considered in this section. The first panel shows the confidence sets for the mean growth rate of TFP; these coincide with the Bayes credible sets shown earlier in Table 1. The next panel shows confidence intervals for the correlation between TFP and GDP and the regression coefficient for the regression of GDP onto TFP. These can be compared to the results shown earlier in Table 1. Notice that the frequentist confidence intervals for β in Table 5 coincide with the Bayes credible intervals shown in Table 1; this is another example of the coincident of flat-prior Bayes and frequentist inference. The top panel of Table 5 contains frequentist prediction

Parameter	Coverage	
	67%	**90%**
(a) TFP Growth Rate, $I(0)$ model		
μ	0.98 to 1.43	0.81 to 1.61
(b) Bivariate TFP and GDP Growth Rates		
(i) $I(0)$ model		
ρ^{LF}	0.75 to 0.90	0.65 to 0.93
β^{LF}	1.00 to 1.41	0.85 to 1.57
(ii) Bivariate (A,B,c,d) model from Müller Watson (2018)		
ρ^{LF}	0.70 to 0.89	0.54 to 0.93
β^{LF}	0.97 to 1.43	0.80 to 1.62
(c) \$/£ Real Exchange Rate, LTU model		
(i) F: $\rho \sim$ U(0.5,1.0)		
ρ	0.89 to 0.97	0.86 to 0.99
half-life	5.7 to 24.0	4.6 to 101.8
(ii) F: $h \sim$ U(0,100)		
ρ	0.94 to 0.98	0.91 to 0.99
half-life	10.8 to 37.9	7.2 to 75.7
(d) Daily Realized Volatility, $I(d)$ model		
d	0.44 to 0.72	0.36 to 0.81
(e) TFP Growth Rate, LLM		
g	2.6 to 17.9	1.5 to 25.1
(f) Unemployment Rate, LTU model		
ρ	0.87 to 0.98	0.50 to 0.99
μ	xxxxx	xxxxx

Table 4: Confidence intervals for selected parameters

intervals for future average growth rates of TFP and GDP taken directly from the flat prior Bayes predictive distributions presented earlier in Table 3.

5.2 Frequentist Inference about Persistence

This section discusses hypothesis tests and confidence intervals for the persistence parameters in the LTU, LL and $I(d)$ models. We consider location and scale invariant procedures, so that we can treat $\mathbf{X}_T^s = \mathbf{X}_T / \sqrt{\mathbf{X}_T' \mathbf{X}_T}$ as the effective observation. From (38) the large-sample likelihood is

$$f\left(\mathbf{X}_T^s \,|\vartheta\right) \propto \left|\mathbf{\Omega}_{XX}(\vartheta)\right|^{-1/2} \left(\mathbf{X}_T^{s\prime} \mathbf{\Omega}_{XX}^{-1}(\vartheta) \mathbf{X}_T^s\right)^{-q/2} \tag{56}$$

where $\mathbf{\Omega}_{XX}$ and the parameter ϑ are model specific: ϑ equals $c, g,$ and d for the LTU, LL and $I(d)$ models, respectively.

Variable	Forecast Horizon (Years)	Coverage	
		67%	**90%**
(a) $I(0)$ model			
TFP	25	0.76 to 1.65	0.42 to 1.99
	50	0.85 to 1.56	0.59 to 1.83
GDP	25	1.20 to 2.41	0.74 to 2.87
	50	1.32 to 2.28	0.96 to 2.65
(b) (*b,c,d*) model from Müller Watson (2016) with $q = 12$			
TFP	25	-0.38 to 1.85	-0.64 to 2.10
	50	-1.31 to 2.57	-1.80 to 3.05
GDP	25	0.40 to 2.52	-0.74 to 3.30
	50	0.10 to 2.50	-1.22 to 3.74

Table 5: Prediction sets for long-run forecasts

5.2.1 Point-Optimal Tests for the $I(1)$ and $I(0)$ Models

Point optimal tests of $H_0 : \vartheta = \vartheta_0$ versus $H_1 : \vartheta = \vartheta_1$ reject the null hypothesis for large values of the likelihood ratio statistic $f(\mathbf{X}_T^s|\vartheta_1)/f(\mathbf{X}_T^s, \vartheta_0)$. Equivalently, given the form of the likelihood (56), the tests reject for large values of the ratio of the generalized sum of squares

$$\frac{\mathbf{X}_T'\mathbf{\Omega}_{XX}^{-1}(\vartheta_0)\mathbf{X}_T}{\mathbf{X}_T'\mathbf{\Omega}_{XX}^{-1}(\vartheta_1)\mathbf{X}_T}. \tag{57}$$

Two important applications are tests for the $I(1)$ null (*unit-root tests*) and tests for the $I(0)$ null (*stationarity tests*).

Unit Root ($I(1)$) Test: Unit root tests such as the well-known augmented-Dickey-Fuller test (Dickey and Fuller (1979)) or its efficient counterpart (Dufour and King (1991), Elliott et al. (1996), Elliott (1999)) are tests of $H_0 : c = 0$ in the LTU model. Implementing these tests requires choosing the number of autoregessive lags to capture the $I(0)$ dynamics in the process. Choosing too few lags results in size distortions, while choosing too many lags results in a loss of power (see Ng and Perron (1995) for discussion); this is analogous to the bandwidth problem in HAC estimation discussed in Section 3.4. Following the discussion there, an alternative is to base the test on the q low-frequency averages $\mathbf{X}_T$. From (57), the point-optimal *low-frequency unit root test* $H_0 : c = 0$ versus $H_1 : c = c_1$ rejects for large

values of the test statistic

$$LFUR = \frac{\mathbf{X}_T' \mathbf{\Omega}_{XX}^{LTU}(0)^{-1} \mathbf{X}_T}{\mathbf{X}_T' \mathbf{\Omega}_{XX}^{LTU}(c_1)^{-1} \mathbf{X}_T}. \tag{58}$$

The full-frequency analysis in Elliott (1999) shows that the point-optimal test using $c_1 = 10$ has good power properties for a wide range of values of c under the alternative, and this holds also for the low-frequency version. Table 6 presents 10%, 5%, and 1% critical values for $LFUR$ for values of q ranging from $q = 2$ to $q = 30$.

Example: Applying this test to the real exchange rate data plotted in Figure 1 with $q = 22$ yields $LFUR = 0.829$. From Table 6 the unit root null is rejected at the 10% but not 5% level. The p-value is 0.053.

This $LFUR$ test supposes $x_t = \mu + u_t$, where u_t follows a LTU process, so it allows x_t to have a non-zero mean under the alternative. Thus, it corresponds to Dickey-Fuller tests that include a constant term. If instead, x_t includes a time trend, so that $x_t = \mu_0 + \mu_1 t + u_t$, a modification is required. This modification is discussed in Müller and Watson (2008) and involves changing the low-frequency weights in $\mathbf{\Psi}_T$ so that they are orthogonal to both the time trend and the constant. Before presenting the modification, it is useful to highlight a feature of the cosine weights in $\mathbf{\Psi}_T$ that was not mentioned earlier: namely, that the columns of $\mathbf{\Psi}_T$ are the eigenvectors corresponding to the largest q eigenvalues of the $T \times T$ covariance matrix of a demeaned random walk. The time-trend modification replaces these with the corresponding eigenvectors of the covariance matrix of the demeaned and detrended random walk, that is, the eigenvectors of $T^{-2}\mathbf{M}_T\mathbf{A}_T\mathbf{A}_T'\mathbf{M}'$, where $\mathbf{A}_T\mathbf{A}_T'$ is the random-walk covariance matrix of Section 3.2 with i,j-th element equal to $\min(i,j)$, $\mathbf{M}_T = \mathbf{I}_T - \mathbf{Z}_T(\mathbf{Z}_T'\mathbf{Z}_T)^{-1}\mathbf{Z}_T'$, and $\mathbf{Z}_T$ the $T \times 2$ matrix with $[1, t]$ in its t-th row. For concreteness, call these new weights $\mathbf{\Psi}_T^\tau$, the new low-frequency averages $\mathbf{X}_T^\tau = \mathbf{\Psi}_T^{\tau\prime}\mathbf{x}_{1:T}$, and, via (15) the new covariance matrix $\mathbf{\Omega}_{XX}^{LTU,\tau}(c)$. The resulting $LFUR^\tau$ test is the same as (58) after replacing $\mathbf{X}_T$ and $\mathbf{\Omega}$ with these trend-adjusted values. Table 6 presents critical values for the $LFUR^\tau$ statistic.

Stationarity ($I(0)$) Test: Tests of the $I(0)$ null have been developed in the context of the LL model, where $H_0 : g = 0$ corresponds to the $I(0)$ model (see Nyblom (1989), Kwiatkowski et al. (1992) and Elliott and Müller (2006)). Implementation of these tests requires consistent estimation of the long-run variance, again raising issues of the appropriate choice of

	LFUR								LFST		
	LFUR (mean)				LFUR^{τ} (trend)						
q	10%	5%	1%		10%	5%	1%		10%	5%	1%
2	0.230	0.240	0.243		0.395	0.400	0.401		10.575	10.987	11.126
3	0.314	0.359	0.410		0.500	0.533	0.565		7.086	8.601	10.501
4	0.389	0.433	0.512		0.565	0.595	0.645		4.825	6.002	8.552
5	0.446	0.499	0.579		0.615	0.649	0.702		3.698	4.493	6.577
6	0.495	0.549	0.633		0.656	0.690	0.742		3.080	3.619	5.179
7	0.536	0.590	0.675		0.688	0.722	0.773		2.668	3.081	4.260
8	0.571	0.625	0.707		0.716	0.749	0.798		2.383	2.721	3.651
9	0.600	0.654	0.735		0.739	0.770	0.817		2.176	2.454	3.185
10	0.627	0.679	0.757		0.759	0.789	0.834		2.020	2.255	2.879
11	0.650	0.701	0.775		0.775	0.805	0.848		1.900	2.101	2.635
12	0.670	0.720	0.793		0.790	0.818	0.860		1.804	1.980	2.444
13	0.688	0.737	0.807		0.804	0.830	0.870		1.726	1.883	2.291
14	0.705	0.752	0.819		0.815	0.841	0.879		1.661	1.801	2.161
15	0.719	0.765	0.831		0.825	0.850	0.886		1.607	1.734	2.058
16	0.733	0.777	0.840		0.835	0.859	0.894		1.561	1.678	1.972
17	0.745	0.788	0.850		0.843	0.866	0.900		1.521	1.628	1.896
18	0.756	0.798	0.857		0.851	0.873	0.905		1.486	1.584	1.828
19	0.766	0.807	0.864		0.858	0.879	0.910		1.455	1.547	1.774
20	0.775	0.815	0.870		0.864	0.884	0.915		1.429	1.514	1.726
21	0.784	0.823	0.877		0.870	0.890	0.919		1.406	1.485	1.684
22	0.793	0.831	0.882		0.875	0.895	0.922		1.384	1.459	1.646
23	0.801	0.837	0.887		0.880	0.899	0.926		1.364	1.435	1.611
24	0.807	0.843	0.892		0.885	0.903	0.929		1.347	1.414	1.579
25	0.814	0.849	0.896		0.889	0.907	0.932		1.331	1.395	1.552
26	0.820	0.854	0.900		0.893	0.911	0.935		1.316	1.377	1.526
27	0.826	0.859	0.904		0.897	0.914	0.938		1.303	1.361	1.501
28	0.831	0.864	0.908		0.901	0.917	0.940		1.291	1.346	1.481
29	0.836	0.868	0.911		0.904	0.920	0.942		1.279	1.332	1.461
30	0.841	0.872	0.914		0.907	0.923	0.945		1.269	1.320	1.442

Table 6: LFUR and LFST critical values

bandwidth with implications for the test's size and power, and as before, an alternative is to use the low-frequency averages $\mathbf{X}_T$. The corresponding point optimal test for the alternative $H_1 : g = g_1$ takes the form (57) using the covariance matrix $\mathbf{\Omega}_{XX}^{LL}(g) = \mathbf{I}_q + g^2\mathbf{\Omega}_{XX}^{I(1)}$, where $\mathbf{\Omega}_{XX}^{I(1)}$ is given in (14). This yields the *low-frequency stationarity* ($LFST$) test that rejects for large values of

$$LFST = \frac{\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LL}(0)^{-1}\mathbf{X}_T}{\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LL}(g_1)^{-1}\mathbf{X}_T} = \frac{\mathbf{X}_T'\mathbf{X}_T}{\mathbf{X}_T'(\mathbf{I}_q + g_1^2\mathbf{\Omega}_{XX}^{I(1)})^{-1}\mathbf{X}_T} = \frac{\sum_{j=1}^q X_{j,T}^2}{\sum_{j=1}^q X_{j,T}(1 + g_1^2/(j\pi)^2)^{-1}}. \quad (59)$$

Müller and Watson (2008) show that the test using $g_1 = 10$ has good power for a wide range of values values of g. Critical values for $LFST$ with $g_1 = 10$ and for different values of q are listed in Table 6.

Example: Applying this test to the TFP growth rate data plotted in Figure 1 with $q = 14$ yields $LFST = 1.87$. From Table 6, the $I(0)$ null is rejected at the 5% level. The p-value is 0.035.

5.2.2 Confidence Sets for Persistence Parameters

As discussed in Section 7, confidence intervals can be constructed by inverting tests, and confidence intervals with small average length are obtained by inverting powerful tests. Here we present confidence intervals for persistence parameters based on inverting the family of tests $H_0 : \vartheta = \vartheta_0$, indexed by ϑ_0. An issue is the appropriate alternative for each of these tests. As discussed in Section 7, a sensible to way proceed is to use a single alternative that encompasses many values of ϑ; the associated composite alternative can be transformed into a simple mixture alternative with associated mixing weights described by the c.d.f. F. The resulting alternative becomes $H_a : \vartheta$ is drawn from F. The weight function F is application specific and determines the values of ϑ where the test has greatest power. For computational convenience, it is useful to choose F with discrete support, so it places weight f_i on, say n_ϑ values of ϑ. The resulting likelihood ratio statistic is

$$LR_F(\mathbf{X}_T^s, \vartheta_0) = \frac{\sum_{i=1}^{n_\vartheta} f(\mathbf{X}_T^s|\vartheta_i)f_i}{f(\mathbf{X}_T^s|\vartheta_0)} = \frac{\sum_{i=1}^{n_\vartheta} |\mathbf{\Omega}_{XX}(\vartheta_i)|^{-1/2}\left(\mathbf{X}_T'\mathbf{\Omega}_{XX}^{-1}(\vartheta_i)\mathbf{X}_T\right)^{-q/2} f_i}{|\mathbf{\Omega}_{XX}(\vartheta_0)|^{-1/2}\left(\mathbf{X}_T'\mathbf{\Omega}_{XX}^{-1}(\vartheta_0)\mathbf{X}_T\right)^{-q/2}}. \quad (60)$$

Critical values for the resulting tests will depend on ϑ_0 and F, but are readily computed via Monte Carlo simulations from $\mathbf{X}_T \sim \mathcal{N}(0, \mathbf{\Omega}_{XX}(\vartheta_0))$. Thus, the test rejects the null

hypothesis when $LR_F(\mathbf{X}_T^s, \vartheta_0) > cv_F(\vartheta_0)$, and the confidence set collects the values of ϑ_0 that are not rejected.

Examples: We show four examples, with results summarized in Table 5 that were introduced in the Bayes section. The first two examples construct confidence intervals for c in the LTU model using the real exchange rate data, where we report results using two more familiar parameters, $\rho = 1 - c/T$ and the half-life parameter $h = e^{-c(h/T)}$. The first example uses a weighting function F chosen so that $\rho = 1 - c/T$ is uniformly distributed on 0.5 to 1.0, so the tests focus power on these values of the AR(1) parameter. Recall this distribution was used as the prior in the Bayes version of this problem. The second example is the same as the first, but now F is chosen so the implied half-life, h, is distributed $\mathcal{U}(0, 100)$; again, this was used as a prior in the Bayes version of this problem. The other examples also use the Bayes priors for F: the next example constructs confidence intervals for d in the $I(d)$ model using the realized volatility data $d \sim \mathcal{U}(0.5, 1.5)$ under F; the final example constructs confidence intervals for g in the LL model using the TFP growth rate data where $\ln(g) \sim \mathcal{U}(0.1, 500)$ under F. In all cases, the F distributions are approximated using a 200-point grid, the test is implemented for 500 values of ϑ_0, and the values not-rejected are collected to form the confidence set.

We make two comments about the frequentist confidence intervals reported in Table 4. First, they are similar to, albeit somewhat wider than their Bayes counterparts in Table 1. This is the price that frequentist inference pays for uniform coverage. Second, and focusing on the confidence intervals for the LTU persistence parameters for the real exchange rate: the confidence intervals depend on the weighting function F just as the Bayes credible sets depend on the prior. The confidence interval for ρ that is constructed to minimize expected average length under the weight function F with $\rho \sim \mathcal{U}(0.5, 1.0)$ is wider and shifted to the left compared to the expected length minimizing confidence interval using the weight function $h \sim \mathcal{U}(0, 100)$. This is similar to the associated Bayes credible intervals for ρ shown in Table 1.

5.3 Frequentist Inference Using Least Favorable Distributions

The low-frequency univariate models are characterized by a location parameter μ, a scale parameter σ, and a persistence parameter ϑ. When interest focuses on ϑ, the additional

parameters (μ, σ) complicate frequentist tests (which are required to control size for all values of the (ϑ, μ, σ) included in the null hypothesis); the parameters (μ, σ) are hence suitably referred to as *nuisance* parameters. The last sub-section focused on ϑ, but invariance restrictions fortuitously eliminated the other two parameters, μ and σ. Said differently, location and scale invariance led us to use the transformed data $\mathbf{X}_T^s$ instead of the original data $\mathbf{X}_T^0$, and the probability distribution of $\mathbf{X}_T^s$ depended only on ϑ. These types of invariance and related equivariance restrictions turn out to be useful tools for simplifying frequentist inference in the face of nuisance parameters. Section 7 provides a general discussion. Another useful concept discussed in Section 7 is a *least favorable distribution* (LFD) for the nuisance parameters that, essentially, allows them to be averaged out of the problem. In some problems involving inference about low-frequency parameters, these LFDs can be approximated using the numerical methods discussed in Section 7.5. This leads to specialized problem-specific frequentist inference procedures that can be embedded into software for carrying out tests and forming confidence intervals. The remainder of this section discusses examples of this approach, where the required software is available in the online appendix for this chapter.

5.3.1 Inference about the Mean in a Highly Persistent Stationary Process

Section 4.3.1 presented Bayes methods for inference about the mean of stationary LTU process. Here we revisit the problem and apply frequentist methods. The density of $\mathbf{X}_T^0$ (see (41)) is characterized by μ, the parameter of interest, and (σ, c) are nuisance parameters. The derivation of frequentist hypothesis tests and confidence intervals involves invariance considerations and the determination of an approximate least favorable distribution for c. See Section 7 and Müller (2014) for details. We have implemented these methods and they are incorporated in the software available in the online appendix

Example: Table 4 shows results from applying these methods to unemployment rate data. XXX Discussion to be added XXX

5.3.2 Inference about Low-Frequency Covariability

Section 4.4 discussed Bayes inference about low-frequency covariability. As highlighted there, a key challenge involves the potentially large number of parameters needed to describe the low-frequency variability for a vector of time series. This problem is particularly challenging

for frequentist inference because of the need to control the size of tests and the coverage rate of confidence intervals uniformly over these parameters.

Müller and Watson (2018) takes up this problem for $n = 2$. Specifically, they consider a model in which the 2×1 vector $\mathbf{x}_t$ evolves as

$$\mathbf{x}_t = \mu + \mathbf{A}\mathbf{b}_t + \mathbf{e}_t$$

where $\mathbf{e}_t$ follows a bivariate $I(0)$ process with long-run covariance matrix $\mathbf{\Sigma}_e$, $\mathbf{A}$ is an unconstrained 2×2 matrix, and $\mathbf{b}_t = (b_{1,t}, b_{2,t})'$ with $b_{i,t}$ independent processes that generalize and nest the LTU and $I(d)$ processes described by a scale parameter and two low-frequency persistence parameters. The low-frequency evolution of $\mathbf{x}_t$ is thus characterized by a 13-dimensional parameter. Invariance and equivariance restrictions reduce this number, but a high dimensional nuisance parameters remains. Finding an approximate least favorable distribution for such a high dimensional nuisance parameter space is a computational challenge and requires careful numerical methods described in Section 7.5.3. Müller and Watson (2018) carry out the required calculations for inference about low-frequency correlations (ρ^{LF}) and bivariate linear regression coefficients (β^{LF}) for various significance levels and values of q. The online appendix includes software that carries out the resulting tests and constructs the associated confidence intervals.

Example: Table 4 shows 67% and 90% for the low-frequency correlation between TFP and GDP growth rates and linear regression coefficient. These intervals are similar to, although slightly wider, than the confidence intervals predicated on the $I(0)$ assumption. The increased width of the intervals reflect their guaranteed coverage over the wide range persistence, location, and scale parameters allowed in the 13-parameter model.

5.3.3 Long-Horizon Prediction Intervals

Section 4.5 described Bayes methods for constructing predictive distributions for $\bar{\mathbf{x}}_{T+1:T+h}$, the average values of $\mathbf{x}_t$ over the out-of-sample period $T+1$ through $T+h$, based on the sample data $\mathbf{X}_T^0$. As discussed there, when T and h are large, $(\mathbf{X}_T^0, \bar{\mathbf{x}}_{T+1:T+h})$ are approximately jointly normally distributed. If the parameters of this distribution were known, the distribution of $\bar{\mathbf{x}}_{T+1:T+h}|\mathbf{X}_T^0$ is approximately normal, with mean and variance given by the familiar conditional normal formula. When the parameters are unknown, the predictive distribution becomes a mixture of these conditional normal distributions, with mixing weights equal to

the posterior distribution of the joint law of $(\mathbf{X}_T^0, \bar{\mathbf{x}}_{T+1:T+h})$. Bayes prediction intervals, that is intervals that contain future realizations of $\bar{\mathbf{x}}_{T+1,T+h}$ with a pre-specified posterior probability can be computed directly the from this mixture. Frequentist inference about $\bar{\mathbf{x}}_{T+1,T+h}$ is similarly simplified by the joint normality of $(\mathbf{X}_T^0, \bar{\mathbf{x}}_{T+1:T+h})$, but frequentist prediction intervals with level $1-\alpha$ must contain the future value with probability of at least $1-\alpha$ for all fixed model parameters over repeated samples.

Müller and Watson (2016) takes up this problem in the context of a univariate time series model that encompasses and generalizes the LTU, $I(d)$ and LL models from Section 2. This encompassing model is characterized by five parameters, three persistence parameters, a location parameter, and a scale parameter. The combination of invariance considerations together with numerically determined least favorable distributions are used to produce prediction intervals that are nearly efficient (in the sense of having smallest weighted average length) among all invariant intervals of the required coverage probability. Software for constructing these prediction sets is included in the online appendix.

Example: Table 5 shows 67% and 90% univariate prediction sets for the average growth rates of TFP and GDP over the next $h = 25$ and 50 years. These are computed using $q = 12$, because they rely the least favorable distributions computed in Müller and Watson (2016), which used this value of q. These prediction intervals, which allow for a general model of persistence, are markedly wider than the $I(0)$ intervals or the Bayes prediction intervals reported in Table 3. Evidently, the data are compatible with low-frequency dynamics that are more persistent than the $I(0)$ model, so that the frequentist coverage guarantee forces the interval to include the wide range of future growth rates that correspond to those dynamics.

6 Bayesian Inference: Concepts and Methods

This section reviews Bayesian concepts and methods. While the examples considered throughout the section relate to low-frequency inference, the concepts and methods are generally applicable in econometrics. Much of the material presented here is discussed in textbooks, often with more details and alternative examples, such as in Gelman et al. (2004), Geweke (2005) and Robert (2007).

6.1 Likelihood, Prior, Posterior

Bayesian analysis is conceptually straightforward: the uncertainty about all unknowns are described using the language of probability, and this uncertainty is updated from the data by applying Bayes rule. Formally, Bayesian analysis involves three ingredients: the prior, the likelihood, and the posterior. In the context of a parametric model with parameter $\theta \in \Theta \subset \mathbb{R}^k$, the researcher posits a *prior* distribution with density p that describes the uncertainty about θ before taking the data information into account. For notational simplicity, we assume that p is the probability density function of a continuous random variable, but almost all of the subsequent discussion goes through with p representing the probability mass function if Θ is finite or countable, or, more generally, a density relative to some dominating measure.

The parametric model describes the distribution of the data $\mathbf{Y} \in \mathcal{Y}$ by its density $f(\mathbf{y}|\theta)$, which is indexed by θ. Again, we treat this as the density of a continuous random variable, but this is easily generalized. For a given realization $\mathbf{Y} = \mathbf{y}$, $f(\mathbf{y}|\theta)$ viewed as a function of θ, is called the *likelihood*.

By Bayes rule, the *posterior* density $p(\theta|\mathbf{y})$ is proportional to the product of the prior and the likelihood

$$p(\theta|\mathbf{y}) \propto p(\theta) f(\mathbf{y}|\theta) \tag{61}$$

where the constant of proportionality is the reciprocal of the *marginal likelihood* $\int_\Theta p(\theta) f(\mathbf{y}|\theta) d\theta$, so that by construction, $p(\theta|\mathbf{y})$ is a probability density function.

We illustrate the concepts in this chapter by the following example.

Example MEAN(a). Consider inference about the mean in a local-to-unity model, as in Section 4.3.1 above. Under the approximation (13), $\mathbf{X}_T^0 \sim \mathcal{N}(\iota_{q+1}\mu, T^{-1}\sigma^2 \mathbf{\Omega}^{LTU}(c))$. Here $\theta = (\mu, \sigma, c) \in \Theta = \mathbb{R} \times (0, \infty)^2$. With prior density $p(\theta)$, the posterior density is proportional to

$$p(\theta|\mathbf{x}_T^0) \propto p(\theta) \sigma^{-q-1} |\mathbf{\Omega}^{LTU}(c)|^{-1/2} \exp\left[-\tfrac{1}{2} T (\mathbf{x}_T^0 - \iota_{q+1}\mu)' \mathbf{\Omega}^{LTU}(c)^{-1} (\mathbf{x}_T^0 - \iota_{q+1}\mu)/\sigma^2\right]. \tag{62}$$

With (μ, σ) treated as continuous and c constrained to take values on a finite grid $\{c_i\}_{i=1}^m$, the right hand side viewed as a function of $c \in \{c_i\}_{i=1}^m$ is proportional to the posterior probability mass function conditional on (μ, σ). ▲

The computational challenge in Bayesian statistics is that it is often difficult to obtain closed-form expressions for the marginal posterior distribution of elements of θ. And as soon

as θ is moderately high-dimensional, one cannot simply rely on numerical integration to obtain such marginal densities from (61). A very large literature has developed numerous numerical approaches to deal with this difficulty. Subsection 6.4 below reviews some basic methods that are often sufficient to obtain accurate posterior approximations in applications like those considered in this chapter.

6.2 Credible Sets

Suppose we are interested in a particular real-valued function of the parameter θ, $\gamma = h(\theta)$, with $h : \Theta \mapsto \Gamma \subset \mathbb{R}$. Let $p(\gamma|\mathbf{y})$ be the posterior density of γ induced by the posterior density $p(\theta|\mathbf{y})$. The posterior uncertainty about γ is usefully described by a set that covers, say, 95% of the posterior mass. Such a set $\hat{\Gamma}(\mathbf{y})$ is called a *credible set* of level 95%. Technically, $\hat{\Gamma}$ is a function that maps data into subsets of Γ, $\hat{\Gamma} : \mathcal{Y} \mapsto \mathcal{G}$, where $\mathcal{G}$ collects all Borel sets on Γ.

The two most common forms of level $1-\alpha$ credible sets are the highest posterior density (HPD) set

$$\hat{\Gamma}_{HPD}(\mathbf{y}) = \{\gamma : p(\gamma|\mathbf{y}) > C\}$$

where the constant C is chosen such that $\int_{\hat{\Gamma}_{HPD}(\mathbf{y})} p(\gamma|\mathbf{y})d\gamma = 1-\alpha$, and the equal-tailed interval

$$\hat{\Gamma}_{ET}(\mathbf{y}) = [L_{ET}(\mathbf{y}), U_{ET}(\mathbf{y})], \ \int_{-\infty}^{L_{ET}(\mathbf{y})} p(\gamma|\mathbf{y})d\gamma = \int_{U_{ET}(\mathbf{y})}^{\infty} p(\gamma|\mathbf{y})d\gamma = \alpha/2.$$

The HPD set is the shortest credible set of given level $1-\alpha$. The equal-tailed interval is relatively easier to compute from a random sample of posterior draws. It also can be appealing in some contexts for the endpoints to have the interpretation of posterior quantiles, so that from the perspective of the posterior distribution, it is equally likely that γ falls below the lower endpoint, or above the upper endpoint of the credible interval.

6.3 Uninformative Priors and Invariance

For some likelihoods, (61) defines a posterior distribution even when the prior density is not integrable, that is when $\int_\Theta p(\theta)d\theta$ does not exist. Such priors are called *improper*. It is sometimes useful to consider improper priors when attempting to be *uninformative* about θ.

One systematic approach to obtaining uninformative priors is the theory of invariance. Intuitively, invariance imposes the restriction that the posterior distribution, and correspondingly the credible set, change in a predetermined fashion when the data is transformed in a particular way. For instance, it might make sense to require that the posterior distribution of a persistence parameter remains unaffected by scale changes of the data, such as those induced by changing the units of measurement.

A general discussion of the derivation of invariant priors is outside the scope of this chapter. See Chapter 6 of Berger (1985) or Chapter 7 of Robert (2007) for introductions and references. Here we note that the natural invariant prior for a (scalar) location parameter is the improper constant prior density, and the natural invariant prior density for a scale parameter σ is the improper density $1/\sigma$. In the presence of both a location and scale parameter, the joint invariant prior on (μ, σ) has density $1/\sigma$.

Example MEAN(b). From now on we consider a prior with density $p(\theta) = p_c(c)/\sigma$. The posterior density (62) is then proportional to

$$p(\theta|\mathbf{x}_T^0) \propto p_c(c)\sigma^{-q-2}|\mathbf{\Omega}^{LTU}(c)|^{-1/2} \exp\left[-\tfrac{1}{2}T(\mathbf{x}_T^0 - \iota_{q+1}\mu)'\mathbf{\Omega}^{LTU}(c)^{-1}(\mathbf{x}_T^0 - \iota_{q+1}\mu)/\sigma^2\right]. \tag{63}$$

The marginal posterior for (μ, c) is obtained by integrating out σ. After the change of variables $\nu = \frac{1}{2\sigma^2}T(\mathbf{x}_T^0 - \iota_{q+1}\mu)'\mathbf{\Omega}^{LTU}(c)^{-1}(\mathbf{x}_T^0 - \iota_{q+1}\mu)$, we find

$$\begin{aligned} p(\mu, c|\mathbf{x}_T^0) &\propto p_c(c)|\mathbf{\Omega}^{LTU}(c)|^{-1/2}\left[(\mathbf{x}_T^0 - \iota_{q+1}\mu)'\mathbf{\Omega}^{LTU}(c)^{-1}(\mathbf{x}_T^0 - \iota_{q+1}\mu)\right]^{-(q+1)/2}\int_0^\infty \nu^{(q-1)/2}e^{-\nu}d\nu \\ &\propto p_c(c)|\mathbf{\Omega}^{LTU}(c)|^{-1/2}\left[(\mathbf{x}_T^0 - \iota_{q+1}\mu)'\mathbf{\Omega}^{LTU}(c)^{-1}(\mathbf{x}_T^0 - \iota_{q+1}\mu)\right]^{-(q+1)/2} \\ &\propto p_c(c)|\mathbf{\Omega}_{XX}^{LTU}(c)|^{-1/2}(\mathbf{x}_T'\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{x}_T)^{-q/2}s(\mathbf{x}_T, c)^{-q}\left[\frac{(\mu - m(\mathbf{x}_T^0, c))^2}{q \cdot s(\mathbf{x}_T^0, c)^2} + 1\right]^{-(q+1)/2} \end{aligned} \tag{64}$$

where

$$m(\mathbf{x}_T^0, c) = \bar{x}_{1:T} + \mathbf{\Omega}_{\bar{x}X}^{LTU}(c)\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{x}_T \tag{65}$$

$$s(\mathbf{x}_T^0, c)^2 = (\mathbf{\Omega}_{\bar{x}\bar{x}}^{LTU}(c) - \mathbf{\Omega}_{\bar{x}X}^{LTU}(c)\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{\Omega}_{X\bar{x}}^{LTU}(c))\mathbf{x}_T'\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{x}_T/q \tag{66}$$

and the last line follows from tedious but straightforward algebra. Thus, conditional on c, the posterior distribution of $(\mu - m(\mathbf{x}_T^0, c))/s(\mathbf{x}_T^0, c)$ is Student-t with q degrees of freedom, or, equivalently, the distribution of μ is equal to a Student-t distribution scaled by $s(\mathbf{x}_T^0, c)$ and shifted by $m(\mathbf{x}_T^0, c)$.

The marginal posterior distribution for c may be obtained by integrating out μ in (64). Noting that the scaled and shifted Student-t density integrates to unity, we obtain that the posterior density for c is proportional to

$$p_c(c)|\mathbf{\Omega}_{XX}^{LTU}(c)|^{-1/2}(\mathbf{x}_T'\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{x}_T)^{-q/2}. \tag{67}$$

In particular, with $c \in \{c_i\}_{i=1}^m$, the unconditional posterior distribution for μ is a finite mixture of Student-t distributions scaled by $s(\mathbf{x}_T, c_i)$ and shifted by $m(\mathbf{x}_T, c_i)$, with mixing weights proportional to (67) evaluated at $c \in \{c_i\}_{i=1}^m$.

Note that the transformation of $\mathbf{x}_T^0 \to a_\sigma \mathbf{x}_T^0 + \iota_{q+1} a_\mu$ with $(a_\mu, a_\sigma) \in \mathbb{R} \times (0, \infty)$ induces the transformations $m(\mathbf{x}_T^0, c) \to a_\sigma m(\mathbf{x}_T^0, c) + a_\mu$ and $s(\mathbf{x}_T^0, c) \to a_\sigma s(\mathbf{x}_T^0, c)$, for all c. Since scale changes of $s(\mathbf{x}_T^0, c)$ do not change the relative values of (67), the posterior distribution of c is unaffected. Thus, in this example, the improper prior density $1/\sigma$ on (μ, σ) induces a posterior distribution and corresponding equal-tailed or HPD sets for μ that change in accordance with scale and location changes of the observation $\mathbf{x}_T^0$. ▲

6.4 Markov Chain Monte Carlo Posterior Samplers

As noted above, for many inference problems it is not possible to analytically determine the marginal posterior distribution of the parameters of interest. Instead, posterior distributions are typically obtained by generating a random sample $\theta^{(l)}$, $l = 1, \ldots, N$ from the posterior distribution for some large N. Quantiles and moments of the posterior distribution of θ or a function of θ, $\gamma = h(\theta)$, are then approximated by the corresponding quantiles and moments of $\theta^{(l)}$ and $\gamma^{(l)} = h(\theta^{(l)})$, $l = 1, \ldots, N$.

Most algorithms do not generate an *i.i.d.* sample $\theta^{(l)}$, but rather a Markov chain with the posterior distribution as the stationary distribution; these are called Markov chain Monte Carlo (MCMC) simulators. Since the posterior distribution is unknown, the starting value $\theta^{(0)}$ is initialized at a value that is reasonably close to the posterior mode, a first batch of *burn-in* draws is discarded (say, 20% of the total draws) to mitigate the effect of this initial condition, and posterior quantities of interest are computed from the remaining draws. Strongly autocorrelated chains are said to *mix poorly*, and the stronger the autocorrelation in the draws $\theta^{(l)}$, the larger N needs to be for the sample moments and quantiles to be accurate estimates of their population counterparts. In practice, one can get a sense of the approximation quality by computing autocorrelation robust confidence intervals. It is also

useful to check convergence by visually inspecting the evolution of $\theta^{(l)}$ as a function of l (so-called *trace plots*), and to check that the initial $\theta^{(0)}$ has indeed no lasting influence by generating multiple chains with different starting values.

6.4.1 Gibbs Sampling

Decompose the parameter $\theta \in \Theta \subset \mathbb{R}^k$ into two components $\theta = (\theta_I, \theta_{II}) \in \mathbb{R}^{k_I} \times \mathbb{R}^{k_{II}}$ with $k_I + k_{II} = k$. Suppose the posterior distribution is such that we know how to draw θ_I conditional on the value of θ_{II}, and we also know how to draw θ_{II} conditional on θ_I. If $\theta^{(l)} = (\theta_I^{(l)}, \theta_{II}^{(l)})$ is a draw from the posterior distribution, and we draw $\theta_I^{(l+1)}$ from the conditional distribution of $\theta_I | \theta_{II} = \theta_{II}^{(l)}$, then by the definition of the conditional distribution, $(\theta_I^{(l+1)}, \theta_{II}^{(l)})$ also has distribution equal to the posterior distribution. If we further draw $\theta_{II}^{(l+1)}$ from the conditional distribution $\theta_{II} | \theta_I = \theta_I^{(l+1)}$, then by the same logic, $\theta^{(l+1)} = (\theta_I^{(l+1)}, \theta_{II}^{(l+1)})$ has distribution equal to the posterior distribution. Repeating these two steps thus yields a Markov chain with stationary distribution equal to the posterior distribution of θ. The same approach readily extends to the case where θ is decomposed into more than two components.

Example MEAN(c). In Example MEAN(b) we derived the posterior distribution for μ analytically, but it is easy to derive a Gibbs sampler that treats each element of $\theta = (\mu, \sigma, c)$ as its own block. Comparing the posterior (63) as a function of μ with a normal density, we see that the conditional distribution of μ given (σ, c) is normal with mean $\iota_{q+1}' \mathbf{\Omega}^{LTU}(c)^{-1} \mathbf{x}_T^0 / (\iota_{q+1}' \mathbf{\Omega}^{LTU}(c)^{-1} \iota_{q+1})$ and variance $\sigma^2 (T \iota_{q+1}' \mathbf{\Omega}^{LTU}(c)^{-1} \iota_{q+1})^{-1}$. Similarly, the posterior distribution of $\sigma^{-2} T (\mathbf{x}_T^0 - \iota_{q+1}\mu)' \mathbf{\Omega}^{LTU}(c)^{-1} (\mathbf{x}_T^0 - \iota_{q+1}\mu)$ conditional on (μ, c) is recognized as a chi-squared distribution with $q+1$ degrees of freedom, so that a conditional draw of σ can be generated by dividing $\sqrt{T(\mathbf{x}_T^0 - \iota_{q+1}\mu)' \mathbf{\Omega}^{LTU}(c)^{-1} (\mathbf{x}_T^0 - \iota_{q+1}\mu)}$ by the square root of a randomly generated chi-squared random variable. Finally, conditional on (μ, σ), the posterior for c is a discrete random variable taking on values in $\{c_i\}_{i=1}^m$ with probabilities proportional to (63). Let v_i, $i = 1, \ldots, m$ be equal to right hand side of (63) with $c \in \{c_i\}_{i=1}^m$, and let $V_i = \sum_{j \leq i} v_j$. A random draw from this conditional distribution is then given by c_J where $J = 1 + \sum_{i=1}^m \mathbf{1}[V_i \leq U V_m]$, where U is uniform $U \sim \mathcal{U}[0, 1)$. ▲

Gibbs sampling can be a very powerful technique. However, it can produce highly correlated draws. Suppose, for instance, that the posterior distribution for $\theta = (\theta_I, \theta_{II}) \in \mathbb{R}^2$ is bivariate normal with a correlation coefficient that is close to one. Then the conditional

distributions $\theta_I|\theta_{II}$ and $\theta_{II}|\theta_I$ are much less variable than their unconditional distributions, so that Gibbs sampling only very slowly visits the entire posterior distribution, resulting in a poorly mixing Markov chain. The remedy is to draw θ jointly, which in this case is easily done (and Gibbs sampling is unnecessary). This message holds more generally, though: mixing is improved by combining blocks of parameters that are highly correlated.

6.4.2 Metropolis Hastings Algorithm

The goal of MCMC is to draw from the posterior distribution with density $p(\theta|\mathbf{y})$. Now suppose we were to draw $\tilde{\theta}^{(l)}$ *i.i.d.* from density $g(\theta)$ instead, where g has same support as $p(\theta|\mathbf{y})$, but $g(\theta) \neq p(\theta|\mathbf{y})$. Then averages of $\tilde{\gamma}^{(l)} = h(\tilde{\theta}^{(l)})$ obviously do not converge to the corresponding functions of the posterior distribution of $\gamma = h(\theta)$. In particular, values of θ where $p(\theta|\mathbf{y})$ is larger than $g(\theta)$ are undersampled. One approach to address this imbalance is to give these draws a correspondingly larger weight when computing averages of $\tilde{\theta}^{(l)}$. This is the key idea underlying *importance sampling* that is reviewed in Section 7.5.1 below in a different context.

An alternative approach to address the imbalance is count the undersampled draws $\tilde{\theta}^{(l)}$ repeatedly in the computation of the averages. For instance, if $p(\theta_2|\mathbf{y})/g(\theta_2)$ is twice as large as $p(\theta_1|\mathbf{y})/g(\theta_1)$, then we would regain balance between these two values if in the construction of $\theta^{(l)}$ from $\tilde{\theta}^{(l)}$, we always included realizations of $\tilde{\theta}^{(l)} = \theta_2$ twice. Alternatively, we could generate a Markov Chain $\theta^{(l)}$ as follows: Whenever $\theta^{(l)} = \theta_2$, we flip a coin and set $\theta^{(l+1)} = \theta_2$ if the coin shows heads, and only take another random draw $\theta^{(l+1)}$ with probability density g if the coin shows tails. In this manner, the reweighting is accomplished by random repetitions, with the expected number of repetitions equal to two. If this "random repetition" correction is applied to all possible pairs of $\theta_1, \theta_2 \in \Theta$ then the resulting Markov Chain is overall properly balanced, with the intended posterior distribution as the stationary distribution.

This random repetition is the basis for the Metropolis-Hastings simulation method. Formally the Metropolis-Hastings algorithm for an independent proposal $g(\theta)$ is shown in Algorithm 1.

Note that the acceptance probability u in (68) of the proposed move θ_p depends on the posterior density only through the ratio

$$\frac{p(\theta_p|\mathbf{y})}{p(\theta^{(l)}|\mathbf{y})} = \frac{p(\theta_p)f(\mathbf{y}|\theta_p)}{p(\theta^{(l)})f(\mathbf{y}|\theta^{(l)})}.$$

Algorithm 1 Metropolis Hastings with independent proposal

1. Draw θ_p from density g, and let $U \sim \mathcal{U}[0,1]$ be independent of θ_p.

2. Compute
$$u = \min\left(\frac{p(\theta_p|\mathbf{y})}{g(\theta_p)}\frac{g(\theta^{(l)})}{p(\theta^{(l)}|\mathbf{y})}, 1\right) \tag{68}$$

3. If $U < u$, set $\theta^{(l+1)}$ equal to θ_p. Otherwise, set $\theta^{(l+1)} = \theta^{(l)}$.

The Metropolis-Hastings algorithm thus does not involve the value of the marginal likelihood. Since computing the marginal likelihood can be a challenging task, this is a highly appealing feature. In practice, the proposal density g in the independent Metropolis-Hastings algorithm must be a somewhat reasonable guess of $p(\theta|\mathbf{y})$ to ensure that the resulting chain mixes well.

Example MEAN(d). Consider Example MEAN(b) with (μ, σ) already integrated out, so that c is the only remaining parameter in the problem, and the posterior is proportional to (67). There is no need to employ the Metropolis Hastings algorithm to characterize this discrete distribution, but for illustration purposes, suppose we wanted to. Let g be the uniform distribution on the m points $\{c_i\}_{i=1}^m$. Given $c^{(l)}$, we generate $c^{(l+1)}$ as follows: Let J be a uniformly drawn index $J \in \{1, \ldots, m\}$. We accept the move from $c^{(l)}$ to the proposed value c_J if U is smaller than the ratio of (67) evaluated at c_J and $c^{(l)}$, respectively, and set $c^{(l+1)} = c^{(l)}$ otherwise. ▲

The Metropolis Hastings idea extends to the case where θ_p is drawn from a distribution that depends on the current value $\theta^{(l)}$, so that g becomes a conditional density. In that case, the balancing through repetition is more delicate, since one must also take into account how often the current value would have been generated from the potential new value θ_p. This is shown in Algorithm 2.

A common choice for the general proposal is a conditional distribution centered at $\theta^{(l)}$. This ensures that with $\theta^{(l)}$ reasonably close to the posterior mode, the proposal automatically focusses on part of the parameter space with relatively high posterior density. The choice for the step size (that is the size of the move from $\theta^{(l)}$ to θ_p) faces a trade-off: too small steps lead to too little exploration of the posterior distribution, and too large steps lead to too few accepted moves. A reasonable compromise in many problems is a step size that leads to

Algorithm 2 Metropolis-Hastings with general proposal

1. Draw θ_p from density $g(\cdot|\theta^{(l)})$, and let $U \sim \mathcal{U}[0,1]$ be independent of θ_p.

2. Compute
$$u = \min\left(\frac{p(\theta_p|\mathbf{y})}{g(\theta_p|\theta^{(l)})}\frac{g(\theta^{(l)}|\theta^p)}{p(\theta^{(l)}|\mathbf{y})}, 1\right) \tag{69}$$

3. If $U < u$, set $\theta^{(l+1)}$ equal to θ_p. Otherwise, set $\theta^{(l+1)} = \theta^{(l)}$.

about 30% acceptance probability.

Note that if the density $g(\theta_p|\theta^{(l)})$ is such that $\theta_p \sim \theta^{(l)} + \nu$, where ν has a symmetric distribution and is independent of $\theta_{(l)}$, then $g(\theta^{(l)}|\theta^p) = g(\theta_p|\theta^{(l)})$. The acceptance probability in (69) then simplifies to $\min(p(\theta_p|\mathbf{y})/p(\theta^{(l)}|\mathbf{y}), 1)$, yielding the classic Random Walk Metropolis-Hastings algorithm. In general, this simplification will only be possible in unbounded parameter spaces Θ, and sometimes θ is reparametrized for that purpose. Alternatively, one can adjust the random walk proposal at the endpoints and apply the general formula (69).

Example MEAN(e). As an alternative to Example MEAN(d), suppose we employ a random walk type proposal for $c^{(l)}$. Let $J^{(l)} \in \{1, \ldots, m\}$ be the current index, $c^{(l)} = c_{J^{(l)}}$, and let J_p the index of the proposed move for c. Let the proposal g be as follows: If $2 \leq J^{(l)} \leq m-1$, then J_p is equal to $J^{(l+1)} \pm 1$ with equal probability. If $J^{(l)} = 1$, then $J_p = 2$ always, and if $J^{(l)} = m$, then $J_p = m-1$ always. The ratio $g(\theta^{(l)}|\theta^p)/g(\theta_p|\theta^{(l)})$ in (69) then becomes equal to 1 for $2 \leq J_p, J^{(l)} \leq m-1$, it is equal to 2 if $J^{(l)} \in \{1, m\}$, and it is equal to 1/2 if $(J^{(l)}, J_p) \in \{(2,1), (m-1, m)\}$. ▲

It is also straightforward to apply the Metropolis-Hastings algorithm within a Gibbs sampler, that is to generate a draw of, say, $\theta_I^{(l+1)}$ given $\theta_{II}^{(l)}$ in the notation of Section 6.4.1. The algorithms then still applies with θ now playing the role of θ_I, and $p(\cdot|\mathbf{y})$ equal to the density $p(\cdot|\mathbf{y}, \theta_{II}^{(l)})$.

Example MEAN(f). In the context of Example MEAN(c), instead of drawing c directly from the conditional discrete distribution in each Gibbs step, we could instead employ a Metropolis-Hastings approach. The only difference to the discussion in Examples MEAN(d) and MEAN(e) would be that now, the ratio $p(\theta_p|\mathbf{y})/p(\theta^{(l)}|\mathbf{y})$ in (68) and (69) corresponds

to (63) viewed as a function of c and conditional on (μ, σ), rather than (67), evaluated at the proposed and current value of c. ▲

6.5 Geweke Test

The MCMC methods outlined above are conceptually straightforward, but coding mistakes are inevitable. Geweke (2004) discusses how the internal logic of MCMC posterior simulators can be used to produce practical checks on MCMC computer code. The algorithm is a simple application of Gibbs sampling reviewed above, except that the sampling is augmented by also sequentially drawing $\mathbf{Y}$: Let $\theta^{(l+1)}$ be a draw from $p(\theta|\mathbf{y}^{(l)})$, and let $\mathbf{y}^{(l+1)}$ be a draw from $\mathbf{Y}|\theta^{(l+1)}$. The stationary distribution of $\theta^{(l)}$ in this chain is the prior distribution. If these augmented MCMC simulations produces a distribution of draws for θ that differs from the prior, then the code that purports to generate a draw from $p(\theta|\mathbf{y}^{(l)})$ is wrong. This produces a powerful check on the correctness of the code.

Example MEAN(g). In the context of Example MEAN(c), the augmentation step simply consists of drawing $\mathbf{x}_T^{0,(l+1)}$ from $\mathcal{N}(\mu\iota_{q+1}, T^1\sigma^2\Omega^{LTU}(c))$. ▲

One might be tempted to visually inspect the similarity of the prior distribution with the distribution from the augmented chain. We found that Algorithm 3 more reliably detects problems.

This implementation of the Geweke tests has two attractive features. First, by considering the prior percentiles of weighted averages of the elements of θ, the algorithm checks the joint distribution properties of θ. Second, the batched t-statistic in Step 3 collects more and more draws of $\theta^{(l)}$ in each of the 100 batches. Thus, even with high serial correlation in the draws, batch averages $\hat{\alpha}_j$ become independent across j (and individually Gaussian by a central limit theorem) as N grows sufficiently large. Thus, if everything is coded correctly, any given t-statistic converges in distribution to a Student-t distribution with 99 degrees of freedom, making values above 4 in absolute value highly unlikely even after computing the maximum of a fairly large set of t-statistics.

When the Geweke test indicates a problem, it is often possible to isolate the coding error by considering each Gibbs block separately, that is, by keeping subsets of the parameter fixed in Algorithm 3.

Algorithm 3 Geweke (2004) Test

1. Draw many (say, 100,000) *i.i.d.* draws $\theta_P^{(l)}$ from the prior distribution with density $p(\theta)$. Compute the $5, 10, \ldots, 95$ percentiles of $\upsilon_i'\theta_P^{(l)}$ from these draws for $i = 1, \ldots, 20 + k$, where υ_i is the ith column of I_k for $i \leq k$, and $\upsilon_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_P^{-1})$ with $\boldsymbol{\Sigma}_P$ the prior variance estimated from $\theta_P^{(l)}$ for $20 < i \leq 20 + k$.

2. Initialize the joint state $(\theta, \mathbf{y})$ via a random draw of $\theta^{(0)}$ from the prior, and $\mathbf{y}^{(0)}$ drawn from $\mathbf{Y}|\theta^{(0)}$. Now draw $\theta^{(l+1)}|\theta^{(l)}, \mathbf{y}^{(l)}$ as if one wanted to generate a MCMC chain for the posterior distribution for the data $\mathbf{Y} = \mathbf{y}^{(l)}$. For each l, perform the augmentation step with $\mathbf{y}^{(l+1)}$ drawn from $\mathbf{Y}|\theta^{(l+1)}$.

3. If the total number of draws N of $\theta^{(l)}$ generated so far is divisible by 1000, compute t-statistics for the percentiles of $\upsilon_i'\theta^{(l)}$ relative to the corresponding values computed in Step 1, treating blocks of size $B = N/100$ as being independent. Concretely, if ψ_i is the 100η prior percentile of $\upsilon_i'\theta$ computed in Step 1, then

$$t_i = \frac{\sum_{j=1}^{100}(\hat{\eta}_j - \eta)}{\sqrt{\sum_{j=1}^{100}(\hat{\eta}_j - \overline{\hat{\eta}})^2}} \tag{70}$$

   with $\hat{\eta}_j = B^{-1}\sum_{l=(j-1)b+1}^{jB} \mathbf{1}[\upsilon_i'\theta^{(l)}) < \psi_i]$ and $\overline{\hat{\eta}} = 100^{-1}\sum_{j=1}^{100}\hat{\eta}_j$.

4. Continue to run the chain and monitor the values of the t-statistics (70). As N becomes large, the largest absolute value of these t-statistics over all percentiles and i should become and remain reasonably small (say, smaller than 4).

6.6 Hierarchical Models with Dirichlet Prior on a Discretized Parameter

When dealing with multiple observations from the same data generating process, it sometimes make sense to impose that the parameter values for the different observations are related. For example, suppose we are interested in the persistence properties of n real exchange rates in the context of a local-to-unity model, with fairly large n. For each series, we compute the q dimensional cosine transform, so that the observations are $\mathbf{X}_{T,j} \sim \mathcal{N}(0, T\sigma_j^2 \mathbf{\Omega}_{XX}^{LTU}(c_j))$, $j = 1, \ldots, n$, where c_j is the local-to-unity parameter of the jth series. Under the (unrealistic) assumption that the low-frequency properties of exchange rates are independent, we have a fully specified model, and the only challenge is to form a prior on σ_j and c_j, $j = 1, \ldots, n$. For simplicity, suppose we simply use the uninformative prior with density proportional $1/\sigma_j$ for all σ_j, $j = 1, \ldots, n$, so that the focus is on the prior for c_j.

Note that a prior for $\{c_j\}_{j=1}^n$ that is data independent and *i.i.d.* across j is necessarily highly informative about certain aspects for n large: By the law of large numbers, the implied prior for the average value of c_j across the n countries, $n^{-1}\sum_{j=1}^n c_j$, for instance, is very tightly concentrated around the prior mean for c. And, since the sample contains fairly limited amount of information about each c_j, this is tight prior seems undesirable for assessing the typical degree of mean reversion across the series.

To avoid this feature one has to allow for some flexibility about the common prior distribution of the c_j's. One way to achieve this is by specifying a parametric distribution, and then form a prior over the parameters. For instance, one could specify the c_j to be *i.i.d.* log-normal, and specify a prior over the parameters of the common log-normal distribution. The effective prior over the common distribution of the c_j is then a mixture of log-normal distributions, with mixing weights determined by the prior over the log-normal parameters.

A less parametric approach directly specifies a prior over the common distribution. Suppose the parameter space for c_j is discretized into a finite grid with m elements $\{c_i\}_{i=1}^m$, as in the running example. In the hierarchical model, there is a prior over the distribution $\pi_c(\cdot)$ of c_j, and conditional on π_c, c_j is *i.i.d.* with probability mass function π_c. By definition, $\sum_{i=1}^m \pi_c(c_i) = 1$, so π_c is a point in the m dimensional simplex. A prior over π_c is thus a prior over points in the m dimensional simplex.

The Dirichlet prior is a particularly convenient choice in this context. Under this prior,

π_c may be generated as follows (cf. Gelman et al. (2004)): Draw m independent chi-square distributed random variables W_i with ν/m degrees of freedom. Then $\pi_c(c_i) = W_i / \sum_{l=1}^m W_l$. Note that if ν is large, then $\pi_c(c_i)$ is close to $1/m$ for all i, and the degree of prior variability in this distribution is small. On the other hand, if ν is small, then the distribution of W_i is relatively fat-tailed, and typical realizations of π_c under this prior are far from uniform, with most mass on a few values of c_i. Thus, ν controls the shrinkage of the prior towards the uniform distribution on the simplex. We use ν/m in the specification of the degrees of freedom, since with this choice, the degree of shrinkage for the implied c.d.f. of π_c does not depend on m for m large (since $\sum_{i=1}^{\lfloor rm \rfloor} W_i$ for $r \in [0,1]$ is distributed chi square with $\lfloor rm \rfloor \nu/m \approx r\nu$ degrees of freedom). The choice of grid for c_i controls what this uniform distribution in the simplex implies for the effective distribution of c_i. For instance, if $c_i = F_c^{-1}(u_i)$ with u_i uniform on $[0,1]$, $i = 1, \ldots, m$ and F_c^{-1} the quantile function of the c.d.f. F_c, then a uniform distribution in the simplex corresponds to a (discretely approximated) distribution F_c for c.

For the posterior of π_c, consider first the extreme case where the data perfectly reveals the value of c_j from each observation $\mathbf{X}_j$. The posterior distribution is then as follows: Let $J_i \in \{0, 1, \ldots, n\}$ count the number of c_j equal to c_i, $J_i = \sum_{j=1}^n \mathbf{1}[c_j = c_i]$. Let W_i be independent chi-squared random variables with $J_i + \nu/m$ degrees of freedom, $i = 1, \ldots, m$. Then the posterior distribution for π_c is equal to $\pi_c(c_i) = W_i / \sum_{l=1}^m W_l$. If n is large, the variation in the posterior distribution is dominated by the J_i terms, and, as is appropriate, the posterior c.d.f. converges to the empirical distribution of $\{c_j\}_{j=1}^n$ for any value of ν.

In general, the data $\mathbf{X}_j$ will not reveal the value of c_j perfectly. But it is straightforward to combine the above posterior characterization in a Gibbs sampler: Conditional on π_c and σ_j, the posterior probability mass function for c_j is proportional to $\pi_c(\cdot)(\Omega_{XX}^{LTU}(\cdot))^{-1/2} \exp[-T\frac{1}{2}\mathbf{X}_j' \Omega_{XX}^{LTU}(\cdot)^{-1}\mathbf{X}_j/\sigma_j^2]$, which is easy to draw from. Conditional on c_j, we can draw σ_j similar to Example MEAN(c). And conditional on $\{c_j\}_{j=1}^m$, drawing π_c is just as described in the last paragraph.

This approach may be extended to vector valued parameters. For instance, in the real exchange example, we could discretize both $\sigma_j \in \{\sigma_i\}_{i=1}^{m_\sigma}$ and $c_j \in \{c_i\}_{i=1}^{m_c}$, and posit two independent Dirichlet priors on the m_σ and m_c dimensional simplexes, respectively. If one would like to allow for the possibility that there is a systematic relationship between the variability of the exchange rate and the amount of mean reversion, then one could treat

the pair of parameters (σ_j, c_j) as being independent in the prior across j conditionally on a joint probability mass function that takes on values in the $m_\sigma m_c$ simplex (and with a regular grid $(\sigma_j, c_j) \in \{\sigma_i\}_{i=1}^{m_\sigma} \times \{c_i\}_{i=1}^{m_c}$, the Dirichlet prior shrinks towards a probability mass function that implies independence between σ_j and c_j). Another generalization is a Dirichlet Process prior relative to some continuously distributed base distribution, which avoids the discretization into a grid of values. See for instance, Neal (2000) for an introduction.

7 Frequentist Analysis: Concepts and Methods

This section discusses several frequentist concepts with a focus on methods for constructing powerful hypothesis tests and efficient confidence intervals. The celebrated Neyman-Pearson (NP) lemma is the cornerstone for these methods. NP tests are designed for situations in which both the null and alternative are *simple*, in the sense that they completely specify the probability distribution of the random variable under study. In practice, hypotheses are often *composite*, in the sense that they restrict, but do not completely specify the probability distribution. Complications associated with composite hypotheses can be handled by averaging over the probability distributions, leading to concepts called *weighted average power* and *least favorable distributions*, or by imposing *invariance* restrictions. This section provides a detailed discussion of these frequentist concepts. The section's running examples are low-frequency inference problems, but aside from these examples, the discussion is general and can be read independently of the other sections of this chapter.

7.1 Hypothesis Tests and Confidence Sets

Bayesian analysis uses the formalism of probability to make statements about the distribution of parameter values in a given sample. In contrast, frequentist analysis seeks to provide guarantees about the properties of inference in repeated samples from the same model and parameters.

As in the last section, the data $\mathbf{Y} \in \mathcal{Y}$ has density $f(\mathbf{y}|\theta)$, indexed by the parameter $\theta \in \Theta \subset \mathbb{R}^k$. Let $\Theta_0, \Theta_1 \subset \Theta$ be proper subsets of the parameter space Θ. A *hypothesis test* φ of

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1 \tag{71}$$

is a function $\varphi : \mathcal{Y} \mapsto [0,1]$, where $\varphi(\mathbf{y}) = 1$ means "H_0 is rejected" and $\varphi(\mathbf{y}) = 0$ means "H_0 is not rejected" for the realization $\mathbf{Y} = \mathbf{y}$. It is mathematically convenient to allow tests $\varphi(\mathbf{y})$ to take on values between zero and one, which are interpreted as the probability of rejecting H_0 given data $\mathbf{Y} = \mathbf{y}$, since this ensures the existence of most powerful tests in great generality, as discussed in the next section. In practice, however, such *randomized tests* are unattractive, as different researchers are not guaranteed to arrive at the same conclusion after having observed the same data.

In this notation, the rejection probability of a test is simply its expectation, $E_\theta[\varphi(\mathbf{Y})] = \int \varphi(\mathbf{y}) f(\mathbf{y}|\theta) d\mathbf{y}$, where the θ subscript in the expectation denotes the data generating value. The *size* of the test is the largest rejection probability under the null hypothesis, $\sup_{\theta \in \Theta_0} E_\theta[\varphi(\mathbf{Y})]$, and a test is of *level* α if its size is smaller or equal to α. The *power function* of the test is given by the $\Theta_1 \mapsto [0,1]$ function $E_\theta[\varphi(\mathbf{Y})]$. The aim in hypothesis testing is to construct a test φ with high power given a particular level α, such as the commonly used value of $\alpha = 0.05$. A low pre-specified level α guarantees that mistaken rejections of H_0 are rare in repeated samples, no matter which value $\theta \in \Theta_0$ generated the data.

An important special case of a hypothesis test specifies the value of a parameter of interest $\gamma = h(\theta)$, where $h : \Theta \mapsto \Gamma \subset \mathbb{R}$. The null hypothesis then becomes $H_0 : \gamma = \gamma_0$, or, in the above notation, $H_0 : \theta \in \Theta_0 = \{\theta : h(\theta) = \gamma_0\}$. Suppose now that we have a family of non-randomized level α hypothesis tests $\varphi_{\gamma_0} : \mathcal{Y} \mapsto \{0,1\}$ indexed by the value of γ under the null hypothesis, so that for each γ_0, $E_\theta[\varphi_{\gamma_0}(\mathbf{Y})] \leq \alpha$ for all $\theta \in \{\theta : h(\theta) = \gamma_0\}$. Given a particular realization $\mathbf{Y} = \mathbf{y}$, suppose we collect the values of γ_0 for which the test φ_{γ_0} does not reject in a set $\hat{\Gamma}(\mathbf{y}) = \{\gamma_0 : \varphi_{\gamma_0}(\mathbf{y}) = 0\}$, so that $\hat{\Gamma} : \mathcal{Y} \mapsto \mathcal{G}$ maps data into Borel subsets of Γ. This *inversion* of a family of level α tests φ_{γ_0} yields a *confidence set of level* $1 - \alpha$, since in repeated samples, $\hat{\Gamma}(\mathbf{Y})$ contains the true value of γ with probability of at least $1 - \alpha$:

$$P_\theta(\gamma_0 \in \hat{\Gamma}(\mathbf{Y})) = 1 - P_\theta(\varphi_{\gamma_0}(\mathbf{Y}) = 1) \geq 1 - \alpha \text{ for all } \theta \in \{\theta : h(\theta) = \gamma_0\} \text{ and } \gamma_0 \in \Gamma. \quad (72)$$

Intuitively, the set $\hat{\Gamma}$ is constructed by trying out all values of $\gamma_0 \in \Gamma$, so at some point, the true value is considered. But by definition of a level α test, the true value is then rejected with probability of at most α. Thus, with at least $1 - \alpha$ probability over repeated samples, the true value is contained in the set $\hat{\Gamma}(\mathbf{Y})$. Similarly, given a confidence set $\hat{\Gamma}$ satisfying the inequality in (72), we can define a corresponding family of tests via $\varphi_{\gamma_0}(\mathbf{y}) = \mathbf{1}[\gamma_0 \notin \hat{\Gamma}(\mathbf{y})]$,

which by construction is of level α. There is thus a perfect equivalence between a family of level α hypothesis tests and a level $1-\alpha$ confidence set, as one can always obtain one from the other.

7.2 Most Powerful Tests

7.2.1 Neyman-Pearson Lemma

If a null or alternative hypothesis fully specifies the data density, then the hypothesis is called *simple*, and otherwise it is called *composite*. The most straightforward hypothesis testing problem involves a simple null hypothesis and a simple alternative hypothesis,

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1. \tag{73}$$

For this case, the Neyman-Pearson Lemma shows that the most powerful test rejects for large values of the likelihood ratio statistic $f(\mathbf{y}|\theta_1)/f(\mathbf{y}|\theta_0)$. Intuitively, the region of $\mathbf{y}$ where $f(\mathbf{y}|\theta_1)/f(\mathbf{y}|\theta_0)$ is large provides the largest rejection probability under θ_1 for a given constraint on the rejection probability under θ_0.

Lemma 1. *The most powerful level α test of (73) is of the form*

$$\varphi^*(\mathbf{y}) = \begin{cases} 1 \text{ if } f(\mathbf{y}|\theta_1) > cv f(\mathbf{y}|\theta_0) \\ \kappa \text{ if } f(\mathbf{y}|\theta_1) = cv f(\mathbf{y}|\theta_0) \\ 0 \text{ if } f(\mathbf{y}|\theta_1) < cv f(\mathbf{y}|\theta_0) \end{cases}$$

where $0 \le \kappa \le 1$ and $cv \ge 0$ are such that $E_{\theta_0}[\varphi(\mathbf{Y})] = \alpha$, that is, for any other level α test φ, $E_{\theta_1}[\varphi^*(\mathbf{Y})] \ge E_{\theta_1}[\varphi(\mathbf{Y})]$.

Proof. If $cv = 0$, $E_{\theta_1}[\varphi^*(\mathbf{Y})] = \int f(\mathbf{y}|\theta_1)d\mathbf{y} = 1$, so there is nothing to prove. Thus assume $cv > 0$ in the following. Let φ be any other level α test. Then by definition of φ^*, $(f(\mathbf{y}|\theta_1) - cv f(\mathbf{y}|\theta_0))(\varphi^*(\mathbf{y}) - \varphi(\mathbf{y})) \ge 0$ for all $\mathbf{y} \in \mathcal{Y}$. Therefore

$$\int f(\mathbf{y}|\theta_1)(\varphi^*(\mathbf{y}) - \varphi(\mathbf{y}))d\mathbf{y} - cv \int f(\mathbf{y}|\theta_0)(\varphi^*(\mathbf{y}) - \varphi(\mathbf{y}))d\mathbf{y} \ge 0. \tag{74}$$

From $\int f(\mathbf{y}|\theta_0)\varphi^*(\mathbf{y})d\mathbf{y} = \alpha$ and $\int f(\mathbf{y}|\theta_0)\varphi(\mathbf{y})d\mathbf{y} \le \alpha$, $\int f(\mathbf{y}|\theta_0)(\varphi^*(\mathbf{y}) - \varphi(\mathbf{y}))d\mathbf{y} \ge 0$. Thus (74) implies $\int f(\mathbf{y}|\theta_1)(\varphi^*(\mathbf{y}) - \varphi(\mathbf{y}))d\mathbf{y} \ge 0$ or equivalently, $\int f(\mathbf{y}|\theta_1)\varphi^*(\mathbf{y})d\mathbf{y} \ge \int f(\mathbf{y}|\theta_1)\varphi(\mathbf{y})d\mathbf{y}$, which was to be shown. ■

When $f(\mathbf{Y}|\theta_1)/f(\mathbf{Y}|\theta_0)$ has a continuous distribution under θ_0, then no randomization via κ is necessary, and the optimal test $\varphi^*(\mathbf{y}) = \mathbf{1}[f(\mathbf{y}|\theta_1) \geq cv f(\mathbf{y}|\theta_0)]$ is simply characterized by the *critical value* cv.[19]

Example PERS(a). Consider inference about the value of c in the local-to-unity model, using the cosine transform as the observations. Then under approximation (13), we have

$$\mathbf{X}_T \sim \mathcal{N}(0, T^{-1}\sigma^2 \mathbf{\Omega}_{XX}^{LTU}(c)). \tag{75}$$

For now, suppose $\sigma^2 = 1$ is known; we discuss the unknown σ^2 case in Section 7.3 below. By Lemma 1, the most powerful test of $H_0 : c = c_0$ against $H_1 : c = c_1$ rejects for large values of

$$\frac{|\mathbf{\Omega}_{XX}^{LTU}(c_1)|^{-1/2}}{|\mathbf{\Omega}_{XX}^{LTU}(c_0)|^{-1/2}} \frac{\exp[-\frac{1}{2}T\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LTU}(c_1)^{-1}\mathbf{X}_T]}{\exp[-\frac{1}{2}T\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LTU}(c_0)^{-1}\mathbf{X}_T]}$$

or, equivalently, for large values of the test statistic

$$T\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LTU}(c_0)^{-1}\mathbf{X}_T - T\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LTU}(c_1)^{-1}\mathbf{X}_T \tag{76}$$

Since (76) has a continuous distribution under any $c \geq 0$, the optimal test rejects if and only if (76) is larger than the critical value. The critical value is simply the $1-\alpha$ quantile of (76) with $\mathbf{X}_T$ drawn from (75) with $c = c_0$. In particular, for $c_0 = 0$, this yields the point-optimal unit root test with $\sigma^2 = 1$ known. ▲

7.2.2 Weighted Average Power Maximizing Tests

Suppose the null hypothesis is simple, but the alternative hypothesis is composite

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1. \tag{77}$$

The NP lemma provides the form of the best test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, for a specific value of $\theta_1 \in \Theta_1$. If these tests happen to be the same function of $\mathbf{y}$ for all values of $\theta_1 \in \Theta_1$, then this test is *uniformly most powerful*. In many instances, however, a uniformly most powerful test does not exist, since the power functions of the NP tests cross. Sometimes the *point-optimal* NP test that maximizes power against a single alternative $\theta_1 \in \Theta_1$ turns out to have good power against all alternatives; see King (1987) for discussion and examples.

[19] In practice, cv may be approximated by taking N i.i.d. draws $\mathbf{Y}^{(l)}$ under $\theta = \theta_0$, and by computing the $1-\alpha$ quantile of $\{f(\mathbf{Y}^{(l)}|\theta_1)/f(\mathbf{Y}^{(l)}|\theta_0)\}_{l=1}^N$.

A more systematic approach to handle composite alternatives draws on the classic decision theory approach of minimizing weighted average risk. In the context of (77), this leads to the objective of maximizing *weighted average power*, a solution concept that was prominently applied in econometrics by Andrews and Ploberger (1994). Formally, with a weighting function that correspond to the c.d.f.[20] F on Θ_1, the weighted average power of a test φ is

$$\text{WAP} = \int_\Theta E_\theta[\varphi(\mathbf{Y})]dF(\theta).$$

The WAP criterion is a scalar summary of the power properties of φ, with F describing the importance of various alternatives. By a change of the order of integration, we obtain

$$\begin{aligned}\text{WAP} &= \int_\Theta \int \varphi(\mathbf{y}) f(\mathbf{y}|\theta) d\mathbf{y} \cdot dF(\theta) \\ &= \int \varphi(\mathbf{y}) \int_\Theta f(\mathbf{y}|\theta) dF(\theta) \cdot d\mathbf{y} \\ &= \int \varphi(\mathbf{y}) f_1(\mathbf{y}) d\mathbf{y}\end{aligned}$$

with $f_1(\mathbf{y}) = \int_\Theta f(\mathbf{y}|\theta)dF(\theta)$. Note that f_1 is a probability density function, since it is nonnegative and $\int f_1(\mathbf{y})d\mathbf{y} = \int \int_\Theta f(\mathbf{y}|\theta)dF(\theta) \cdot d\mathbf{y} = \int_\Theta \int f(\mathbf{y}|\theta)d\mathbf{y} \cdot dF(\theta) = 1$. Thus maximizing WAP is equivalent to maximizing power against the single alternative H_1^*: "the density of $\mathbf{Y}$ is given by f_1." But the NP lemma provides the form of the optimal test between two single alternatives. The WAP maximizing test of (77) thus rejects for large values of

$$\frac{f_1(\mathbf{y})}{f(\mathbf{y}|\theta_0)} = \frac{\int f(\mathbf{y}|\theta)dF(\theta)}{f(\mathbf{y}|\theta_0)}$$

and if $f_1(\mathbf{Y})/f(\mathbf{Y}|\theta_0)$ has a continuous distribution, then the WAP maximizing test is of the simple form $\mathbf{1}[f_1(\mathbf{y}) > cvf(\mathbf{y}|\theta_0)]$.

Example PERS(b). The weighted average power maximizing test of $H_0 : c = c_0$ rejects for large values of

$$\frac{\int |\mathbf{\Omega}_{XX}^{LTU}(c)|^{-1/2} \exp[-\frac{1}{2}T\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{X}_T]dF(c)}{|\mathbf{\Omega}_{XX}^{LTU}(c_0)|^{-1/2} \exp[-\frac{1}{2}T\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LTU}(c_0)^{-1}\mathbf{X}_T]}$$

[20] We allow F to be the c.d.f. of a continuous or discrete random variable, and write the expectation of the function $\psi : \Theta \mapsto \mathbb{R}$ with respect to F as the Riemann-Stieltjes integral $\int \psi(\theta)dF(\theta)$. If F is characterized by the p.d.f. f, then $\int \psi(\theta)dF(\theta)$ corresponds to $\int \psi(\theta)f(\theta)d\theta$, and if F describes a discrete distribution with support $\{\theta_i\}_{i=1}^m$ and p.m.f. $f(\theta)$, then $\int \psi(\theta)dF(\theta)$ is shorthand for $\sum_{i=1}^m \psi(\theta_i)f(\theta_i)$.

or, equivalently, for large values of

$$\int |\mathbf{\Omega}_{XX}^{LTU}(c)|^{-1/2} \exp[-\tfrac{1}{2}T\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{X}_T + \tfrac{1}{2}T\mathbf{X}_T'\mathbf{\Omega}_{XX}^{LTU}(c_0)^{-1}\mathbf{X}_T]dF(c) \tag{78}$$

with critical value equal to the $1-\alpha$ quantile of (78) with $\mathbf{X}_T$ distributed as (75) under $c = c_0$. ▲

Now consider the construction of confidence intervals. Specifically, suppose $k = 1$, so that $\theta \in \Theta \subset \mathbb{R}$, and assume that $\gamma = h(\theta)$ is a one-to-one transformation with inverse function $h^{-1} : \Gamma \mapsto \Theta$. Let $\varphi_{\gamma_0}^*$ be a level α non-randomized WAP maximizing test of $H_0 : h(\theta) = \gamma_0$, which is equivalent to $H_0 : \theta = \theta_0 = h^{-1}(\gamma_0)$, with weighting function F on Θ that does not depend on γ_0. Let $\hat{\Gamma}^*(\mathbf{y}) = \{\gamma_0 : \varphi_{\gamma_0}^*(\mathbf{y}) = 0\}$ be the resulting level $1-\alpha$ confidence set. Note that the length of $\hat{\Gamma}^*(\mathbf{y})$ is given by $\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}^*(\mathbf{y})]d\gamma_0 = \int(1-\varphi_{\gamma_0}^*(\mathbf{y}))d\gamma_0$. The expected length of $\hat{\Gamma}^*$ under θ is thus given by $E_\theta[\int(1-\varphi_{\gamma_0}^*(\mathbf{Y}))d\gamma_0]$, and F-weighted average expected length is equal to (cf. Pratt (1961))

$$\int_\Theta E_\theta\left[\int(1-\varphi_{\gamma_0}^*(\mathbf{Y}))d\gamma_0\right]dF(\theta) = \int\int(1-\varphi_{\gamma_0}^*(\mathbf{y}))f_1(\mathbf{y})d\mathbf{y}d\gamma_0. \tag{79}$$

Since $\varphi_{\gamma_0}^*$ maximizes $\int \varphi_{\gamma_0}^*(\mathbf{y})f_1(\mathbf{y})d\mathbf{y}$ for all γ_0 among all level α tests, we conclude that the confidence set $\hat{\Gamma}^*$ minimizes F-weighted average expected length among all level $1-\alpha$ confidence sets.

Example PERS(c). The F-weighted average expected length minimizing confidence set for c simply collects all values of c_0 for which the test based on (78) does not reject. Note that for a given F, this requires determination of the critical values of (78) for all c_0. ▲

7.2.3 Least Favorable Distributions

Now suppose the null hypothesis is composite, and the alternative hypothesis is simple,

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \text{the density of } \mathbf{Y} \text{ is } f_1. \tag{80}$$

While we use the density formulation for the alternative hypothesis from the last subsection, $f_1(\mathbf{y}) = f(\mathbf{y}|\theta_1)$ is included as a special case.

Consider this testing problem from the perspective of game theory: An econometrician seeks to discriminate between H_0 and H_1 using data $\mathbf{Y}$, and plays against an adversarial

nature that controls the value of θ under the null hypothesis. A randomized strategy of nature corresponds to a distribution Λ with support in Θ_0. The econometrician's optimal response to nature playing Λ is to use the NP test φ_Λ of

$$H_\Lambda : \text{the density of } \mathbf{Y} \text{ is } f_\Lambda(\mathbf{y}) = \int f(\mathbf{y}|\theta) d\Lambda(\theta) \quad \text{against} \quad H_1 : \text{the density of } \mathbf{Y} \text{ is } f_1. \tag{81}$$

Note that any level α test of H_0 is necessarily a level α test of H_Λ, for any Λ, since $\sup_{\theta\in\Theta_0} E_\theta[\varphi(\mathbf{Y})] \leq \alpha$ implies $\int \varphi(\mathbf{y}) f_\Lambda(\mathbf{y}) d\mathbf{y} = \int E_\theta[\varphi(\mathbf{Y})] d\Lambda(\theta) \leq \alpha$. The level constraint in (80) is therefore more severe than in (81), so that the econometrician's best response to nature playing Λ generically allows for a more powerful test. Nature knows this, so its best strategy is to play a *least favorable distribution* Λ^* that induces the best response φ_Λ to also satisfy the more severe level constraint $\sup_{\theta\in\Theta_0} E_\theta[\varphi(\mathbf{Y})] \leq \alpha$, and thus induces low power.

More formally, we have the following result taken from Elliott et al. (2015) (also see Theorem 3.8.1 of Lehmann and Romano (2005)).

Theorem 1. *Suppose Λ and Λ^* are probability distributions with support equal to a subset of Θ_0.*

(a) Let φ be any level α test of (80). For any Λ, the best level α test φ_Λ of (81) has at least as much power as φ.

(b) If Λ^* is such that the best level α test of (81) with $\Lambda = \Lambda^*$, φ_{Λ^*}, is also of level α in (80), then φ_{Λ^*} is the best level α test of H_0.

Proof. (a) The test φ is also of level α under H_Λ, since

$$\begin{aligned} \int \varphi(\mathbf{y}) f_\Lambda(\mathbf{y}) d\mathbf{y} &= \int_{\Theta_0} \int \varphi(\mathbf{y}) f(\mathbf{y}|\theta) d\mathbf{y} \cdot d\Lambda(\theta) \\ &\leq \sup_{\theta\in\Theta_0} \int \varphi(\mathbf{y}) f(\mathbf{y}|\theta) d\mathbf{y} \leq \alpha. \end{aligned}$$

But φ_Λ is the best level α test of H_Λ against H_1, so its power is no smaller than the power of φ.

(b) From part (a), no level α test φ of (80) can have higher power than φ_{Λ^*}. ■

Note that part (a) of Theorem 1 provides a set of upper bounds on power in the original problem (80), indexed by arbitrary probability distributions Λ with support in Θ_0. Since in many problems it is hard to analytically derive a least favorable distribution, these upper

bounds provide a useful benchmark for the relative efficiency of a numerically determined test. We discuss how to exploit this for the construction of a demonstrably nearly power maximizing test in Sections 7.5.2 and 7.5.3 below.

Example MEAN(h). Reconsider inference about the mean in the local-to-unity model, but now viewed from a frequentist perspective. Suppose $\sigma^2 = 1$ is known, so that

$$\mathbf{X}_T^0 \sim \mathcal{N}(\mu\iota_{q+1}, T^{-1}\mathbf{\Omega}^{LTU}(c)). \tag{82}$$

Suppose we want to test $H_0 : \mu = \mu_0$, and maximize weighted average power for some weighting function F that corresponds to a probability distribution over (μ, c). By the reasoning of Section 7.2.2, this corresponds to maximizing power against the single alternative H_1 : the density of $\mathbf{X}_T^0$ is $f_1(\mathbf{x}_T^0)$, where

$$f_1(\mathbf{x}_T^0) \propto \int |\mathbf{\Omega}^{LTU}(c)|^{-1/2} \exp\left[-\tfrac{1}{2}T(\mathbf{x}_T^0 - \mu\iota_{q+1})'\mathbf{\Omega}^{LTU}(c)^{-1}(\mathbf{x}_T^0 - \mu\iota_{q+1})\right] dF(\mu, c)$$

and maximizing power against a single alternative $H_1 : (\mu, c) = (\mu_1, c_1)$ is a special case with F degenerate at the point (μ_1, c_1).

This hypothesis testing problem has the nuisance parameter c under the null hypothesis. Let Λ be a probability distribution for c, such as a distribution that puts all its mass on a single value c_0, or a distribution that puts 30% probability on $c_{0,1}$ and 70% probability on $c_{0,2}$, and so forth. By the NP lemma, the best level α test φ_Λ of H_Λ :"$\mu = 0$ and $c \sim \Lambda$" against H_1 rejects for large values of

$$\frac{\int |\mathbf{\Omega}^{LTU}(c)|^{-1/2} \exp\left[-\tfrac{1}{2}T(\mathbf{X}_T^0 - \mu\iota_{q+1})'\mathbf{\Omega}^{LTU}(c)^{-1}(\mathbf{X}_T^0 - \mu\iota_{q+1})\right] dF(\mu, c)}{\int |\mathbf{\Omega}^{LTU}(c)|^{-1/2} \exp\left[-\tfrac{1}{2}T(\mathbf{X}_T^0 - \mu_0\iota_{q+1})'\mathbf{\Omega}^{LTU}(c)^{-1}(\mathbf{X}_T^0 - \mu_0\iota_{q+1})\right] d\Lambda(c)}, \tag{83}$$

and the critical value is equal to the $1-\alpha$ quantile of (83) with $\mathbf{X}_T^0$ distributed as (82) under $\mu = \mu_0$ and c randomly drawn from Λ. The power of this test φ_Λ is the probability that it exceeds the critical value with $\mathbf{X}_T^0$ distributed as (82) under (μ, c) randomly drawn from F. By Theorem 1 (a), there cannot exist a level α test φ of $H_0 : \mu = \mu_0$ with $c > 0$ whose power against H_1 exceeds the power of φ_Λ, and this holds for any choice of Λ. Part (b) says that if the test φ_Λ happened to also be of level α under the composite null hypotheses H_0 :"$\mu = \mu_0$, $c > 0$", then φ_Λ is in fact the best level α test of H_0 against H_1. ▲

Consider the related confidence interval problem. Note that if f_1 in (80) is equal to $f_1(\mathbf{y}) = \int f(\mathbf{y}|\theta)dF(\theta)$, then φ_{Λ^*} of Theorem 1 (b) is the WAP maximizing test of (71).

Furthermore, similar to the discussion at the end of Section 7.2.2, consider the family of hypothesis tests indexed by $\gamma_0 \in \Gamma$

$$H_0 : h(\theta) = \gamma_0 \quad \text{against} \quad H_1 : \text{the density of } \mathbf{Y} \text{ is } \int_\Theta f(\mathbf{y}|\theta) dF(\theta). \tag{84}$$

Then as in (79), the F-weighted average expected length of a level $1-\alpha$ confidence set $\hat{\Gamma}(\mathbf{y}) = \{\mathbf{y} : \varphi_{\gamma_0}(\mathbf{y}) = 0\}$ obtained by inverting level α tests of (84) is equal to

$$\int_\Theta E_\theta[\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{y})] d\gamma_0] dF(\theta) = \int\int (1 - \varphi_{\gamma_0}(\mathbf{y})) f_1(\mathbf{y}) d\mathbf{y} d\gamma_0. \tag{85}$$

Let $\varphi_{\Lambda_{\gamma_0}}$ be the non-randomized level α test of of the null hypothesis $H_{\Lambda_{\gamma_0}}$: the density of $\mathbf{Y}$ is $\int f(\mathbf{y}|\theta) d\Lambda_{\gamma_0}(\theta)$ for some distribution Λ_{γ_0} with support on $\{\theta : h(\theta) = \gamma_0\}$. By Theorem 1 (a), $\int (1 - \varphi_{\gamma_0}(\mathbf{y})) f_1(\mathbf{y}) d\mathbf{y} \geq \int (1 - \varphi_{\Lambda_{\gamma_0}}(\mathbf{y})) f_1(\mathbf{y}) d\mathbf{y}$, so $\int\int (1 - \varphi_{\varphi_{\Lambda_{\gamma_0}}}(\mathbf{y})) f_1(\mathbf{y}) d\mathbf{y} d\gamma_0$ is a lower bound for the F-weighted average expected length for any level $1-\alpha$ confidence set of γ. Thus, if there exists a family of least favorable distribution $\Lambda^*_{\gamma_0}$ and corresponding level α non-randomized tests $\varphi_{\Lambda^*_{\gamma_0}}$, then the resulting confidence set minimizes (85) among all level $1-\alpha$ confidence sets. We conclude that length optimal confidence sets are obtained by determining a family of least favorable distributions for the family of hypothesis tests (84).

Example MEAN(i). If for each μ_0, we specify an arbitrary distribution $\Lambda = \Lambda_{\mu_0}$ on c, then the F-weighted expected length of the set for μ obtained by inverting level $1-\alpha$ tests based on (83) is a lower bound on the F-weighted expected length of any valid level $1-\alpha$ confidence set. In absence of specific information about plausible values of μ, it might be unattractive to having to specify a weighting function for μ. An alternative approach based on invariance is discussed in Section 7.3 below. ▲

Example BRK(a). Consider the problem of inference for the break date, as in Section 4.3.2. Suppose for now that the pre-break mean $\mu = 0$ and the variance $\sigma^2 = 1$ are known, so that from (44)

$$\mathbf{X}^0_T \sim \mathcal{N}\left(\delta \mathbf{v}^0(r), T^{-1}\mathbf{I}_{q+1}\right). \tag{86}$$

Let F be the weighting function on (r, δ) such that under F, $\delta \sim \mathcal{N}(0, T^{-1}\varpi^2)$ and $r \sim \mathcal{U}[0,1]$. Thus, under F and conditional on r,

$$\mathbf{X}^0_T \sim \mathcal{N}\left(0, T^{-1}(\mathbf{I}_{q+1} + \varpi^2 \mathbf{v}^0(r)\mathbf{v}^0(r)')\right).$$

The best level α test of $H_0 : r = r_0$, $\delta \sim \Lambda_{r_0}$ against $H_1 : (r, \delta) \sim F$ then rejects for large values of

$$\frac{\int_0^1 |\mathbf{I}_{q+1} + \varpi^2 \mathbf{v}^0(r)\mathbf{v}^0(r)'|^{-1/2} \exp[-\frac{1}{2}T\mathbf{X}_T^{0\prime}(\mathbf{I}_{q+1} + \varpi^2 \mathbf{v}^0(r)\mathbf{v}^0(r)')^{-1}\mathbf{X}_T^0]dr}{\int \exp[-\frac{1}{2}T(\mathbf{X}_T^0 - \delta\mathbf{v}^0(r))'(\mathbf{X}_T^0 - \delta\mathbf{v}^0(r))]d\Lambda_{r_0}(\delta)} \tag{87}$$

with critical value equal to the $1-\alpha$ quantile of (87) with $\mathbf{X}_T^0$ distributed as (86) with $r = r_0$ and δ drawn randomly from Λ_{r_0}. The F-weighted expected length of any valid level $1 - \alpha$ confidence set for r is bounded below by the expected length of the set obtained by collecting the values of r_0 for which the tests based on (87) don't reject. ▲

7.3 Invariance

As already briefly mentioned in the discussion of invariant priors, in many inference problems it can make sense to impose the restriction that a transformation of the data leads to a corresponding transformation of confidence sets, or does not alter the decision of a hypothesis test. For example, it might be reasonable to impose the restriction that alternative units of measurement of the data $\mathbf{Y}$ should not affect the decision to reject a hypothesis. In the context of frequentist inference, invariance can be useful because it often reduces the dimension of the effective parameter space, which in turn simplifies the inference problem. See Chapter 6 of Lehmann and Romano (2005), Chapter 3 of Lehmann and Casella (1998) and Chapter 6 of Berger (1985) for alternative expositions and additional references.

7.3.1 Groups and Transformations

First, some background on transformations. Let $g : A \times \mathcal{Y} \mapsto \mathcal{Y}$ be a group of transformations of the data, with group actions indexed by $a \in A$. Recall that a group satisfies the three properties: (i) for all $a_1, a_2 \in A$, there exists $a_3 \in A$ such that $g(a_2, g(a_1, \mathbf{y})) = g(a_3, \mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$; (ii) for all $a \in A$, there exists an inverse element $a^{-1} \in A$ such that $g(a^{-1}, g(a, \mathbf{y})) = \mathbf{y}$ for all $\mathbf{y} \in \mathcal{Y}$; and (iii) there exists the identity transformation, that is for some $a \in A$, $g(a, \mathbf{y}) = \mathbf{y}$ for all $\mathbf{y} \in \mathcal{Y}$. (Associativity is guaranteed, because a composition of functions is always associative.)

Let $f(\cdot|\theta)$ denote the density of $\mathbf{Y}$. Assume that the probability model is *formally invariant* in the sense that the density of $g(a, \mathbf{Y})$ is $f(\cdot|\bar{g}(a, \theta))$ for all $a \in A$ and $\theta \in \Theta$ and some class of transformations $\bar{g} : A \times \Theta \mapsto \Theta$; said differently, the density of $\mathbf{Y}$ under

θ is the same as the density of $g(a, \mathbf{Y})$ under $\bar{g}(a, \theta)$. In other words, the distribution of transformed data is equal to the distribution of untransformed data under a transformed parameter value. As is easily checked, the class of transformations $\bar{g}$ on the parameter space then also forms a group.

Finally, assume that, as the parameter θ is transformed to $\bar{g}(a, \theta)$, the parameter of interest $\gamma = h(\theta)$ is transformed in a way that only depends on a and $h(\theta)$, that is

$$h(\bar{g}(a, \theta)) = \hat{g}(a, h(\theta)), \text{ for all } \theta \in \Theta \text{ and } a \in A \tag{88}$$

for some $\hat{g} : A \times \Gamma \mapsto \Gamma$. The class of transformations $\hat{g}$ again forms a group. We assume in the following that $\hat{g}$ is one-to-one in the sense that for all $a \in A$ and $\gamma_1, \gamma_2 \in \Gamma$

$$\hat{g}(a, \gamma_1) = \hat{g}(a, \gamma_2) \text{ implies } \gamma_1 = \gamma_2. \tag{89}$$

The parameter of interest remains unaffected by the transformation if $\gamma \mapsto \hat{g}(a, \gamma)$ is the identity transformation, that is, if $\hat{g}(a, h(\theta)) = h(\theta)$ for all θ and $a \in A$. In that case one may want to restrict attention to tests of the null hypothesis $H_0 : h(\theta) = \gamma_0$ that remain correspondingly *invariant*, that is

$$\varphi(g(a, \mathbf{y})) = \varphi(\mathbf{y}), \text{ for all } a \in A. \tag{90}$$

Example PERS(d). Recall the problem of inference about c in the local-to-unity model with observation $\mathbf{X}_T \sim \mathcal{N}(0, T^{-1}\sigma^2 \mathbf{\Omega}_{XX}^{LTU}(c))$, but without assuming σ^2 is known. This problem is indexed by the two dimensional parameter $\theta = (c, \sigma) \in (0, \infty)^2$. Consider the group of scale transformations $g(a, \mathbf{x}_T) = a\mathbf{x}_T$, for $a \in A = (0, \infty)$. These transformations form a group, since (i) $g(a_2, g(a_1, \mathbf{x}_T)) = g(a_2 a_1, \mathbf{x}_T)$, (ii) $1/a$ is the inverse element of the transformation indexed by a and (iii) $a = 1$ is the identity transformation. Furthermore, from $g(a, \mathbf{X}_T) \sim \mathcal{N}(0, T^{-1}a^2\sigma^2 \mathbf{\Omega}_{XX}^{LTU}(c))$, we see that the problem is formally invariant with $\bar{g}(a, \theta) = (c, a\sigma)$. With c the parameter of interest, $h(\theta) = \gamma = c$, trivially $h(\bar{g}(a, \theta)) = c$ for all $a \in A$, so that $\hat{g}(a, h(\theta)) = c$ is the identity transformation. Thus, one might want to restrict attention to tests of $H_0 : c = c_0$ that are scale invariant, $\varphi(a\mathbf{x}_T) = \varphi(\mathbf{x}_T)$ for all $a > 0$. ▲

In other problems, the transformations affect the parameter of interest, that is $\hat{g}(a, \gamma)$ is not the identity transformation. It then does not make sense to impose (90). Rather, one

might reasonably demand that the decision to reject $H_0 : h(\theta) = \gamma_0$ when observing $\mathbf{Y} = \mathbf{y}$ should be the same as the decision to reject $H_0 : h(\theta) = \hat{g}(a, \gamma_0)$ when observing $\mathbf{Y} = g(a, \mathbf{y})$, for all $a \in A$. This amounts to a constraint on the entire *family* of hypothesis tests, indexed by γ_0, that define a confidence set $\hat{\Gamma}$ for γ. In particular, a set estimator $\hat{\Gamma} : \mathcal{Y} \mapsto \mathcal{G}$ of $\gamma = h(\theta)$ is *invariant* (or *equivariant*) if

$$\hat{\Gamma}(g(a, \mathbf{y})) = \hat{g}(a, \hat{\Gamma}(\mathbf{y})) \tag{91}$$

for all $a \in A$ and $\mathbf{y} \in \mathcal{Y}$, where $\hat{g}(a, \Gamma_0) = \{\hat{g}(a, \gamma) : \gamma \in \Gamma_0\}$ for all Borel subsets $\Gamma_0 \subset \Gamma$.

Example MEAN(j). Without the assumption that σ^2 is known, the problem becomes one of observing $\mathbf{X}_T^0 \sim \mathcal{N}(\mu\iota_{q+1}, T^{-1}\sigma^2\mathbf{\Omega}^{LTU}(c))$, $\theta = (\mu, \sigma, c) \in \Theta = \mathbb{R} \times (0, \infty)^2$ and $\gamma = h(\theta) = \mu$. Consider the group of transformations that changes the scale and location of $\mathbf{x}_T^0$. Formally, let $a = (a_\mu, a_\sigma) \in A = \mathbb{R} \times (0, \infty)$, so that the three groups introduced above are given by

$$\begin{aligned} g(a, \mathbf{x}_T^0) &= a_\sigma \mathbf{x}_T^0 + \iota_{q+1} a_\mu \\ \bar{g}(a, \theta) &= (a_\sigma \mu + a_\mu, a_\sigma \sigma, c) \\ \hat{g}(a, \gamma) &= a_\sigma \gamma + a_\mu. \end{aligned}$$

This problem is formally invariant, since the distribution (and hence density) of $g(a, \mathbf{X}_T^0) = a_\sigma \mathbf{X}_T^0 + \iota_{q+1} a_\mu \sim \mathcal{N}((a_\sigma \mu + a_\mu)\iota_{q+1}, T^{-1} a_\sigma^2 \sigma^2 \Omega^{LTU}(c))$ corresponds to the distribution of $\mathbf{X}_T^0$ under the parameter $\bar{g}(a, \theta) = (a_\sigma \mu + a_\mu, a_\sigma \sigma, c)$. An invariant confidence set $\hat{\Gamma}(\mathbf{x}_T^0)$ satisfies $\hat{\Gamma}(a_\sigma \mathbf{x}_T^0 + \iota_{q+1} a_\mu) = \{a_\sigma \gamma + a_\mu : \gamma \in \hat{\Gamma}(\mathbf{x}_T^0)\}$, for all $a \in A$ and $\mathbf{x}_T^0 \in \mathbb{R}^{q+1}$. ▲

Note that (91) incudes the special case where $\hat{g}$ is the identity transformation. The invariance requirement of the confidence set $\hat{\Gamma}$ for γ then merely amounts to having to invert a sequence of invariant tests $H_0 : h(\theta) = \gamma_0$, where invariant tests are tests that satisfy (90).

Example PERS(e). All scale invariant confidence sets $\hat{\Gamma}(\mathbf{x}_T) \subset [0, \infty)$ for c, that is, confidence sets that satisfy $\hat{\Gamma}(\mathbf{x}_T) = \hat{\Gamma}(a\mathbf{x}_T)$ for all $a > 0$, can be obtained by inverting a family of scale invariant tests of $H_0 : c = c_0$. ▲

Example BRK(b). Reconsider inference for the break date, but without assuming μ and σ^2 are known, so that $\mathbf{X}_T^0 \sim \mathcal{N}(\mu\iota_{q+1} + \delta\mathbf{v}^0(r), T^{-1}\sigma^2\mathbf{I}_{q+1})$, $\theta = (\mu, \delta, \sigma, r) \in \Theta = \mathbb{R}^2 \times (0, \infty) \times [0, 1]$ and $\gamma = h(\theta) = r$. With $g(a, \mathbf{x}_T^0)$ the group of scale and location transformations, as

in Example MEAN(a) above, we have $\bar{g}(a,\theta) = (a_\sigma\mu + a_\mu, a_\sigma\delta, a_\sigma\sigma, r)$ and $\hat{g}(a,\gamma) = \gamma$. An invariant confidence $\hat{\Gamma}(\mathbf{X}_T^0)$ set for r satisfies $\hat{\Gamma}(a_\sigma\mathbf{x}_T^0 + \iota_{q+1}a_\mu) = \hat{\Gamma}(\mathbf{x}_T^0)$, for all $a \in A$ and $\mathbf{x}_T^0 \in \mathbb{R}^{q+1}$. ▲

7.3.2 Maximal Invariants

Intuitively, transformations by g send a given value $\mathbf{y}_0 \in \mathcal{Y}$ around an orbit $\mathcal{Y}_0 \subset \mathcal{Y}$, that is, any value in $\mathcal{Y}_0$ is equivalent to $\mathbf{y}_0$ up to a transformation, $\mathcal{Y}_0 = \{g(a,\mathbf{y}_0) : a \in A\}$. Thus, any value of $\mathbf{y}$ can be decomposed as

$$\mathbf{y} = g(O(\mathbf{y}), M(\mathbf{y})) \tag{92}$$

with $M : \mathcal{Y} \mapsto \mathcal{Y}$ and $O : \mathcal{Y} \mapsto A$. The function $M(\mathbf{y}_0)$ selects a particular value on the orbit $\mathcal{Y}_0$, and $O(\mathbf{y}_0)$ indexes the group action that recovers the original $\mathbf{y}$ from $M(\mathbf{y}_0)$. Formally, $M(\mathbf{y})$ is a *maximal invariant* that satisfies (i) invariance: $M(g(a,\mathbf{y})) = M(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$, $a \in A$; and (ii) maximality: $M(\mathbf{y}_1) = M(\mathbf{y}_2)$ implies that $\mathbf{y}_1 = g(a,\mathbf{y}_2)$ for some $a \in A$.

Substituting (92) into (90) yields the conclusion that any invariant test φ can be written as a function of $M(\mathbf{y})$,

$$\varphi(\mathbf{y}) = \varphi(M(\mathbf{y})). \tag{93}$$

Similarly, (92) and (91) imply that any invariant set estimator $\hat{\Gamma}$ can be written in the form

$$\hat{\Gamma}(\mathbf{y}) = \hat{g}(O(\mathbf{y}), \hat{\Gamma}(M(\mathbf{y}))). \tag{94}$$

If one is committed to constructing an invariant test or invariant set estimator $\hat{\Gamma}$, it therefore suffices to determine their value on $M(\mathcal{Y}) = \{M(\mathbf{y}) : \mathbf{y} \in \mathcal{Y}\}$, since the values of φ and $\hat{\Gamma}$ for any $\mathbf{y} \in \mathcal{Y}\backslash M(\mathcal{Y})$ are determined by (93) and (94). Thus, the invariance structure makes $M(\mathbf{Y}) \in M(\mathcal{Y})$ the effective observation.

Example PERS(f). Consider scale invariant inference about c in the LTU model. One choice for $M(\mathbf{x}_T)$ is $M(\mathbf{x}_T) = \mathbf{x}_T^s = \mathbf{x}_T/\sqrt{\mathbf{x}_T'\mathbf{x}_T}$, and correspondingly, $O(\mathbf{x}_T) = \sqrt{\mathbf{x}_T'\mathbf{x}_T}$. Thus (91) shows that all scale invariant tests of $H_0 : c = c_0$ can be written as functions of $M(\mathbf{x}_T) = \mathbf{x}_T^s$, and, recalling that $\hat{g}$ is the identity transformation in this example, also all invariant confidence sets for c can be written in the form $\hat{\Gamma}(\mathbf{x}_T) = \hat{\Gamma}(\mathbf{x}_T^s)$. ▲

Example MEAN(k). One choice for $M(\mathbf{x}_T^0)$ is $M(\mathbf{x}_T^0) = (0, \mathbf{x}_T^{s\prime})'$ with $\mathbf{x}_T^s = \mathbf{x}_T/\sqrt{\mathbf{x}_T'\mathbf{x}_T}$, so that $O(\mathbf{x}_T^0) = (\bar{x}_{1:T}, \sqrt{\mathbf{x}_T'\mathbf{x}_T})$. Equation (94) implies that all invariant set estimators for μ are of the form

$$\left\{\bar{x}_{1:T} + \sqrt{\mathbf{x}_T'\mathbf{x}_T}\gamma : \gamma \in \hat{\Gamma}((0, \mathbf{x}_T^{s\prime})')\right\}. \quad \blacktriangle \tag{95}$$

Example BRK(c). With the same choice for $M(\mathbf{x}_T^0)$ and $O(\mathbf{x}_T^0)$ as in Example MEAN(k), we obtain that any invariant confidence set for r is of the form $\hat{\Gamma}((0, \mathbf{x}_T^{s\prime})')$. Note that $\mathbf{X}_T^s \sim \mathbf{X}_T/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$ with $\mathbf{X}_T \sim \mathcal{N}(\delta\mathbf{v}(r), T^{-1}\sigma\mathbf{I}_q)$, where $\mathbf{v}(r)$ are the last q elements of $\mathbf{v}^0(r)$. $\blacktriangle$

We now show how invariance also reduces the effective parameter space. Let $\bar{M} : \Theta \mapsto \Theta$ be the maximal invariant for the group $\bar{g}$ that acts on the parameter space. In analogy to (92), we assume that for some function $\bar{O} : \Theta \mapsto A$

$$\theta = \bar{g}(\bar{O}(\theta), \bar{M}(\theta)). \tag{96}$$

The formal invariance of the problem now yields the following result (cf. Theorem 6.3.2 of Lehmann and Romano (2005) and Lemma 3 of Müller and Norets (2016)).

Lemma 2. *If a problem is formally invariant, then*

(a) the distribution of $M(\mathbf{Y})$ depends on θ only through $\bar{M}(\theta)$;

(b) for any invariant $\hat{\Gamma}$, the distribution of $(\mathbf{1}[h(\theta) \in \hat{\Gamma}(\mathbf{Y})], M(\mathbf{Y}))$ with $\mathbf{Y}$ drawn from density $f(\cdot|\theta)$ is the same as the distribution of $(\mathbf{1}[h(\bar{M}(\theta)) \in \hat{\Gamma}(\mathbf{Y})], M(\mathbf{Y}))$ with $\mathbf{Y}$ drawn from density $f(\cdot|\bar{M}(\theta))$.

Proof.

(a) For any Borel set $\mathcal{B}$ on $\mathcal{Y}$,

$$\begin{aligned} P_\theta(\mathbf{Y} \in \mathcal{B}) &= P_{\bar{g}(\bar{O}(\theta),\bar{M}(\theta))}(\mathbf{Y} \in \mathcal{B}) \quad \text{(by (96))} \\ &= P_{\bar{M}(\theta)}(M(\bar{g}(\bar{O}(\theta), \mathbf{Y})) \in \mathcal{B}) \text{ (by formal invariance)} \\ &= P_{\bar{M}(\theta)}(M(\mathbf{Y}) \in \mathcal{B}) \text{ (by invariance of } M). \end{aligned}$$

(b) Let $\mathcal{B}_0$ be a Borel set on $\{0,1\} \times \mathcal{Y}$. Then

$$\begin{aligned} &P_\theta((\mathbf{1}[h(\theta) \in \hat{\Gamma}(\mathbf{Y})], M(\mathbf{Y})) \in \mathcal{B}_0) \\ &= P_{\bar{g}(\bar{O}(\theta),\bar{M}(\theta))}((\mathbf{1}[h(\bar{g}(\bar{O}(\theta), \bar{M}(\theta))) \in \hat{\Gamma}(\mathbf{Y})], M(\mathbf{Y})) \in \mathcal{B}_0) \quad \text{(by (96))} \end{aligned}$$

$$
\begin{aligned}
&= P_{\bar{M}(\theta)}((\mathbf{1}[h(\bar{g}(\bar{O}(\theta),\bar{M}(\theta))) \in \hat{\Gamma}(g(\bar{O}(\theta),\mathbf{Y}))], M(\bar{g}(\bar{O}(\theta),\mathbf{Y}))) \in \mathcal{B}_0) \text{ (by formal invariance)} \\
&= P_{\bar{M}(\theta)}((\mathbf{1}[h(\bar{g}(\bar{O}(\theta),\bar{M}(\theta))) \in \hat{\Gamma}(g(\bar{O}(\theta),\mathbf{Y}))], M(\mathbf{Y})) \in \mathcal{B}_0) \text{ (by invariance of } M) \\
&= P_{\bar{M}(\theta)}((\mathbf{1}[\hat{g}(\bar{O}(\theta),h(\bar{M}(\theta))) \in \hat{\Gamma}(g(\bar{O}(\theta),\mathbf{Y}))], M(\mathbf{Y})) \in \mathcal{B}_0) \text{ (by (88))} \\
&= P_{\bar{M}(\theta)}((\mathbf{1}[\hat{g}(\bar{O}(\theta),h(\bar{M}(\theta))) \in \hat{g}(\bar{O}(\theta),\hat{\Gamma}(\mathbf{Y}))], M(\mathbf{Y})) \in \mathcal{B}_0) \text{ (by (91))} \\
&= P_{\bar{M}(\theta)}((\mathbf{1}[h(\bar{M}(\theta))) \in \hat{\Gamma}(\mathbf{Y})], M(\mathbf{Y})) \in \mathcal{B}_0) \text{ (by (89)). } \blacksquare
\end{aligned}
$$

Lemma 2 (a) shows that the distribution of the maximal invariant $M(\mathbf{Y})$ is fully characterized by the value of $\bar{M}(\theta)$. Thus, in deriving the optimal invariant test, one only needs to consider the set of distributions for the effective observation $M(\mathbf{Y})$ in the effective parameter space $\theta \in \bar{M}(\Theta) = \{\bar{M}(\theta), \theta \in \Theta\}$. This can be a much simpler problem.

Example PERS(g). We can choose $\bar{M}(\theta) = (1, c)$ and $\bar{O}(\theta) = \sigma$, so Lemma 2 (a) asserts that the distribution of $M(\mathbf{X}_T) = \mathbf{X}_T^s$ only depends c. Since all tests can be written as functions of $M(\mathbf{x}_T)$, the problem of deriving a scale invariant tests of $H_0 : c = c_0$ and unknown σ given observation $\mathbf{X}_T$ has been transformed to the problem of testing $H_0 : c = c_0$ given observation $\mathbf{X}_T^s$, whose distribution does depend on nuisance parameters. Since the density of $\mathbf{X}_T^s$ is given by (38), a straightforward application of the NP lemma shows that the best test of $H_0 : c = c_0$ against $H_1 : c = c_1$ in this latter problem rejects for large values of

$$\left|\frac{\mathbf{\Omega}_{XX}^{LTU}(c_1)}{\mathbf{\Omega}_{XX}^{LTU}(c_0)}\right|^{-1/2} \frac{(\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c_1)^{-1}\mathbf{X}_T^s)^{-q/2}}{(\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c_0)^{-1}\mathbf{X}_T^s)^{-q/2}}.$$

This is equivalent to rejecting for large values of

$$\frac{\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c_0)^{-1}\mathbf{X}_T^s}{\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c_1)^{-1}\mathbf{X}_T^s}$$

and with $c_0 = 0$ this yields the point-optimal scale invariant unit root test (58). ▲

Similarly, Lemma 2 (b) extends the conclusion of part (a) to the event that a given invariant set estimator $\hat{\Gamma}(\mathbf{Y})$ contains the true value. Thus, as in the testing problem, in deriving invariant set estimators, one can treat $M(\mathbf{Y})$ as the effective observation and $\bar{M}(\Theta)$ as the effective parameter space.

Example PERS(h). With $h(\bar{M}(\theta)) = h(\theta) = c$, part (b) of Lemma 2 asserts that the distribution of $(\mathbf{1}[c \in \hat{\Gamma}(\mathbf{X}_T)], M(\mathbf{X}_T)) = (\mathbf{1}[c \in \hat{\Gamma}(\mathbf{X}_T^s)], \mathbf{X}_T^s)$ only depends on c for any scale invariant confidence interval $\hat{\Gamma}$. ▲

Example MEAN(l). We can choose $\bar{M}(\theta) = (0, 1, c)$ and $\bar{O}(\theta) = (\mu, \sigma)$. With $h(\bar{M}(\theta)) = h((0, 1, c)) = 0$, Lemma 2 asserts that the distribution of $(\mathbf{1}[\mu \in \hat{\Gamma}(\mathbf{X}_T^0)], M(\mathbf{X}_T^0))$ under $\theta = (\mu, \sigma, c)$ is the same as the distribution of

$$(\mathbf{1}[0 \in \hat{\Gamma}(\mathbf{X}_T^0)], M(\mathbf{X}_T^0)))$$

under $\theta = (0, 1, c)$. Invariance has reduced the effective parameter space to the one-dimensional unknown $c \in (0, \infty)$. ▲

Example BRK(d). We can choose $\bar{M}(\theta) = (0, \delta/\sigma, 1, r)$ and $\bar{O}(\theta) = (\mu, \sigma)$. With $h(\bar{M}(\theta)) = r$, part (b) of Lemma 2 asserts that it is without loss of generality to consider the distribution of $(\mathbf{1}[r \in \hat{\Gamma}(\mathbf{X}_T^0)], M(\mathbf{X}_T^0))$ only for values $\theta = (0, \delta/\sigma, 1, r)$ (and indeed, the distribution of $M(\mathbf{X}_T^0) = (0, \mathbf{X}_T^{s\prime})'$ only depends on $(\delta/\sigma, r)$). ▲

7.3.3 Length Optimal Invariant Confidence Sets

The preceding section showed that in the construction of invariant confidence sets, the effective observation is $M(\mathbf{Y})$, and the effective parameter space is $\bar{M}(\Theta)$. We now discuss how to use these insights to construct length optimal invariant confidence sets (cf. Müller and Norets (2016)).

If $\hat{g}$ is the identity transformation, then from (94), $\hat{\Gamma}(\mathbf{y}) = \hat{\Gamma}(M(\mathbf{y}))$, and by Lemma 2 (b)

$$E_\theta \left[\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{Y})] d\gamma_0 \right] = E_{\bar{M}(\theta)} \left[\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(M(\mathbf{Y}))] d\gamma_0 \right]$$

so that the construction of length optimal *invariant* confidence sets in the problem of observing $\mathbf{Y}$ with parameter space Θ becomes identical to constructing length optimal confidence sets in the problem of observing $M(\mathbf{Y})$ with parameter space $\bar{M}(\Theta)$: Let $f_M(\mathbf{y}|\theta)$ be the density of $M(\mathbf{Y})$ under $\theta \in \bar{M}(\Theta)$ relative to some dominating measure ν_M (typically, $M(\mathbf{Y})$ is not a continuous random vector on $\mathcal{Y}$, even if $\mathbf{Y}$ is). For some given weighting function $\bar{F}$ on $\bar{M}(\Theta)$, the $\bar{F}$-weighted average expected length minimizing invariant level $1 - \alpha$ confidence set is then obtained by inverting a sequence of level α WAP maximizing tests of $H_0 : h(\theta) = \gamma_0$ against H_1 :the density of $M(\mathbf{Y})$ is $\int f_M(\mathbf{y}|\theta) d\bar{F}(\theta)$, based on the observation $M(\mathbf{Y})$. If h is one-to-one on the effective parameter space $\bar{M}(\Theta)$ with inverse function $h^{-1} : \Gamma \mapsto \bar{M}(\Theta)$, then assuming no randomization is necessary, these WAP maximizing

tests $\varphi^*_{\gamma_0} : M(\mathcal{Y}) \mapsto \{0,1\}$ are of the form

$$\varphi^*_{\gamma_0}(\mathbf{y}) = \mathbf{1}\left[\int f_M(\mathbf{y}|\theta)d\bar{F}(\theta) > cv_{\gamma_0} f_M(\mathbf{y}|h^{-1}(\gamma_0))\right], \tag{97}$$

as discussed at the end of Section 7.2.2. If the effective parameter space $\bar{M}(\Theta)$ contains nuisance parameters, then each WAP maximizing test, indexed by γ_0, is characterized by a least favorable distribution $\Lambda^*_{\gamma_0}$, and the test is of the form

$$\varphi^*_{\gamma_0}(\mathbf{y}) = \mathbf{1}\left[\int f_M(\mathbf{y}|\theta)d\bar{F}(\theta) > cv_{\gamma_0}\int f_M(\mathbf{y}|\theta)d\Lambda^*_{\gamma_0}(\theta)\right] \tag{98}$$

as discussed at the end of Section 7.2.3. In either case, once $\varphi^*_{\gamma_0}$ is determined on $M(\mathcal{Y})$, it is extended to the original domain $\mathcal{Y}$ via (90), that is, $\varphi^*_{\gamma_0}(\mathbf{y}) = \varphi^*_{\gamma_0}(M(\mathbf{y}))$ for $\mathbf{y} \in \mathcal{Y}$.

Example PERS(i). The confidence interval for c obtained by inverting the family of level α tests of $H_0 : c = c_0$ based on

$$\int |\mathbf{\Omega}^{LTU}_{XX}(c)|^{-1/2}\left(\frac{\mathbf{X}_T'\mathbf{\Omega}^{LTU}_{XX}(c)^{-1}\mathbf{X}_T}{\mathbf{X}_T'\mathbf{\Omega}^{LTU}_{XX}(c_0)^{-1}\mathbf{X}_T}\right)^{-q/2} d\bar{F}(c)$$

for some probability distribution $\bar{F}$ yields the minimal $\bar{F}$-weighted expected length level $1-\alpha$ confidence interval for c among all scale invariant intervals. ▲

Example BRK(e). For notational ease, define $\delta^s = \delta/\sigma$. Let $\bar{F}$ be a probability distribution for (δ^s, r) such that under $\bar{F}$, $r \sim \mathcal{U}[0,1]$ and $\delta^s \sim \mathcal{N}(0, T^{-1}\varpi^2)$. Consider testing $H_0 : r = r_0$ against $H_1 : (\delta^s, r) \sim \bar{F}$ based on observing $M(\mathbf{X}^0_T) = (0, \mathbf{X}^{s\prime}_T)'$. Proceeding as in Appendix B of Müller and Watson (2016), we find that the density of $\mathbf{X}^s_T$ (relative to the Haar density on the surface of the q-dimensional unit sphere) is equal to

$$\begin{aligned} f(\mathbf{x}^s_T|\delta^s, r) &= \int_0^\infty (2\pi)^{-q/2}u^{q-1}\exp\left[-\tfrac{1}{2}(u\mathbf{x}^s_T - \delta^s\mathbf{v}(r))'(u\mathbf{x}^s_T - \delta^s\mathbf{v}(r))\right]du \\ &= (2\pi)^{-(q-1)/2}\exp\left[\tfrac{1}{2}(\delta^s)^2[(\mathbf{v}(r)'\mathbf{x}^s_T)^2 - \mathbf{v}(r)'\mathbf{v}(r)]\right]\int_0^\infty \phi(u - \delta^s\mathbf{v}(r)'\mathbf{x}^s_T)u^{q-1}du \end{aligned} \tag{99}$$

where ϕ is the p.d.f. of a standard normal variate, and the remaining integral may be computed in closed form as in Appendix C.2 of Elliott et al. (2015). If the probability distribution $\Lambda^*_{r_0}$ for δ^s is least favorable at level α, the tests reject for large values of

$$\frac{\int f(\mathbf{X}^s_T|\delta^s, r)d\bar{F}(r, \delta^s)}{\int f(\mathbf{X}^s_T|\delta^s, r_0)d\Lambda^*_{r_0}(\delta^s)}.$$

Equivalently, integrating out $\delta^s \sim \mathcal{N}(0, T^{-1}\varpi^2)$ analytically by changing the order of integration in (99), the tests reject for large values of

$$\frac{\int |\mathbf{I}_q + \varpi^2 \mathbf{v}(r)\mathbf{v}(r)'|^{-1/2} [\mathbf{X}_T^{s\prime} \left(\mathbf{I}_q + \varpi^2 \mathbf{v}(r)\mathbf{v}(r)'\right)^{-1} \mathbf{X}_T^s]^{-q/2} dr}{\int f(\mathbf{X}_T^s | \delta^s, r_0) d\Lambda_{r_0}^*(\delta^s)} \tag{100}$$

with critical value equal to the $1-\alpha$ quantile of (100) with $\mathbf{X}_T^s \sim \mathbf{X}_T / \sqrt{\mathbf{X}_T' \mathbf{X}_T}$, $\mathbf{X}_T \sim \mathcal{N}(\delta^s \mathbf{v}(r_0), T^{-1}\mathbf{I}_q)$, and $\delta^s \sim \Lambda_{r_0}^*$. The inversion of these tests yields the confidence set with smallest expected length under $\delta^s \sim \mathcal{N}(0, T^{-1}\varpi^2)$ and $r \sim \mathcal{U}[0,1]$ among all invariant level $1-\alpha$ confidence sets for r. Also, with $\Lambda_{r_0}^*$ replaced by an arbitrary distributions for δ^s, the inversion of these tests yields a set for r whose expected length under $\delta^s \sim \mathcal{N}(0, T^{-1}\varpi^2)$ and $r \sim \mathcal{U}[0,1]$ provides a lower bound on the expected length of any valid level $1-\alpha$ invariant confidence set. ▲

The case where $\hat{g}$ is not the identity transformation is more subtle. For an invariant set $\hat{\Gamma}$, we have

$$\begin{aligned}
\mathbf{1}[h(\bar{M}(\theta))) \in \hat{\Gamma}(\mathbf{y})] &= \mathbf{1}[h(\bar{M}(\theta)) \in \hat{g}(O(\mathbf{y}), \hat{\Gamma}(M(\mathbf{y})))] \quad \text{(by (94))} \\
&= \mathbf{1}[\hat{g}(O(\mathbf{y}), \hat{g}(O(\mathbf{y})^{-1}, h(\bar{M}(\theta)))) \in \hat{g}(O(\mathbf{y}), \hat{\Gamma}(M(\mathbf{y})))] \\
&= \mathbf{1}[\hat{g}(O(\mathbf{y})^{-1}, h(\bar{M}(\theta))) \in \hat{\Gamma}(M(\mathbf{y}))] \quad \text{(by (89))}
\end{aligned} \tag{101}$$

so that from Lemma 2 (b), its coverage probability is given by

$$E_\theta[\mathbf{1}[h(\theta) \in \hat{\Gamma}(\mathbf{Y})]] = E_{\bar{M}(\theta)}\left[\mathbf{1}[\hat{g}(O(\mathbf{Y})^{-1}, h(\bar{M}(\theta))) \in \hat{\Gamma}(M(\mathbf{Y}))]\right]. \tag{102}$$

Furthermore, assume that there exists a function $g_l : A \mapsto \mathbb{R}$ such that for any Borel subset $\Gamma_0 \subset \Gamma$,

$$\int \mathbf{1}[\gamma_0 \in \hat{g}(a, \Gamma_0)] d\gamma_0 = g_l(a) \int \mathbf{1}[\gamma_0 \in \Gamma_0] d\gamma_0, \tag{103}$$

that is, g_l records how the length of $\Gamma_0 \subset \Gamma$ changes by the transformation $\hat{g}(a, \Gamma_0)$. Then, by invoking Lemma 2 (b), the expected length of $\hat{\Gamma}$ can be written as

$$\begin{aligned}
E_\theta\left[\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{Y})] d\gamma_0\right] &= E_{\bar{M}(\theta)}\left[g_l(O(\mathbf{Y})) \int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(M(\mathbf{Y}))] d\gamma_0\right] \\
&= E_{\bar{M}(\theta)}\left[e(M(\mathbf{Y})|\bar{M}(\theta)) \int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(M(\mathbf{Y}))] d\gamma_0\right]
\end{aligned} \tag{104}$$

with $e(M(\mathbf{Y})|\bar{M}(\theta)) = E_{\bar{M}(\theta)}[g_l(O(\mathbf{Y}))|M(\mathbf{Y})]$ by the law of iterated expectations.

Recalling that $f_M(\mathbf{y}|\theta)$ is the density of $M(\mathbf{Y})$ under $\theta \in \bar{M}(\Theta)$, we can rewrite (104) as

$$E_\theta\left[\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{Y})]d\gamma_0\right] = \int\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{y})]e(\mathbf{y}|\theta)f_M(\mathbf{y}|\theta)d\nu_M(\mathbf{y})d\gamma_0 \tag{105}$$

and with $f_O(\cdot|\mathbf{y},\theta)$ denoting the conditional p.d.f. of the scalar random variable $\hat{g}(O(\mathbf{Y})^{-1}, h(\theta))$ given $M(\mathbf{Y}) = \mathbf{y}$ under $\theta \in \bar{M}(\Theta)$, the coverage probability (102) becomes

$$E_\theta[\mathbf{1}[h(\theta) \in \hat{\Gamma}(\mathbf{Y})]] = \int\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{y})]f_O(\gamma_0|\mathbf{y},\theta)f_M(\mathbf{y}|\theta)d\nu_M(\mathbf{y})d\gamma_0. \tag{106}$$

If the effective parameter space after imposing invariance $\bar{M}(\Theta)$ is a singleton, then minimizing (105) subject to (106) is simply solved by the set

$$\hat{\Gamma}^*(\mathbf{y}) = \{\gamma_0 : e(\mathbf{y}|\theta)f_M(\mathbf{y}|\theta) \leq cv^* f_O(\gamma_0|\mathbf{y},\theta)f_M(\mathbf{y}|\theta)\} \tag{107}$$

where $cv^* \geq 0$ is such that (102) (or, equivalently, (106)) are equal to $1-\alpha$, at least as long as no randomization is necessary. Note that (107) amounts to inverting the tests $\varphi_{\gamma_0} : M(\mathbf{Y}) \mapsto \{0,1\}$

$$\varphi_{\gamma_0}(\mathbf{y}) = \mathbf{1}[e(\mathbf{y}|\theta)f_M(\mathbf{y}|\theta) > cv^* f_O(\gamma_0|\mathbf{y},\theta)f_M(\mathbf{y}|\theta)] \tag{108}$$

which have the same structure as NP tests, with $f_O(\gamma_0|\mathbf{y},\theta)f_M(\mathbf{y}|\theta)$ playing the role of the density under the null hypothesis, and $e(\mathbf{y}|\theta)f_M(\mathbf{y}|\theta)$ the density under the alternative. The analogy is not exact, however, since these are no longer necessarily p.d.f.s of $\mathbf{Y}$ (or $M(\mathbf{Y})$). Also, the critical value cv^* is not specific to γ_0, but rather ensures that $E_\theta[\varphi_{\hat{g}(O(\mathbf{Y})^{-1},h(\theta))}(M(\mathbf{Y}))] = \alpha$ (cf. (102)).

Example MEAN(m). Suppose c is known, so that the effective parameter space after imposing invariance is the singleton $\bar{M}(\theta) = (0,1,c)$. With $\hat{g}(a,\gamma) = a_\sigma\gamma + a_\mu$, (103) holds with $g_l(a) = a_\sigma$, and recalling that $O(\mathbf{X}_T^0) = (\bar{x}_{1:T}, \sqrt{\mathbf{X}_T'\mathbf{X}_T})$, $e(M(\mathbf{X}_T^0)|\bar{M}(\theta)) = E[\sqrt{\mathbf{X}_T'\mathbf{X}_T}|\mathbf{X}_T^s]$ and $f_O(\cdot|\mathbf{X}_T^s, c)$ is the conditional density of $\hat{g}(O(\mathbf{X}_T^0)^{-1}, h(\theta)) = -\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$ given $\mathbf{X}_T^s$. Calculations detailed in Appendix B of Müller and Watson (2016) show that $E[\sqrt{\mathbf{X}_T'\mathbf{X}_T}|\mathbf{X}_T^s]$ is proportional to $(\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{X}_T^s)^{-1/2}f_M(\mathbf{X}_T^s|c)$, and that the conditional distribution of

$$\frac{\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T} - m((0,\mathbf{X}_T^{s\prime})',c)}{s((0,\mathbf{X}_T^{s\prime})',c)} \tag{109}$$

given $\mathbf{X}_T^s$ is Student-t with q degrees of freedom, with $m(\mathbf{x}_T^0, c)$ and $s(\mathbf{x}_T^0, c)$ defined in (65) and (66). Thus, with $f_O(\cdot|\mathbf{X}_T^s, \theta)$ the implied scaled and shifted Student-t density of $-\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$ given $\mathbf{X}_T^s$, the length optimal confidence set is of the form

$$\hat{\Gamma}^*((0, \mathbf{X}_T^{s\prime})') = \left\{\gamma_0 : cv^* \sqrt{\mathbf{X}_T^{s\prime} \mathbf{\Omega}_{XX}^{LTU}(c)^{-1} \mathbf{X}_T^s} f_O(\gamma_0|\mathbf{X}_T^s, c) \geq 1\right\}$$

where cv^* is chosen such that $P(-\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T} \in \hat{\Gamma}^*(\mathbf{X}_T^s)) = 1 - \alpha$, and for generic $\mathbf{X}_T^0$, this yields the set

$$\hat{\Gamma}^*(\mathbf{X}_T^0) = \left\{\bar{x}_{1:T} + \sqrt{\mathbf{X}_T'\mathbf{X}_T}\gamma_0 : cv^* \sqrt{\mathbf{X}_T^{s\prime} \mathbf{\Omega}_{XX}^{LTU}(c)^{-1} \mathbf{X}_T^s} f_O(\gamma_0|\mathbf{X}_T^s, c) \geq 1\right\} \tag{110}$$

via (95). Since $\sqrt{\mathbf{X}_T^{s\prime} \mathbf{\Omega}_{XX}^{LTU}(c)^{-1} \mathbf{X}_T^s}$ is proportional to the scale of the conditional scaled and shifted Student-t distribution of $\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$, this is recognized as being equal to the interval $m(\mathbf{x}_T^0, c) \pm cv_{q+1}^T s(\mathbf{x}_T^0, c)$, where cv_q^T is the usual two-sided level $1 - \alpha$ critical value of a Student-t with q degrees of freedom, so it is simply the Bayesian credible set of Section 6.3. ▲

If $\bar{M}(\Theta)$ is not a singleton, then the changes to the analysis are analogous to the construction of confidence sets in the presence of nuisance parameters, as discussed in Sections 7.2.2 and 7.2.3. Specifically, with $\bar{F}$ a probability weighting function on $\bar{M}(\Theta)$, the $\bar{F}$-weighted average expected length of the invariant confidence set $\hat{\Gamma}$ is given by

$$\int E_{\bar{M}(\theta)}\left[e(M(\mathbf{Y})|\theta) \int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(M(\mathbf{Y}))] d\gamma_0\right] d\bar{F}(\theta) = \tag{111}$$
$$\int\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{y})] \left(\int_{\bar{M}(\Theta)} e(\mathbf{y}|\theta) f_M(\mathbf{y}|\theta) d\bar{F}(\theta)\right) d\nu_M(\mathbf{y}) d\gamma_0.$$

And given a probability distribution $\bar{\Lambda}$ of $\bar{M}(\Theta)$, the $\bar{\Lambda}$ weighted average coverage is equal to

$$\int E_\theta[\mathbf{1}[h(\theta) \in \hat{\Gamma}(\mathbf{Y})]] d\Lambda(\theta) = \int\int \mathbf{1}[y \in \hat{\Gamma}(\mathbf{y})] \left(\int_{\bar{M}(\Theta)} f_O(y|\mathbf{y}, \theta) f_M(\mathbf{y}|\theta) d\bar{\Lambda}(\theta)\right) d\nu_M(\mathbf{y}) dy. \tag{112}$$

Minimizing (111) among all functions $\hat{\Gamma} : M(\mathcal{Y}) \mapsto \mathcal{G}$ subject to (112) to be at least $1 - \alpha$ yields sets of the form

$$\hat{\Gamma}_{\bar{\Lambda}}(\mathbf{y}) = \left\{\gamma_0 : \int_{\bar{M}(\Theta)} e(\mathbf{y}|\theta) f_M(\mathbf{y}|\theta) d\bar{F}(\theta) \leq cv_{\bar{\Lambda}} \int_{\bar{M}(\Theta)} f_O(\gamma_0|\mathbf{y}, \theta) f_M(\mathbf{y}|\theta) d\bar{\Lambda}(\theta))\right\} \tag{113}$$

where $cv_{\bar{\Lambda}} \geq 0$ is such that (112) is equal to $1-\alpha$, at least as long as no randomization is necessary. This amounts to inverting tests of the form

$$\varphi_{\gamma_0}(\mathbf{y}) = \mathbf{1}[\int_{\bar{M}(\Theta)} e(\mathbf{y}|\theta) f_M(\mathbf{y}|\theta) d\bar{F}(\theta) > cv_{\bar{\Lambda}} \int_{\bar{M}(\Theta)} f_O(\gamma_0|\mathbf{y},\theta) f_M(\mathbf{y}|\theta) d\bar{\Lambda}(\theta))]. \tag{114}$$

By the analogous reasoning of Section 7.2.3, any $\hat{\Gamma}_{\bar{\Lambda}}$ in (113) provides a lower bound on the objective (111) among all level $1-\alpha$ confidence sets $\hat{\Gamma} : M(\mathcal{Y}) \mapsto \mathcal{G}$ of $\gamma = h(\theta)$ in the parameter space $\bar{M}(\Theta)$. And if $\bar{\Lambda} = \bar{\Lambda}^*$ is such that $\hat{\Gamma}_{\bar{\Lambda}^*}$ has coverage (102) that is at least as large as $1-\alpha$ for all $\theta \in \bar{M}(\theta)$, that is if

$$\sup_{\theta \in \bar{M}(\theta)} E_\theta[\varphi_{\hat{g}(O(\mathbf{Y})^{-1}, h(\theta))}(M(\mathbf{Y}))] = \alpha$$

then $\bar{\Lambda}^*$ is the least favorable distribution, and the resulting invariant confidence set (cf. (94))

$$\hat{g}(O(\mathbf{y}), \hat{\Gamma}_{\bar{\Lambda}^*}(M(\mathbf{y}))) \tag{115}$$

for $\mathbf{y} \in \mathcal{Y}$ minimizes $\bar{F}$-weighted average expected length among all level $1-\alpha$ invariant confidence sets.

This solution might seem complicated. But note that (115) is fully determined by the least favorable distribution $\bar{\Lambda}^*$ on $\bar{M}(\Theta)$ and associated critical value $cv_{\bar{\Lambda}^*}$. As such, the problem is about as hard to solve as determining a *single* hypothesis test with a composite null hypothesis against a composite alternative, as discussed in Section 7.2.3. In contrast, the case where $\hat{g}$ is the identity transformation and $\bar{M}(\Theta)$ still involves nuisance parameters requires solving a *family* of such problems, indexed by γ_0.

Example MEAN(n). With c unknown, the length-optimal tests (114) are of the form

$$\varphi_{\gamma_0}((0, \mathbf{X}_T^{s\prime})') = \mathbf{1}\left[\int (\mathbf{X}_T^{s\prime} \mathbf{\Omega}_{XX}^{LTU}(c)^{-1} \mathbf{X}_T^s)^{-1/2} f_M(\mathbf{X}_T^s|c) d\bar{F}(c) \geq cv_{\bar{\Lambda}} \int f_O(\gamma_0|\mathbf{X}_T^s, c) f_M(\mathbf{X}_T^s|c) d\bar{\Lambda}(c)\right] \tag{116}$$

where $cv_{\bar{\Lambda}}$ is such that $\int E_c[\mathbf{1}[\varphi_{-\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T}}((0, \mathbf{X}_T^{s\prime})')]] d\bar{\Lambda}(c) = \alpha$, and a level $1-\alpha$ optimal invariant confidence set requires determination of the least favorable distribution $\bar{\Lambda}^*$ such that for φ_{γ_0} as in (116) with $\bar{\Lambda} = \bar{\Lambda}^*$, $E_c[\mathbf{1}[\varphi_{-\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T}}((0, \mathbf{X}_T^{s\prime})')] \leq \alpha$ for all $c \geq 0$.

The weighting function $\bar{F}$ here trades off the expected length of the resulting interval for different values of c. For known c and $\sigma = 1$, the expected length of the optimal invariant

confidence interval (110) is proportional to

$$\sqrt{\mathbf{\Omega}_{\bar{x}\bar{x}}^{LTU}(c) - \mathbf{\Omega}_{\bar{x}X}^{LTU}(c)\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{\Omega}_{X\bar{x}}^{LTU}(c)} \tag{117}$$

(cf. (66)), which becomes very large for small c. Implicitly, a flat weighting function $\bar{F}$ on c thus puts a lot of emphasis on small values of c, since these values contribute the most to the weighted average expected length. To compensate for this mechanical effect, it makes sense to use instead a weighting function $\bar{F}$ whose density is proportional to some baseline choice divided by (117). In this manner, the $\bar{F}$-weighted expected length minimizing confidence set minimizes the baseline weighted average excess expected length ratios, a form of weighted average regret of not knowing c. ▲

7.4 Confidence Sets and Credible Sets

7.4.1 Coverage of Credible Sets

Both confidence sets and credible sets of some parameter of interest $\gamma = h(\theta) \in \Gamma$ are set estimators that map data to subsets of Γ. In practice, both are used to describe the uncertainty about the true value of γ. By definition, a level $1-\alpha$ confidence set $\hat{\Gamma}$ contains the true value with probability of at least $1-\alpha$

$$\inf_{\theta\in\Theta} P_\theta(h(\theta) \in \hat{\Gamma}(\mathbf{Y})) \geq 1-\alpha \tag{118}$$

while a level $1-\alpha$ credible set $\hat{\Gamma}_p$ relative to some prior $p(\theta)$ contains $1-\alpha$ posterior probability mass for all realizations $\mathbf{Y} = \mathbf{y}$

$$\int_\Theta \mathbf{1}[h(\theta) \in \hat{\Gamma}_p(\mathbf{y})]p(\theta|\mathbf{y})d\theta = 1-\alpha. \tag{119}$$

Credible sets are not confidence sets by construction, that is, they do not in general satisfy (118). Level $1-\alpha$ credible sets do, however, always have prior weighted coverage of $1-\alpha$, that is

$$\int_\Theta P_\theta(h(\theta) \in \hat{\Gamma}_p(\mathbf{Y}))p(\theta)d\theta = 1-\alpha \tag{120}$$

which follows from (61) and a change of the order of integration. In a loose sense, confidence sets are thus more "pessimistic", as the infimum in (118) ensures coverage in repeated samples uniformly in θ, while a credible set only has has frequentist coverage by construction with

θ drawn from the prior. The more pessimistic frequentist approach might be considered attractive when several decision makers need to be convinced that $\hat{\Gamma}$ is large enough relative to its level, at least before having seen the data, since (118) implies $\int P_\theta(h(\theta) \in \hat{\Gamma}(\mathbf{Y}))p(\theta)d\theta \geq 1-\alpha$ for all priors p.

7.4.2 Conditional Properties of Confidence Sets

At the same time, a credible set contains $1-\alpha$ posterior probability for *all* realizations of $\mathbf{y}$, so its description of level $1-\alpha$ uncertainty is always meaningful. In contrast, the confidence set property (118) does not rule out that for some realizations $\mathbf{y}$, $\hat{\Gamma}(\mathbf{y})$ is very short, or even empty. These cases are not entirely pathological: Müller and Norets (2016) document examples where length optimal confidence sets are empty with positive probability. Confidence sets are thus not constrained to provide a sensible description of uncertainty for all $\mathbf{y}$, leading to potentially unreasonably short "overoptimistic" descriptions of level $1-\alpha$ uncertainty about γ for some realizations $\mathbf{y}$.

Müller and Norets (2016) provide references and a detailed analysis of this issue. As a practical matter, their suggestion is to start with a level $1-\alpha$ credible set $\hat{\Gamma}_p$ relative to some reasonable prior with density $p(\theta)$, and to then enlarge it to induce the level $1-\alpha$ confidence set property (118). In this way, the description of uncertainty is guaranteed to have some attractive properties both before and after having observed the data.

The enlargement is usefully performed in a way that the weighted average expected length of the resulting confidence set is minimized, now among all level $1-\alpha$ confidence sets that contain the given credible set $\hat{\Gamma}_p(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$. This requires a minor modification of the corresponding discussions at the end of Sections 7.2.2 and 7.2.3 above: If h is one-to-one with inverse $h^{-1} : \Gamma \mapsto \Theta$, as in Section 7.2.2, F-weighted average expected length is now minimized by a confidence set that inverts a family of tests of the form

$$\tilde{\varphi}^*_{\theta_0}(\mathbf{y}) = \mathbf{1}\left[\int f(\mathbf{y}|\theta)dF(\theta) \geq cv_{\gamma_0} f(\mathbf{y}|h^{-1}(\gamma_0))\right]\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{y})] \tag{121}$$

and in the presence of nuisance parameters as in Section 7.2.3

$$\tilde{\varphi}_{\Lambda^*_{\gamma_0}}(\mathbf{y}) = \mathbf{1}\left[\int f(\mathbf{y}|\theta)dF(\theta) \geq cv_{\gamma_0}\int f(\mathbf{y}|\theta)d\Lambda^*_{\gamma_0}(\theta)\right]\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{y})] \tag{122}$$

(at least as long as no randomization is necessary). The $\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{y})]$ term ensures that the tests never reject when the credible set $\hat{\Gamma}_p(\mathbf{y})$ contains γ_0, so inversion of the tests yields

supersets of $\hat{\Gamma}_p(\mathbf{y})$. In general, the critical values cv_{γ_0} in (121) and (122) depend on the choice of prior and form of $\hat{\Gamma}_p(\mathbf{y})$ (equal-tailed or HPD, say), and if the test $\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{y})]$ is already a level α test of $H_0 : h(\theta) = \gamma_0$ for some $\gamma_0 \in \Gamma_0$, then the corresponding critical value cv_{γ_0} is equal to zero (since no enlargement is necessary to ensure level $1-\alpha$ coverage of this value of γ_0). The constraint of always including the given credible set thus does not make the problem of determining length optimal confidence sets any harder.

The weighting function F and the prior density p are in principle unrelated in this construction. But in practice, it often makes sense to use the same distribution, that is, to let F be the distribution with density p. The inversion of (121) or (122) then yields a confidence set $\hat{\Gamma}^*$ that minimizes p weighted average expected length $\int_\Theta E_\theta[\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}^*(\mathbf{y})]d\gamma_0]p(\theta)d\theta$ among all level $1-\alpha$ confidence sets that are supersets of $\hat{\Gamma}_p(\mathbf{y})$ for all $\mathbf{y}$.

7.4.3 Conditional Properties of Confidence Sets under Invariance

The enlargement approach of the last subsection readily extends to length optimal invariant confidence sets, as discussed in Section 7.3.3, as long as the credible set $\hat{\Gamma}_p$ shares the same invariance property. To be precise, if $\hat{g}$ is the identity transformation, then the $\bar{F}$-weighted average expected length minimizing invariant confidence set is obtained by simply inverting tests of the form (97) and (98) multiplied by $\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{y})]$ with appropriately adjusted critical value, just as in (121) and (122). If $\hat{g}$ isn't the identity transformation, length minimizing sets are constructed via (115) from inverting tests of the form (108) or (114) multiplied by $\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{y})]$.

Example MEAN(o). Let $\hat{\Gamma}_p(\mathbf{x}_T^0)$ be the level $1-\alpha$ equal-tailed credible set for μ derived from the mixture of shifted and scaled Student-t distributions obtained from the improper priors discussed in Section 6.3. Note that $\hat{\Gamma}_p(\mathbf{x}_T^0)$ is invariant, so $\hat{\Gamma}_p(\mathbf{x}_T^0) = \{\overline{x}_{1:T} + \sqrt{\mathbf{x}_T'\mathbf{x}_T}\gamma_0 : \gamma_0 \in \hat{\Gamma}((0, \mathbf{x}_T^{s\prime})')\}$. Thus, the length-optimally enlarged level $1-\alpha$ confidence set for inverts tests of the form

$$\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{X}_T^0)]\mathbf{1}[\int (\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{X}_T^s)^{-1/2} f_M(\mathbf{X}_T^s|c)d\bar{F}(c) > cv_{\bar{\Lambda}^*} \int f_O(\gamma_0|\mathbf{X}_T^s, c) f_M(\mathbf{X}_T^s|c)d\bar{\Lambda}^*(c))]$$

where $cv_{\bar{\Lambda}^*}$ is now such that this test is of level α with c drawn randomly from $\bar{\Lambda}^*$, and the least favorable distribution $\bar{\Lambda}^*$ is such that this test is level α for all $c > 0$. ▲

This approach requires specifying a prior on the original parameter space Θ that induces $\hat{\Gamma}_p$ to be invariant, such as the flat prior on a location parameter (cf. the discussion in Section 6.3). Alternatively, one might take a limited information approach to the Bayes problem: Recall from Section 7.3.2 that with invariant confidence sets, the effective observation is $M(\mathbf{Y})$, and the effective parameter space is $\bar{M}(\Theta)$. Under a prior $\bar{p}$ on $\bar{M}(\Theta)$, consider the Bayesian problem of observing $M(\mathbf{Y}) = \mathbf{y}$ generated from some $\theta \in \bar{M}(\Theta)$, and having to form the posterior probability that $h(\bar{M}(\theta))$ falls into some set $\hat{\Gamma}(\mathbf{y}) \subset \Gamma$. If the parameter of interest does not vary with the transformations, that is, if $\hat{g}$ is the identity transformation, then this is simply equal to

$$\int_{\bar{M}(\Theta)} \mathbf{1}[h(\theta) \in \hat{\Gamma}(\mathbf{y})]\bar{p}(\theta|\mathbf{y})d\theta. \tag{123}$$

If $\hat{g}$ is not the identity transformation, then from (101), this is the same as the posterior probability of the event $\hat{g}(O(\mathbf{Y})^{-1}, h(\bar{M}(\theta))) \in \hat{\Gamma}(\mathbf{y})$ conditional on observing $M(\mathbf{Y}) = \mathbf{y}$. In the notation of Section 7.3.3, this probability is

$$\int_{\bar{M}(\Theta)} \left(\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{y})] f_O(\gamma_0|\mathbf{y}, \theta) d\gamma_0 \right) \bar{p}(\theta|\mathbf{y})d\theta. \tag{124}$$

From this expression, a Bayesian limited to observing $M(\mathbf{Y}) \in M(\mathcal{Y})$ (rather than the original $\mathbf{Y} \in \mathcal{Y}$) could determine, for each realization $M(\mathbf{Y}) = \mathbf{y}$, the shortest set $\hat{\Gamma}(\mathbf{y})$ so that (124) is equal to $1 - \alpha$, or the equal-tailed interval with the property that the posterior probability of $\hat{g}(O(\mathbf{Y})^{-1}, h(\bar{M}(\theta)))$ being above or below the upper and lower end points of $\hat{\Gamma}(\mathbf{y})$ is equal to $\alpha/2$, respectively. Either way, the resulting set $\hat{\Gamma}_p(\mathbf{y})$ is a credible set for $\hat{g}(O(\mathbf{Y})^{-1}, h(\bar{M}(\theta)))$ in this limited information problem. Furthermore, extending this set to all $\mathbf{y} \in \mathcal{Y}$ via $\hat{\Gamma}(\mathbf{y}) = \hat{g}(O(\mathbf{y}), \hat{\Gamma}_p(\mathbf{y}))$ yields, again applying (101), an invariant set with the same limited information interpretation. Note that in this construction, one only has to form a prior $\bar{p}$ on the smaller effective parameter space $\bar{M}(\Theta)$, and the invariance of the limited information credible set is obtained by construction.

Example MEAN(p). With $\bar{M}(\Theta) = (0, 1, c)$, we only specify a prior p_c on c. Conditional on c, (109) implies that the posterior distribution of $\hat{g}(O(\mathbf{Y})^{-1}, h(\bar{M}(\theta))) = -\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$ given $M(\mathbf{X}_T^0) = (0, \mathbf{X}_T^{s\prime})' = (0, \mathbf{x}_T^{s\prime})'$ is Student-t scaled by $s((0, \mathbf{x}_T^{s\prime})', c)$ and shifted by $-m((0, \mathbf{x}_T^{s\prime})', c)$. From the form of the density of $\mathbf{X}_T^s$, (38), the posterior distribution for c is proportional to $p_c(\cdot)|\mathbf{\Omega}_{XX}^{LTU}(\cdot)|^{-1/2}(\mathbf{x}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(\cdot)^{-1}\mathbf{x}_T^s)^{-q/2}$, so that the limited information

posterior for the value of $\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$ is the corresponding mixture of the scaled and shifted Student-t distributions. Extending this set to all $\mathbf{X}_T^0$ yields the mixture of Student-t distributions scaled by $s(\mathbf{x}_T^0, c)$ and shifted by $m(\mathbf{x}_T^0, c)$, which is the same as the posterior set derived in Section 6.3 under uninformative priors on (μ, σ). ▲

Example BRK(f). Consider a limited Bayesian analysis for the break data using the prior $\bar{F}$ on (δ^s, r). Since $\hat{g}$ is the identity transformation in this example, the posterior probability (123) can simply be computed from the posterior from (δ^s, r). Integrating out δ^s over the prior $\delta^s \sim \mathcal{N}(0, T^{-1}\varpi^2)$ yields the posterior density for r given the "limited information" observation $M(\mathbf{X}_T^0) = (0, \mathbf{X}_T^{s\prime})'$ is

$$p(r|\mathbf{X}_T^s) \propto |\mathbf{I}_q + \varpi^2 \mathbf{v}(r)\mathbf{v}(r)'|^{-1/2}[\mathbf{X}_T^{s\prime}\left(\mathbf{I}_q + \varpi^2 \mathbf{v}(r)\mathbf{v}(r)'\right)^{-1}\mathbf{X}_T^s]^{-q/2}$$

with constant of proportionality determined by $\int_0^1 p(r|\mathbf{X}_T^s)dr = 1$, so that the equal-tailed level $1-\alpha$ credible set is equal to $\hat{\Gamma}(\mathbf{X}_T^0) = \hat{\Gamma}((0, \mathbf{X}_T^{s\prime})') = [L_{ET}(\mathbf{X}_T^s), U_{ET}(\mathbf{X}_T^s)]$ with $\int_0^{L_{ET}(\mathbf{X}_T^s)} p(r|\mathbf{X}_T^s)dr = \int_{U_{ET}(\mathbf{X}_T^s)}^1 p(r|\mathbf{X}_T^s)dr = \alpha/2$. ▲

7.5 Numerical Determination of Powerful Tests

As discussed above, when nuisance parameters are present under the null, powerful tests can be computed using a least favorable distribution. In some simple testing problems it is possible to "guess and verify" the least favorable distribution, but many inference problems are too complex for this strategy. This section discusses numerical methods for constructing powerful tests using approximate least favorable distributions. The chapter begins with a brief discussion of *importance sampling*, a well-known simulation for estimating the expected value of a random variable (see, for instance, Chapter 4.2.2 of Geweke (2005) for a more detailed discussion). In our context, importance sampling is used to estimate the rejection frequency of tests. We then move on to discuss the numerical methods that we have found useful for approximating least favorable distributions to obtain powerful tests and short confidence intervals.

7.5.1 Importance Sampling

A first-order problem in the numerical determination of tests involves computing a test's rejection frequency. And, inverting tests to determine confidence intervals requires these

rejection frequencies for many values of the model's parameters. For instance, if we want to use a test statistic to construct a level $1-\alpha$ confidence set, one must determine its $1-\alpha$ quantile under all $\theta \in \Theta$. With an appropriate definition of the function $\psi : \mathcal{Y} \mapsto \mathbb{R}$, this amounts to evaluating $E_\theta[\psi(\mathbf{Y})]$ as a function of $\theta \in \Theta$.

One approach is to discretize $\Theta = \{\theta_i\}_{i=1}^m$, and to draw, say, 10,000 (i.e., 10k) *i.i.d.* draws of $\mathbf{Y}$ with density $f(\cdot|\,\theta_i)$, for each $i = 1, \ldots, m$, and to compute $\psi(\mathbf{Y})$ for all of these $m \times 10$k draws. But with m moderately large, this quickly becomes computationally burdensome. It is also inefficient, since the distribution of $\psi(\mathbf{Y})$ under θ_1 is close the distribution of $\psi(\mathbf{Y})$ under θ_2 as long as the distribution of $\mathbf{Y}$ under θ_1 and θ_2 are close to each other. One can therefore (also) use the draws generated under θ_1 to learn about the distribution of $\psi(\mathbf{Y})$ under θ_2.

This idea is formalized by importance sampling. Let f_p be a probability density function for $\mathbf{Y}$ such that $f(\mathbf{y}|\theta)/f_p(\mathbf{y}) < \infty$ for all $\mathbf{y} \in \mathbf{Y}$, $\theta \in \Theta$, that is, the support of the density f_p is at least as large as the support of $f(\mathbf{y}|\theta)$, for all θ. Then

$$\begin{aligned} E_\theta[\psi(\mathbf{Y})] &= \int \psi(\mathbf{y}) f(\mathbf{y}|\theta) d\mathbf{y} \\ &= \int \psi(\mathbf{y}) \frac{f(\mathbf{y}|\theta)}{f_p(\mathbf{y})} f_p(\mathbf{y}) d\mathbf{y} \\ &= E_p\left[\frac{f(\mathbf{Y}|\theta)}{f_p(\mathbf{Y})}\psi(\mathbf{Y})\right] \end{aligned}$$

where we write E_p for integration over f_p. Thus, by the law of large numbers, we can approximate

$$E_\theta[\psi(\mathbf{Y})] \approx N^{-1} \sum_{l=1}^{N} \frac{f(\mathbf{Y}^{(l)}|\theta)}{f_p(\mathbf{Y}^{(l)})} \psi(\mathbf{Y}^{(l)}) \tag{125}$$

for N large, where $\mathbf{Y}^{(l)}$ are *i.i.d.* draws from the *proposal* density $f_p(\cdot)$. Values of $\mathbf{y}$ where $f_p(\mathbf{y}) > f(\mathbf{y}|\theta)$ are oversampled relative to the target distribution, and the *importance sampling weights* $f(\mathbf{Y}^{(l)}|\theta)/f_p(\mathbf{Y}^{(l)})$ in (125) downweigh these draws of $\mathbf{Y}$ in the calculation of the average. Note that one can use the same set of N draws $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(N)}$ to obtain an approximation for $E_\theta[\psi(\mathbf{Y})]$, for all $\theta \in \Theta$. In practice, however, this only works well if the distribution of the importance sample weights are not too skewed for all θ. This in turn requires that the proposal density is never much smaller than the target densities $f(\mathbf{y}|\theta)$, for all θ. For low-dimensional θ, this can often be achieved by discretizing $\Theta = \{\theta_i\}_{i=1}^m$, and

letting

$$f_p(\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{y}|\theta_i). \tag{126}$$

The density (126) is the equal probability mixture of $f(\mathbf{y}|\theta_i)$, $i = 1, \ldots, m$, so that a draw with density f_p is obtained by letting the random index $J \in \{1, \ldots, m\}$ be uniformly distributed, and by then drawing $\mathbf{Y}$ under θ_J.[21]

Even though the evaluation of $f_p(\mathbf{Y}^{(l)})$ under (126) requires m density evaluations, importance sampling is still often computationally more efficient than a direct discretization of Θ with many independent draws for each value: The approximation (125) requires many fewer total random draws (say, $N = 100$k in contrast to $m \times 10$k draws), and the discretization in the construction of the proposal density (126) can be chosen much coarser (say, $m = 50$) while still allowing for accurate approximations of $E_\theta[\psi(\mathbf{Y})]$ for all $\theta \in \Theta$.

The quality of the approximation (125) is usefully assessed by computing its standard error, which is given by

$$N^{-1/2}\sqrt{N^{-1}\sum_{l=1}^{N}\left(\frac{f(\mathbf{Y}^{(l)}|\theta)}{f_p(\mathbf{Y}^{(l)})}\psi(\mathbf{Y}^{(l)})\right)^2 - \left(N^{-1}\sum_{l=1}^{N}\frac{f(\mathbf{Y}^{(l)}|\theta)}{f_p(\mathbf{Y}^{(l)})}\psi(\mathbf{Y}^{(l)})\right)^2}. \tag{127}$$

To catch potential errors in the coding of importance sampling, it is advisable to check that $N^{-1}\sum_{l=1}^{N} f(\mathbf{Y}^{(l)}|\theta)/f_p(\mathbf{Y}^{(l)}) \approx 1$ for all $\theta \in \Theta$ with an approximation accuracy as suggested by this standard error.

Example PERS(j). Consider the problem of numerically obtaining the critical values for the length optimal confidence set with weighting function $\bar{F}$. This amount to having to compute the $1-\alpha$ quantile of

$$S(\mathbf{X}_T^s, c_0) = \int |\mathbf{\Omega}_{XX}^{LTU}(c)|^{-1/2}\left(\frac{\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c)^{-1}\mathbf{X}_T^s}{\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c_0)^{-1}\mathbf{X}_T^s}\right)^{-q/2} d\bar{F}(c) \tag{128}$$

under $\mathbf{X}_T^s = \mathbf{X}/\sqrt{\mathbf{X}'\mathbf{X}}$ with $\mathbf{X} \sim \mathcal{N}(0, \mathbf{\Omega}_{XX}^{LTU}(c_0))$, for all $c_0 \geq 0$. It turns out that values of $c_0 \geq 10^4$ lead to essentially the same $1-\alpha$ quantile of (128), so effectively the parameter space is compact. Here is an importance sampling scheme: generate $N = 100$k draws of

[21] When $\Theta \subset \mathbb{R}$ is a closed interval, it can make sense to include the endpoints repeatedly in the set $\{\theta_i\}_{i=1}^m$ to compensate for the lack of values of θ smaller or larger than the lower bound or upper bound, respectively.

$\mathbf{X}_T^{s,(l)}$, where in each draw, $\mathbf{X} \sim \mathcal{N}(0, \mathbf{\Omega}_{XX}^{LTU}(c))$ with c randomly drawn from $\{c_i\}_{i=1}^{50} = \{10^{-3+7i/50}\}_{i=1}^{50}$. The level $1-\alpha$ critical value cv_{c_0} then solves

$$\sum_{l=1}^{N} \frac{|\mathbf{\Omega}_{XX}^{LTU}(c_0)|^{-1/2}(\mathbf{X}_T^{s,(l)\prime}\mathbf{\Omega}_{XX}^{LTU}(c_0)^{-1}\mathbf{X}_T^{s,(l)})^{-q/2}}{f_p(\mathbf{X}_T^{s,(l)})}\mathbf{1}[S(\mathbf{X}^{s,(l)}, c_0) \leq cv_{c_0}] \approx 1-\alpha \tag{129}$$

where $f_p(\mathbf{X}_T^s) = \sum_{i=1}^{m} |\mathbf{\Omega}_{XX}^{LTU}(c_i)|^{-1/2}(\mathbf{X}_T^{s\prime}\mathbf{\Omega}_{XX}^{LTU}(c_i)^{-1}\mathbf{X}_T^s)^{-q/2}$ (and we omitted constants in the definition of $f_p(\mathbf{X}_T^s)$ that cancel in the ratio (129)). The equation can be solved by a simple bisection algorithm, exploiting the monotonicity in cv_{c_0}. ▲

7.5.2 Approximating Least Favorable Distributions

We now take up the problem of approximating least favorable distributions. This subsection and the next present methods discussed in Elliott et al. (2015), and refinements to those methods that we have subsequently found useful in a variety of contexts. Related numerical methods are discussed in Kempthorne (1987) and Moreira and Moreira (2013).

Recall from Section 7.2.3 that the solution to the hypothesis testing problem

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \text{the density of } \mathbf{Y} \text{ is } f_1. \tag{130}$$

involves the least favorable distribution. We now discuss a numerical approaches to determining a level α test of (130) that comes demonstrably close to maximizing power.

In this section, we assume that the parameter space Θ_0 is sufficiently small such that an approach based on a given discrete approximation $\Theta_0 \approx \{\theta_i\}_{i=1}^m$ is fruitful.[22] More specifically, this requires that the null rejection probability function $E_\theta[\varphi(\mathbf{Y})]$ on $\theta \in \Theta_0$ is reasonably well characterized by its values on $\{\theta_i\}_{i=1}^m$. In practice, this potentially holds if Θ_0 has one or two free parameters, but not if it has five. The next section discusses strategies for larger Θ_0.

Probability distributions Λ on $\{\theta_i\}_{i=1}^m$ are simply points in the m dimensional simplex. The defining property of the least favorable distribution Λ^* is that the level α test φ_Λ of

$$H_\Lambda : \text{the density of } \mathbf{Y} \text{ is } f_\Lambda(\mathbf{y}) = \int f(\mathbf{y}|\theta) d\Lambda(\theta) \tag{131}$$

[22]The number of points m in this section does not need to equate the number of points m used in importance sampling approximations of Section 7.5.1.

for $\Lambda = \Lambda^*$ against H_1 is also of level α under $\{\theta_i\}_{i=1}^m$, that is, $E_{\theta_i}[\varphi_{\Lambda^*}(\mathbf{Y})] \leq \alpha$ for all $i = 1, \ldots, m$. Assuming no randomization is necessary, NP tests of H_Λ against H_1 are of the form

$$\mathbf{1}[f_1(\mathbf{y}) > \sum_{i=1}^m \nu_i f(\mathbf{y}|\theta_i)] \tag{132}$$

where the non-negative weights ν_i are the product of the critical value and the probability mass of Λ on θ_i. Let $\{\nu_i\}_{i=1}^m \in [0, \infty)^m$ be a guess for the values of ν_i that characterize the test φ_{Λ^*}. If we find that the test (132) has null rejection probability larger (smaller) than α for some θ_i, then presumably a better guess is obtained by slightly increasing (decreasing) the corresponding values of ν_i. This suggests Algorithm 4 (cf. Section 3 of Elliott et al. (2015)).

The factors $\omega_i^{(j)}$ control the speed at which the weights $\nu_i^{(j)}$ are adjusted as a function of the discrepancy between the (estimated) rejection probability $\widehat{RP}_i^{(j+1)}$ and the nominal level α. These factors are slowly increased if the adjustment is in the same direction iteration after iteration, but quickly decreased otherwise. This algorithm is computationally fast, even for large N, since after the pre-computations in Step 2, it only involves additions and multiplications of $N \times m$ precomputed values.

Example BRK(g). As noted in Example BRK(e), the minimal expected length invariant confidence set for r under $\delta^s \sim \mathcal{N}(0, T^{-1}\varpi^2)$ inverts tests of $H_0 : r = r_0$ which involve a least favorable distribution $\bar{\Lambda}_{r_0}^*$ for δ^s. Set $\varpi = 10$, and suppose we want to control size on the grid of m values $\delta^s \in \{0, \pm 0.05, \ldots, \pm 20\} = \{\delta_i^s\}_{i=1}^m$ (these are the choices of Elliott et al. (2015) in the "all frequencies" version of the problem). For simplicity, use the same grid also for the proposal density, so that under f_p, $\mathbf{X}_T^s \sim \mathbf{X}_T / \sqrt{\mathbf{X}_T' \mathbf{X}_T}$ with an equal probability mixture of $\mathbf{X}_T \sim \mathcal{N}(\delta_i^s \mathbf{v}(r_0), T^{-1}\mathbf{I}_q)$, $i = 1, \ldots, m$. Let $\mathbf{X}_T^{s,(l)}$ be the corresponding $N = 100$k draws, say. Then Step 2 amounts to computing (dropping all constants that cancel in any ratio of the densities)

$$\begin{aligned} f_p(\mathbf{X}_T^s) &= m^{-1} \sum_{i=1}^m [(\mathbf{X}_T^s - \delta_i^s \mathbf{v}(r_0))'(\mathbf{X}_T^s - \delta_i^s \mathbf{v}(r_0))]^{-q/2} \\ f_1(\mathbf{X}_T^s) &= |\mathbf{I}_q + \varpi^2 \mathbf{v}(r)\mathbf{v}(r)'|^{-1/2} [\mathbf{X}_T^{s\prime} \left(\mathbf{I}_q + \varpi^2 \mathbf{v}(r)\mathbf{v}(r)'\right)^{-1} \mathbf{X}_T^s]^{-q/2} \end{aligned}$$

and $f(\mathbf{X}_T^s | \delta_i^s, r_0)$ as defined in (99), $i = 1, \ldots, m$ for all $\mathbf{X}_T^s = \mathbf{X}_T^{s,(l)}$, and the test (133) in Step 3 equals $\varphi^{(l),(j)} = \mathbf{1}[f_1(\mathbf{X}_T^{s,(l)}) > \sum_{i=1}^m \nu_i^{(j)} f(\mathbf{X}_T^{s,(l)} | \delta_i^s)]$. It would make sense to exploit

Algorithm 4 Approximation of the level α least favorable distribution for $\Theta_0 = \{\theta_i\}_{i=1}^m$ in (130)

1. Draw N *i.i.d.* draws $\mathbf{Y}^{(l)}$ from the proposal density f_p, $l = 1, \ldots, N$. (Note: (126) provides a candidate proposal density.)

2. Compute and store the values of $f_p(\mathbf{Y}^{(l)})$, $f_1(\mathbf{Y}^{(l)})$ and $f(\mathbf{Y}^{(l)}|\theta_i)$ for $i = 1, \ldots, m$, $l = 1, \ldots, N$.

3. Set $\nu_i^{(0)} = 1$, $\widehat{RP}_i^{(0)} = \alpha$, and $\omega_i^{(0)} = 1$ for $i = 1, \ldots, m$. (Note: these serve as initial guesses of ν_i, the rejection probability and a step-size parameter.)

4. Iterating over $j = 0, 1, \ldots, 399$, set

$$\varphi^{(l),(j)} = \mathbf{1}[f_1(\mathbf{Y}^{(l)}) > \sum_{i=1}^m \nu_i^{(j)} f(\mathbf{Y}^{(l)}|\theta_i)], \quad l = 1, \ldots, N \tag{133}$$

and

$$\begin{aligned} \widehat{RP}_i^{(j+1)} &= N^{-1} \sum_{l=1}^N \frac{f(\mathbf{Y}^{(l)}|\theta_i)}{f_p(\mathbf{Y}^{(l)})} \varphi^{(l),(j)} \\ \nu_i^{(j+1)} &= \nu_i^{(j)} \exp\left[\omega_i^{(j)} (\widehat{RP}_i^{(j+1)} - \alpha)\right], \\ \omega_i^{(j+1)} &= \begin{cases} \min(1.03\omega_i^{(j)}, 20) \text{ if } (\widehat{RP}_i^{(j+1)} - \alpha)(\widehat{RP}_i^{(j)} - \alpha) > 0 \\ \max(0.5\omega_i^{(j)}, 0.01) \text{ otherwise} \end{cases} \end{aligned} \tag{134}$$

for $i = 1, \ldots, m$.

5. Let $\hat{\nu}_i = \nu_i^{(400)}$, $\hat{\nu}^S = \sum_{i=1}^m \hat{\nu}_i$ and $\hat{\varphi}(\mathbf{y}) = \mathbf{1}[f_1(\mathbf{y}) > \sum_{i=1}^m \hat{\nu}_i f(\mathbf{y}|\theta_i)]$. The approximation of the LFD Λ^* has probability mass $\hat{\lambda}_i = \hat{\nu}_i / \hat{\nu}^S$ on θ_i, $i = 1, \ldots, m$, and the associated test $\hat{\varphi}(\mathbf{y})$ has critical value equal to $\hat{\nu}^S$. Its rejection probability under θ_i is estimated to equal $\widehat{RP}_i^{(400)}$ with associated standard error (127) evaluated at $\theta = \theta_i$ with $\psi(\mathbf{Y}^{(l)}) = \hat{\varphi}^{(l)} = \varphi^{(l),(400)}$.

the symmetry of the problem in the sign of δ^s to correspondingly impose the same symmetry in the weights $\nu_i^{(j)}$. ▲

Algorithm 4 seeks to identify the least favorable distribution on the discretized parameters space $\{\theta_i\}_{i=1}^m$. Given the numerical approximations involved in estimating the null rejection probabilities, and the finite number of iterations, it will not deliver the exact least favorable distribution. What is more, most nuisance parameter spaces are not discrete.

But recall from Theorem 1 (a) of Section 7.2.3 that the power of the level α test φ_Λ of H_Λ in (131) against H_1 provides an upper bound on the power of level α test of H_0, for *all* Λ. Thus, the distribution on $\{\theta_i\}_{i=1}^m$ obtained in Step 5 of Algorithm 4 can be applied to obtain such an upper bound. What is more, since the largest rejection probability of $\varphi^{(400)}$ for $\theta \in \{\theta_i\}_{i=1}^m$ is close to α, it is reasonable to expect that its largest rejection probability for $\theta \in \Theta_0$ is not much larger, at least as long as $\{\theta_i\}_{i=1}^m$ is a sufficiently fine discretization of Θ. Thus, a small increase of its critical value might be enough to obtain a level α under Θ_0 with power that is only slightly smaller than the power bound.

These considerations suggest Algorithm 5 (cf. Appendix A.3 of Müller and Watson (2018)).[23]

If Algorithm 5 is employed in the context of obtaining a level $1-\alpha$ confidence set by inverting $\hat{\varphi} = \hat{\varphi}_{\gamma_0}$, as discussed at the end of Section 7.2.3, then equation (85) shows how to use the power $\hat{\pi} = \hat{\pi}_{\gamma_0}$ of Step 7 to compute its F-weighted average expected length. What is more, as discussed there, one can also use the power bound $\bar{\pi} = \bar{\pi}_{\gamma_0}$ of Step 7 to obtain a lower bound on this length criterion that holds for all level $1-\alpha$ confidence sets.

Example BRK(h). Running the algorithm repeatedly for some grid of values $r_0 \in \{r_{0,i}\}_{i=1}^{m_r}$ yields a set of level α_0 tests $\hat{\varphi}_{r_{0,i}}$, $\bar{F}$-weighted power $\hat{\pi}_{r_{0,i}}$, and power upper bounds $\bar{\pi}_{r_{0,i}}$, $i = 1, \ldots, m_r$. The $\bar{F}$-weighted expected length of the feasible level $1-\alpha_0$ invariant confidence interval obtained by inverting the tests $\hat{\varphi}_{r_{0,i}}$ is $m_r^{-1} \sum_{i=1}^n (1 - \hat{\pi}_{r_{0,i}})$, and the corresponding lower bound on the $\bar{F}$-weighted expected length of any level $1-\alpha_0$ invariant confidence interval is $m_r^{-1} \sum_{i=1}^n (1 - \bar{\pi}_{r_{0,i}})$. ▲

When the confidence set is constrained to be a superset of a given level $1-\alpha_0$ credible set $\hat{\Gamma}_p$, as discussed in Section 7.4, the algorithm only requires minor modifications:

[23] See Elliott et al. (2015) for an algorithm that more directly targets a given tolerance between the power upper bound and the power of level α test.

Algorithm 5 Determination of a nearly power maximizing level α_0 test of (130)

1. Draw N *i.i.d.* draws $\mathbf{Y}^{(l)}$ from density f_p, $l = 1, \ldots, N$.

2. Form a finite approximation $\{\theta_i\}_{i=1}^m$ to Θ_0.

3. Apply Algorithm 4, with $\alpha = \alpha_0 - \varepsilon$ for some small $\varepsilon > 0$ (say, 0.3% for $\alpha_0 = 5\%$).

4. Compute $\sup_{\theta\in\Theta_0} \widehat{RP}(\theta)$ with

$$\widehat{RP}(\theta) = N^{-1} \sum_{l=1}^{N} \frac{f(\mathbf{Y}^{(l)}|\theta)}{f_p(\mathbf{Y}^{(l)})} \hat{\varphi}^{(l)} \tag{135}$$

   by using a fine discretization of Θ_0.

5. If $\sup_{\theta\in\Theta_0} \widehat{RP}(\theta) \geq \alpha_0$ increase ε and go back to Step 3, or choose a finer discretization and go back to Step 2.

6. Use bisection to determine the value of $\hat{c} < 1$ such that the test $\mathbf{1}[f_1(\mathbf{y}) > \hat{c} \sum_{i=1}^m \hat{\nu}_i f(\mathbf{y}|\theta_i)]$ is of level α_0 under the density $\sum_{i=1}^m \hat{\lambda}_i^* f(\theta_i|\mathbf{y})$, that is, $\hat{c}$ solves

$$N^{-1} \sum_{l=1}^{N} \left(\sum_{i=1}^{m} \hat{\lambda}_i^* \frac{f(\mathbf{Y}^{(l)}|\theta_i)}{f_p(\mathbf{Y}^{(l)})} \right) \mathbf{1}[f_1(\mathbf{Y}^{(l)}) > \hat{c} \sum_{i=1}^{m} \hat{\nu}_i f(\mathbf{Y}^{(l)}|\theta_i)] \approx \alpha_0. \tag{136}$$

7. Estimate the power $\hat{\pi}$ of $\hat{\varphi}(\mathbf{y}) = \mathbf{1}[f_1(\mathbf{y}) > \sum_{i=1}^m \hat{\nu}_i f(\mathbf{y}|\theta_i)]$ and the power $\bar{\pi}$ of $\bar{\varphi}(\mathbf{y}) = \mathbf{1}[f_1(\mathbf{y}) > \hat{c} \sum_{i=1}^m \hat{\nu}_i f(\mathbf{y}|\theta_i)]$ by evaluating the tests on N_1 *i.i.d.* draws from f_1, for some large N_1. If the power $\hat{\pi}$ of the level α_0 test $\hat{\varphi}$ of (130) is sufficiently close to the upper bound on power $\bar{\pi}$ of all level α_0 tests of (130), stop. Otherwise, decrease ε and go back to Step 3.

$\varphi^{(l),(j)}$ in (133) and the left-hand side of (136) are to be multiplied by $\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{Y}^{(l)})]$, and correspondingly, $\hat{\varphi}$ and $\bar{\varphi}$ of Step 7 Algorithm 4 are to be multiplied by $\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{y})]$, and if the left-hand side of (136) is smaller than α_0 even for $\hat{c} = 0$, then set $\hat{c} = 0$, so that $\bar{\varphi}(\mathbf{y}) = \mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{y})]$ (indicating that the inversion of the credible set $\hat{\Gamma}_p$ to a test of $H_0 : \gamma = \gamma_0$ is already of level α_0). These modifications in particular require evaluating whether $\gamma_0 \in \hat{\Gamma}_p(\mathbf{Y}^{(l)})$ for all $l = 1, \ldots, N$, which for a two-sided credible set amounts to comparing the posterior probability of the event $\{\gamma < \gamma_0\}$ given observations $\mathbf{Y} = \mathbf{Y}^{(l)}$ with the bounds $\alpha_0/2$ and $1 - \alpha_0/2$. Note that all draws $\mathbf{Y}^{(l)}$ for which $\gamma_0 \in \hat{\Gamma}_p(\mathbf{Y}^{(l)})$ effectively drop out of the algorithm, as $\mathbf{1}[\gamma_0 \notin \hat{\Gamma}_p(\mathbf{Y}^{(l)})] = 0$. Thus, the sums over the N draws $\mathbf{Y}^{(l)}$ in (134), (135) and (136) may be replaced with sums over the $N_0 \leq N$ values of $\mathbf{Y}^{(l)}$ where $\gamma_0 \notin \hat{\Gamma}_p(\mathbf{Y}^{(l)})$. This can yield a substantial reduction in the computational burden of the algorithm.

Example BRK(i). Let $\hat{\Gamma}_p(\mathbf{X}_T^s)$ be the level $1 - \alpha$ equal-tailed credible set of Example BRK(f), so that $\mathbf{1}[r_0 \in \hat{\Gamma}_p(\mathbf{X}_T^s)] = \mathbf{1}[\alpha_0/2 \leq \int_0^{r_0} p(r|\mathbf{X}_T^s)dr \leq 1 - \alpha_0/2]$. When applying Algorithm 5 to determine the tests of $H_0 : r = r_0$ whose inversion yields the minimal $\bar{F}$-weighted average expected length level $1 - \alpha_0$ superset of $\hat{\Gamma}_p(\mathbf{X}_T^s)$, simply multiply $\varphi^{(l),(j)}$ in (133) and the indicator function in Steps 6 and 7 by $\mathbf{1}[r_0 \notin \hat{\Gamma}_p(\mathbf{X}_T^s)]$. ▲

As discussed at the end of Section 7.3.3, when the transformation affects the parameter of interest, length optimal confidence sets still invert tests (114) that have a NP-like structure. Note that for the determination of the coverage, this test only needs to be evaluated at the "true" value $\hat{g}(O(\mathbf{Y})^{-1}, h(\theta))$. Algorithms 4 and Steps 1-6 of Algorithm 5 can thus be employed by letting $f_{M,p}(M(\mathbf{y}))f_{O,p}(O(\mathbf{y})^{-1}|M(\mathbf{y}))$ be a suitable proposal for the density $f_{M,p}$ of $M(\mathbf{Y})$ and the conditional density $f_{O,p}(\cdot|M(\mathbf{y}))$ of $O(\mathbf{Y})^{-1}$ given $M(\mathbf{Y})$, $f_1(\mathbf{y}) = \int e(\mathbf{y}|\theta)f_M(\mathbf{y}|\theta)d\bar{F}(\theta)$ and $f(\mathbf{y}|\theta) = f_O(\hat{g}(O(\mathbf{y})^{-1}, h(\theta))|\mathbf{y}, \theta)f_M(\mathbf{y}|\theta)$. For the analogue of Step 7 of Algorithm 5, using (105)

$$\begin{aligned} E_\theta\left[\int \mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{Y})]d\gamma_0\right] &= \int\int \frac{\mathbf{1}[\gamma_0 \in \hat{\Gamma}(\mathbf{y})]e(\mathbf{y}|\theta)f_M(\mathbf{y}|\theta)}{f_{M,p}(M(\mathbf{Y}))f_{O,p}(\gamma_0|\mathbf{y})} f_{M,p}(\mathbf{y})f_{O,p}(\gamma_0|\mathbf{y})d\nu_M(\mathbf{y})d\gamma_0 \\ &= E_p\left[\frac{\mathbf{1}[O(\mathbf{Y})^{-1} \in \hat{\Gamma}(M(\mathbf{Y}))]e(M(\mathbf{Y})|\theta)f_M(M(\mathbf{Y})|\theta)}{f_{M,p}(M(\mathbf{Y}))f_{O,p}(O(\mathbf{Y})^{-1}|M(\mathbf{Y}))}\right] \end{aligned}$$

so that we can obtain an importance sampling estimate of the $\bar{F}$-weighted average expected

length of the sets implied by an invariant test φ via

$$N^{-1}\sum_{l=1}^{N}\frac{f_1(\mathbf{Y}^{(l)})}{f_p(\mathbf{Y}^{(l)})}(1-\varphi(\mathbf{Y}^{(l)})) \tag{137}$$

which, in contrast to Step 7 of Algorithm 5, does not require drawing any new random variables. Applying (137) to the tests $\hat{\varphi}$ and $\bar{\varphi}$ of Step 7 yields the $\bar{F}$-weighted average expected length of the feasible level $1-\alpha_0$ invariant confidence set implied by $\hat{\varphi}$, and a lower bound on that length for all level $1-\alpha_0$ invariant confidence sets. Remarkably, these estimates did not involve determining the length of any realized set directly, which would be computationally more demanding. The necessary modifications to the algorithm in order to obtain the length-optimal superset of a given level $1-\alpha_0$ credible set $\hat{\Gamma}_p$ remains exactly as before.

Example MEAN(q). Following the discussion in Example MEAN(n), let $\bar{F}$ be the discrete distribution with point masses $\{\bar{F}_i\}_{i=1}^m$ on $\{c_i\}_{i=1}^m$ proportional to the reciprocal of (117) for $c=c_i$, where the $\{c_i\}_{i=1}^m$ is the grid of values Example PERS(j). For notational simplicity, write $Y^s=\hat{g}(O(\mathbf{Y})^{-1},h(\theta))=-\bar{x}_{1:T}/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$. With $M(\mathbf{X}_T^0)=(0,\mathbf{X}_T^{s\prime})'$, $\mathbf{X}_T^s=\mathbf{X}_T/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$, a natural proposal would be to generate N *i.i.d.* draws $\mathbf{Z}_T^{s,(l)}=(-Y_T^{s,(l)},\mathbf{X}_T^{s,(l)})$ distributed like $\mathbf{Z}_T^s=(-Y_T^s,\mathbf{X}_T^{s\prime})'=\mathbf{X}_T^0/\sqrt{\mathbf{X}_T'\mathbf{X}_T}$, where $\mathbf{X}_T^0\sim\mathcal{N}(0,T^{-1}\boldsymbol{\Omega}^{LTU}(c))$ with c drawn uniformly from $c\in\{c_i\}_{i=1}^m$. In their Appendix B, Müller and Watson (2016) derive the joint density of $\mathbf{Z}_T^s$ as

$$f(\mathbf{z}_T^s|c)=|\boldsymbol{\Omega}^{LTU}(c)|^{-1/2}\left[\mathbf{z}_T^{s\prime}\boldsymbol{\Omega}^{LTU}(c)^{-1}\mathbf{z}_T^s\right]^{-(q+1)/2}$$

so that

$$f_p(\mathbf{z}_T^s)=m^{-1}\sum_{i=1}^{m}|\boldsymbol{\Omega}^{LTU}(c_i)|^{-1/2}\left[\mathbf{z}_T^{s\prime}\boldsymbol{\Omega}^{LTU}(c_i)^{-1}\mathbf{z}_T^s\right]^{-(q+1)/2}$$

and

$$f_1(\mathbf{x}_T^s)=\sum_{i=1}^{m}\bar{F}_i\sqrt{2\pi}(\mathbf{x}_T^{s\prime}\boldsymbol{\Omega}_{XX}^{LTU}(c_i)^{-1}\mathbf{x}_T^s)^{-1/2}f_M(\mathbf{x}_T^s|c_i).$$

In the context of computing the expected length of sets via (137), it is important to use the appropriate constant of proportionality between f_1 and f_p, which is equal to $\sqrt{2\pi}$ here.

In order to impose that the confidence set is a superset $\hat{\Gamma}_p(\mathbf{X}_T^s)$ of a level $1-\alpha_0$ credible set for Y_T^s given the limited information $\mathbf{X}_T^s$ under prior $\bar{F}$, one needs to evaluate $\mathbf{1}[Y_T^{s,(l)}\notin$

Algorithm 6 Proposal determination for large Θ_0

1. Initialize $f_p^{(1)}(\cdot) = f(\mathbf{y}|\theta^{(1)})$ for some $\theta^{(1)} \in \Theta_0$.

2. Iterating over $j = 1, 2 \ldots$,

   (a) Numerically determine

   $$\max_{\mathbf{y}\in\mathcal{Y},\theta\in\Theta_0} \ln \frac{f(\mathbf{y}|\theta)}{f_p^{(j)}(\mathbf{y})} \tag{138}$$

   (b) Set $f_p^{(j+1)}(\mathbf{y}) = (j+1)^{-1} \sum_{i=1}^{j+1} f(\mathbf{y}|\theta^{(i)})$, where $\theta^{(j+1)}$ is the maximizer of (138).

   (c) Stop iterating when the maximized values $\ln f(\mathbf{y}|\theta)/f_p^{(j)}(\mathbf{y})$ in (138) stabilize.

$\hat{\Gamma}_p(\mathbf{X}_T^{s,(l)})]$. With $\hat{\Gamma}_p(\mathbf{X}_T^{s,(l)})$ the equal-tailed set discussed in Example MEAN(p), this is most easily accomplished by computing, for each $\mathbf{Z}_T^{s,(l)}$, the posterior probability of the event $Y_T^s \leq Y_T^{s,(l)}$ from the mixture of Student-t densities described there. ▲

7.5.3 Algorithms for Larger Θ_0

When the parameter space Θ_0 is large, there are several computational challenges. First, it is not obvious how to construct a proposal density that avoids highly skewed importance weights under all θ. Second, for large Θ_0 it is impossible to enumerate all potential points of support for an (approximate) least favorable distribution. Third, given any candidate level α test, it is not possible to check that it is indeed of level α by computing its rejection probability over a fine grid of values. This section discusses two algorithms that address these challenges that were developed in Müller and Watson (2018) and described in their Appendix A.3.

Consider first the construction of an appropriate proposal density. The major threat to accurate approximations via importance sampling is that for some θ and draws $\mathbf{Y}^{(l)}$ from the proposal density f_p, the importance sampling weights $f(\mathbf{Y}^{(l)}|\theta)/f_p(\mathbf{Y}^{(l)})$ become very large. One way to avoid this is to construct f_p in a way that minimizes $\sup_{\mathbf{y}\in\mathcal{Y},\theta\in\Theta_0} f(\mathbf{y}|\theta)/f_p(\mathbf{y})$. This is a potentially hard problem in general, especially if $\mathbf{y}$ is high dimensional. For moderately low dimensional $\mathbf{y}$ (say, less than 50), we found Algorithm 6 to be effective.

The algorithm constructs f_p as an equal probability mixture of $f(\mathbf{y}|\theta)$ with θ taking on a

finite number of distinct values $\theta^{(i)}$, $i = 1, 2, \ldots, j+1$. (It may make sense to form a mixture of a more variable density $\tilde{f}(\mathbf{y}|\theta)$ instead, such as a multivariate normal with moderately larger variance, as long as it remains easy to generate draws from the mixture.) For most problems, $f(\mathbf{y}|\theta)$ is a quickly evaluated smooth function of both θ and $\mathbf{y}$, so that even though the maximization in (138) is high dimensional, standard hill climbing techniques such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm can be applied. It is not necessary to obtain the global maximizer in each iteration; rather it suffices to eventually capture all θ that lead to large importance weights. For θ equal to $\theta^{(i)}$ for some $i \leq j$, $f(\mathbf{y}|\theta)/f_p^{(j)}(\mathbf{y})$ is bounded above by j. Instead of the vague stopping criterion in 2.c, one may alternatively stop if, say, 200 randomly chosen starting values for a BFGS maximization do not yield a value of $f(\mathbf{y}|\theta)/f_p^{(j)}(\mathbf{y})$ that is larger than some fixed value M, such as $M = 50$. In general, importance sampling estimates from N draws are guaranteed to be no less precise than a those from a standard Monte Carlo with N/M *i.i.d.* draws (cf. equation (127)).

Now consider the issue that large Θ_0 cannot usefully be approximated with discrete grids. A potential solution here is again iterative: Build up an appropriate set of support points of the approximate least favorable distribution by including values of θ that lead to overrejections. A major advantage of the importance sampling approximation to the rejection probability is that it is smooth function in θ, so hill climbing techniques can be applied to identify points of overrejection. A basic version of such an algorithm is given as Algorithm 7.

The numerical maximization in (139) is best approached with hill climbing techniques, such as the BFGS algorithm, and often benefits from explicitly programming the derivative. Here discrete parameter spaces (such as a finite set of local-to-unity parameters) is a major hindrance. One approach to obtain a smooth version is to form a quadratic interpolation between the finite set of precomputed (inverted) covariance matrices, which ensures continuity of the first derivative of the likelihood, and thus (139).

In practice, it often makes sense to collect several (distinct) values of θ that lead to an overrejection in Step 4.b as potential support points for the LFD to avoid numerous and relatively costly recomputations of the LFD weights in Algorithm 4 of Step 4.a. Similarly, one might also drop values of $\theta^{(i)}$ to which the LFD assigns near zero weight, although with this modification, the algorithm could potentially cycle (although we have not encountered this in practice).

Algorithm 7 Determination of a nearly power maximizing level α_0 test of (130) for large Θ_0

1. Apply Algorithm 6 to determine f_p.
2. Draw N *i.i.d.* draws $\mathbf{Y}^{(l)}$ from density f_p, $l = 1, \ldots, N$.
3. Set $\theta^{(1)} \in \Theta_0$ arbitrarily.
4. For $j = 1, 2, \ldots,$
   (a) Apply Algorithm 4 with $\{\theta_i\}_{i=1}^m = \{\theta^{(i)}\}_{i=1}^j$, and $\alpha = \alpha_0 - \varepsilon$ for some small $\varepsilon > 0$ (say, 0.5% for $\alpha_0 = 5\%$).
   (b) Numerically maximize

$$\max_{\theta \in \Theta_0} N^{-1} \sum_{l=1}^{N} \frac{f(\mathbf{Y}^{(l)}|\theta)}{f_p(\mathbf{Y}^{(l)})} \varphi^{(j)}(\mathbf{Y}^{(l)}) \tag{139}$$

   where $\varphi^{(j)}$ is the test $\hat{\varphi}$ of Step 7 of Algorithm 4 obtained in Step 4a.
   (c) If the maximized value in (139) is larger than α_0, then set $\theta^{(j+1)}$ equal to the maximizer of (139), and continue the iteration.
   (d) Otherwise, exit the iteration and proceed as in Steps 6 and 7 of Algorithm 5.

8 A Spectral Analysis Perspective

This section draws the connection between the low-frequency analysis of Section 3 and spectral analysis. The first subsection derives the limiting covariance matrix of the cosine transforms in terms of the spectral density of the underlying time series. A calculation shows that the covariance matrix is fully determined by the shape of the spectral density close to the origin, a function we call the *local-to-zero spectrum*. The second subsection presents a central limit theorem for smooth weighted averages under general assumptions about the shape of the local-to-zero spectrum.

8.1 Local-to-Zero Spectra

We are interested in deriving the limit distribution of the cosine transformations of u_t. Suppose u_t can be written as a function of some mean-zero second-order stationary stochastic process v_t with spectral density Υ_T.[24] We first consider the case where u_t is scalar, and $u_t = v_t$. We then turn to $\Delta u_t = v_t$, as, for instance, in the $I(1)$ model, or, more generally, in the $I(d)$ model with $\frac{1}{2} < d < 3/2$, and then we take up the vector case.

Let $T^{-1}\mathbf{V}_T$ be the covariance matrix of cosine weighted averages of v_t in a sample of size T. Note that $\mathbf{V}_T$ is a function of the autocovariances of v_t, which in turn can be written as a function of the spectrum Υ_T. Furthermore, since the cosine weights are smooth, one would expect that the value of $\mathbf{V}_T$ is largely determined by the low-frequency properties of v_t, that is the behavior of Υ_T close to frequency zero. This suggests that one can obtain asymptotic results under the assumption that for arguments close to zero, Υ_T suitably converges, and this limit behavior of Υ_T then determines the limit $\mathbf{V}_T \to \mathbf{V}$.

Specifically, suppose that in the $O(T^{-1})$ neighborhood of the origin, the suitably scaled spectral density of v_t converges (see assumption (iv) of Theorem 2 below for a precise technical assumption on the convergence and the limit function S)

$$\Upsilon_T(\lambda/T) \to S(\lambda) \tag{140}$$

for some $S : \mathbb{R} \mapsto [0, \infty)$. For example, the $I(d)$ model is usually defined by the assumption

[24] The T subscript on Υ_T accomodates "double-array" processes such as the LTU model in which the AR(1) coefficient depends on T. To ease notation, we omit the corresponding T subscript on $v_{T,t} = v_t$ in this subsection.

that $\Upsilon_T(\phi)$ is proportional to $|\phi|^{-2d}$ for small ϕ. Thus, with the proper scale normalization of Υ_T, (140) holds with $S(\lambda) \propto |\lambda|^{-2d}$. As a second example, for Υ_T the spectrum of an AR(1) process with coefficient $1 - c/T$, a straightforward calculation shows that after appropriate scale normalization, $S(\lambda) \propto 1/(\lambda^2 + c^2)$. More generally, the function $S(\lambda)$ is the large sample limit of the shape of the original spectrum Υ_T close to the origin, and we correspondingly refer to it as the *local-to-zero* spectrum.

Classic spectral analysis considers asymptotics where Υ_T is a bounded and continuous function that does not depend on T. In that case, the limit in (140) yields a constant, which is a flat local-to-zero spectrum, and the limiting properties of $\mathbf{V}_T$ are as if v_t was white noise. It is precisely potential curvature of Υ_T in the T^{-1} neighborhood of the origin as captured by non-flat S that makes classic spectral analysis results inapplicable, and lead to non-trivial low-frequency dynamics in the sense of this chapter.

Consider a weighted average of v_t,

$$W_T = T^{-1/2} \sum_{t=1}^{T} b_{T,t} v_t$$

where the weights $b_{T,t}$ are such that $\sup_t |b_{T,t} - b(t/T)| \to 0$ for some Riemann integrable function $b : [0,1] \mapsto \mathbb{R}$. The cosine transforms are an example of W_T. Recalling that the j-th autocovariance of u_t is given by $\int_{-\pi}^{\pi} \Upsilon_T(\phi) e^{-\mathrm{i}\phi j} d\phi$, where $\mathrm{i} = \sqrt{-1}$, we obtain for the covariance between two such weighted averages, W_T^1 and W_T^2

$$\begin{aligned}
E[W_T^1 W_T^2] &= T^{-1} E\left[\left(\sum_{s=1}^{T} b_{T,s}^1 v_s\right)\left(\sum_{t=1}^{T} b_{T,t}^2 v_t\right)\right] \\
&= T^{-1} \sum_{s=1}^{T}\sum_{t=1}^{T} b_{T,s}^1 b_{T,t}^2 E[v_s v_t] \\
&= T^{-1} \sum_{s=1}^{T}\sum_{t=1}^{T} b_{T,s}^1 b_{T,t}^2 \int_{-\pi}^{\pi} \Upsilon_T(\phi) e^{-\mathrm{i}\phi(s-t)} d\phi \\
&= T^{-1} \int_{-\pi}^{\pi} \Upsilon_T(\phi) \left(\sum_{s=1}^{T} b_{T,s}^1 e^{\mathrm{i}\phi s}\right)\left(\sum_{t=1}^{T} b_{T,t}^2 e^{-\mathrm{i}\phi t}\right) d\phi \\
&= \int_{-\pi T}^{\pi T} \Upsilon_T(\lambda/T) \left(T^{-1}\sum_{s=1}^{T} b_{T,s}^1 e^{\mathrm{i}\lambda s/T}\right)\left(T^{-1}\sum_{t=1}^{T} b_{T,t}^2 e^{-\mathrm{i}\lambda t/T}\right) d\lambda \\
&\to \int_{-\infty}^{\infty} S(\lambda) \left(\int_0^1 b^1(s) e^{\mathrm{i}\lambda s} ds\right)\left(\int_0^1 b^2(s) e^{-\mathrm{i}\lambda s} ds\right) d\lambda. \qquad (141)
\end{aligned}$$

Thus, for $u_t = v_t$ and using the cosine weights in Ψ_T as b_T, the i,j-th element of the limiting covariance matrix $\mathbf{V}$ of the cosine transform of u_t are given by weighted averages of S, with weights of the form $\left(\int_0^1 \Psi_i(s)e^{\mathrm{i}\lambda s}ds\right)\left(\int_0^1 \Psi_j(s)e^{-\mathrm{i}\lambda s}ds\right)$.

Now suppose $\Delta u_t = T^{-1}v_t$, with $v_0 = 0$, that is, $u_t = T^{-1}\sum_{s=1}^t v_s$. Then by summation by parts, with $B_{T,t-1} = T^{-1}\sum_{s=1}^{t-1} b_{T,s}$,

$$\begin{aligned} W_T^{\Delta} &= T^{-1/2}\sum_{t=1}^{T} b_{T,t}u_t \\ &= -T^{1/2}\sum_{t=1}^{T} B_{T,t-1}\Delta u_t + T^{1/2}B_{T,T}u_T \\ &= T^{-1/2}\sum_{t=1}^{T}(B_{T,T} - B_{T,t-1})v_t. \end{aligned} \tag{142}$$

Let $B(s) = \int_0^s b(r)dr$, and note that $T^{-1}\sum_{t=1}^T (B_{T,T} - B_{T,t-1})e^{\mathrm{i}\lambda t/T} \to \int_0^1 (B(1) - B(s))e^{\mathrm{i}\lambda s}ds$. Proceeding as above yields

$$E[W_T^{\Delta 1}W_T^{\Delta 2}] \to \int_{-\infty}^{\infty} S(\lambda)\left(\int_0^1 (B^1(1) - B^1(s))e^{\mathrm{i}\lambda s}ds\right)\left(\int_0^1 (B^2(1) - B^2(s))e^{\mathrm{i}\lambda s}ds\right)d\lambda.$$

Furthermore, by integration by parts,

$$\int_0^1 (B(1) - B(s))e^{\mathrm{i}\lambda s}ds = \frac{\int_0^1 b(s)e^{\mathrm{i}\lambda s}ds - B(1)}{\mathrm{i}\lambda}$$

so that we equivalently obtain

$$E[W_T^{\Delta 1}W_T^{\Delta 2}] \to \int_{-\infty}^{\infty} \frac{S(\lambda)}{\lambda^2}\left(\int_0^1 b^1(s)e^{\mathrm{i}\lambda s}ds - B^1(1)\right)\left(\int_0^1 b^2(s)e^{-\mathrm{i}\lambda s}ds - B^2(1)\right)d\lambda. \tag{143}$$

Consider first the case that $B^1(1) = B^2(1) = 0$, which corresponds to weights $b_{T,t}$ that average to zero. As noted before, the cosine weights have this property. In this case, the limit in (143) is the same function of weights b^1, b^2 and the "pseudo" local-to-zero spectrum $S_p(\lambda) = S(\lambda)/\lambda^2$ as the limit in the stationary case (141). In fact, for the parametric models considered in this chapter, this pseudo local-to-zero spectrum is the natural continuous extension of the local-to-zero spectrum for stationary processes: In the suitably scaled LTU model, as $c \to 0$, $S(\lambda) \to 1/\lambda^2$, which is the pseudo-spectrum of the I(1) model with $u_t = T^{-1}\sum_{s=1}^t v_s$ and $v_t \sim I(0)$ (that is, $S(\lambda)$ is constant). Similarly, define the fractional

model with $d \in (1/2, 3/2)$ via $u_t = T^{-1} \sum_{s=1}^{t} v_s$, $v_t \sim I(d-1)$. Then the fractional model $u_t \sim I(d)$, $d(-1/2, 3/2)$ has limiting spectrum $S(\lambda) = \lambda^{-2d} \to \lambda^{-1}$ for $\lambda \to 1/2$ both from above and below. In particular, this implies that the limiting covariance matrix $\mathbf{V}$ in these models is a continuous function of $c \in [0, \infty)$ and $d \in (-1/2, 3/2)$ as long as the constant with associated flat weights $b_{T,t} = 1$ are not included.

If the weights $B^1(1)$ or $B^2(1)$ are not zero, then this equivalence to the analogous expression with pseudo local-to-zero spectrum S_p does not hold. Intuitively, if the weights do not sum to zero, then the variance of the weighted average also loads on the unconditional variance of the process. But that unconditional variance diverges as one approaches non-stationarity: In the stationary local-to-unity model for u_t, for instance, the variance of u_t diverges as $c \to 0$, and the limit of the variance of $T^{-1/2} \sum_{t=1}^{T} u_t$ is not well defined. In contrast, in the $I(1)$ model with $u_t = T^{-1} \sum_{s=1}^{t} v_s$ and $v_t \sim I(0)$, the variance of $T^{-1/2} \sum_{t=1}^{T} u_t = T^{-1/2} \sum_{t=1}^{T} (1 - (t-1)/T) v_t$ converges to a finite limit.

These notions generalize to vector valued $\mathbf{v}_t \in \mathbb{R}^n$. Let $\mathbb{H}_n$ be the space of $n \times n$ Hermitian matrices. Suppose the $n \times n$ spectral density matrix $\mathbf{\Upsilon}_T : [-\pi, \pi] \mapsto \mathbb{H}_n$ of $\mathbf{v}_t$ converges to

$$\mathbf{\Upsilon}_T(\lambda/T) \to \mathbf{S}(\lambda)$$

for some function $\mathbf{S} : \mathbb{R} \mapsto \mathbb{H}_n$. For two sequences of weights $\mathbf{b}^i_{T,t} \in \mathbb{R}^n$, $i = 1, 2$, satisfying $\sup_t ||\mathbf{b}^i_{T,t} - \mathbf{b}^i(t/T)|| \to 0$ for Riemann integrable functions $\mathbf{b}^i : [0, 1] \mapsto \mathbb{R}^k$, define the weighted averages

$$W^i_T = T^{-1/2} \sum_{t=1}^{T} \mathbf{v}'_t \mathbf{b}^i_{T,t} \text{ and } W^{\Delta i}_T = T^{-1/2} \sum_{t=1}^{T} \left(T^{-1} \sum_{s=1}^{t} \mathbf{v}_s \right)' \mathbf{b}^i_{T,t}.$$

Then proceeding as for (141) yields

$$E[W^1_T W^2_T] \to \int_{-\infty}^{\infty} \left(\int_0^1 \mathbf{b}^1(s) e^{\mathrm{i}\lambda s} ds \right)' \mathbf{S}(\lambda) \left(\int_0^1 \mathbf{b}^2(s) e^{-\mathrm{i}\lambda s} ds \right) d\lambda,$$

and

$$E[W^{\Delta 1}_T W^{\Delta 2}_T] \to \int_{-\infty}^{\infty} \left(\int_0^1 \mathbf{b}^1(s)(e^{\mathrm{i}\lambda s} - 1) ds \right)' \frac{\mathbf{S}(\lambda)}{\lambda^2} \left(\int_0^1 \mathbf{b}^2(s)(e^{-\mathrm{i}\lambda s} - 1) ds \right) d\lambda.$$

8.2 A Central Limit Theorem

The inference results of this chapter crucially depend on the large sample *Gaussianity* of the suitably scaled cosine transform $\mathbf{X}_T$, and not just on the value of the limiting covariance

matrix $\mathbf{V}$ discussed in the previous subsection. The following result, due to Müller and Watson (2016) and Müller and Watson (2017), provides a corresponding CLT.

Theorem 2. *Let* $\mathbf{v}_{T,t} = \sum_{s=-\infty}^{\infty} \mathbf{c}_{T,s}\varepsilon_{t-s}$, *where* $\mathbf{c}_{T,s}$ *are* $n \times n$ *and* ε_t *is* $n \times 1$. *Suppose that*

(i) $\{\varepsilon_t, \mathcal{F}_t\}$ *is a martingale difference sequence with* $E[\varepsilon_t\varepsilon_t'] = \mathbf{\Sigma}_\varepsilon$, $\mathbf{\Sigma}_\varepsilon$ *invertible,* $\sup_t E[||\varepsilon_t||^{2+\delta}] < \infty$ *for some* $\delta > 0$, *and*

$$E[||E[\varepsilon_t\varepsilon_t' - \mathbf{\Sigma}_\varepsilon|\mathcal{F}_{t-m}]||] \leq \xi_m \tag{144}$$

for some sequence $\xi_m \to 0$.

(ii) For every $\epsilon > 0$ *the exists an integer* $L_\epsilon > 0$ *such that* $\limsup_{T\to\infty} T^{-1}\sum_{l=L_\epsilon T+1}^{\infty}\left(T\sup_{|s|\geq l}||\mathbf{c}_{T,s}||\right)^2 < \epsilon$.

(iii) $\sum_{s=-\infty}^{\infty}||\mathbf{c}_{T,s}|| < \infty$ *(but not necessarily uniformly in* T*). The spectral density of* $\mathbf{v}_{T,t}$ *thus exists; denote it by* $\mathbf{\Upsilon}_T : [-\pi, \pi] \mapsto \mathbb{H}_n$, *where* $\mathbb{H}_n$ *is the space of* $n \times n$ *Hermitian matrices.*

(iv.a) Assume that there exists a function $\mathbf{S} : \mathbb{R} \mapsto \mathbb{H}_n$ *such that* $\int_0^1 ||\mathbf{S}(\lambda)||d\lambda < \infty$, $\int_1^\infty ||\mathbf{S}(\lambda)||\lambda^{-2}d\lambda < \infty$ *and for all fixed* K,

$$\int_0^K ||\mathbf{\Upsilon}_T(\frac{\lambda}{T}) - \mathbf{S}(\lambda)||d\lambda \to 0. \tag{145}$$

(iv.b) For every diverging sequence $K_T \to \infty$, $K_T \leq \pi T$,

$$T^{-1}\int_{K_T/T}^{\pi} ||\mathbf{\Upsilon}_T(\phi)||\phi^{-2}d\phi = \int_{K_T}^{\pi T} ||\mathbf{\Upsilon}_T(\lambda/T)||\lambda^{-2}d\lambda \to 0. \tag{146}$$

(iv.c)

$$T^{-1/2}\int_{1/T}^{\pi} ||\mathbf{\Upsilon}_T(\phi)||^{1/2}\phi^{-1}d\phi = T^{-1/2}\int_1^{\pi T} ||\mathbf{\Upsilon}_T(\lambda/T)||^{1/2}\lambda^{-1}d\lambda \to 0. \tag{147}$$

(v) The double array of weights $\mathbf{b}_{T,t}$ *is such that* $\sup_t ||\mathbf{b}_{T,t} - \mathbf{b}(t/T)|| \to 0$ *for a Riemann integrable function* $\mathbf{b} : [0,1] \mapsto \mathbb{R}^n$, *and* $\sup_T \sum_{t=2}^T ||\mathbf{b}_{T,t} - \mathbf{b}_{T,t-1}||^2 < \infty$.

Then

$$T^{-1/2}\sum_{t=1}^{T}\mathbf{v}_{T,t}'\mathbf{b}_{T,t} \Rightarrow \mathcal{N}\left(0, \int_{-\infty}^{\infty}\left(\int_0^1 e^{\mathrm{i}\lambda s}\mathbf{b}(s)ds\right)'\mathbf{S}(\lambda)\left(\int_0^1 e^{-\mathrm{i}\lambda s}\mathbf{b}(s)ds\right)d\lambda\right). \tag{148}$$

Proof. The claim of the theorem generalizes Theorem 1 of Müller and Watson (2017) only in the assumptions about the weights $\mathbf{b}_{T,t}$. The properties that are used in the proof of Müller and Watson (2017) are that $\sup_{t,T} ||\mathbf{b}_{T,t}||$ is finite, which follows from assumption (v), that $\sup_T \sum_{t=2}^T ||\mathbf{b}_{T,t} - \mathbf{b}_{T,t-1}||^2$ (in their Lemma 1), which is assumed in part (v), and that $\sup_{0\leq\lambda\leq K} ||T^{-1}\sum_{t=1}^T \mathbf{b}_{T,t} e^{\mathrm{i}\lambda t/T} - \int_0^1 e^{\mathrm{i}\lambda s}\mathbf{b}(s)ds|| \to 0$ (in their Lemma 2), which follows from

$$\sup_{0\leq\lambda\leq K} ||T^{-1}\sum_{t=1}^T (\mathbf{b}_{T,t} - \mathbf{b}(t/T))e^{\mathrm{i}\lambda t/T}|| \leq \sup_t ||\mathbf{b}_{T,t} - \mathbf{b}(t/T)|| \to 0$$

and the Riemann integrability of $\mathbf{b}$. ■

To better understand the role of assumptions (ii) and (iii), consider some leading examples for scalar series, $n = 1$. Suppose first that $v_{T,t}$ is causal and weakly dependent with exponentially decaying $c_{T,s}$, $|c_{T,s}| \leq C_0 e^{-C_1 s}$ for some $C_0, C_1 > 0$, as would arise in causal and invertible ARMA models of any fixed and finite order. Then $T^{-1}\sum_{l=LT+1}^\infty \left(T\sup_{|s|\geq l}|c_{T,s}|\right)^2 \to 0$ for any $L > 0$, $S(\lambda)$ is constant and equal to the long-run variance of $v_{T,t}$ divided by 2π, and (146) and (147) hold, since Υ_T is bounded, $\int_{K_T}^\infty \lambda^{-2}d\lambda \to 0$ for any $K_T \to \infty$ and $T^{-1/2}\int_1^{\pi T} \lambda^{-1}d\lambda = T^{-1/2}\ln(\pi T) \to 0$.

Second, suppose $v_{T,t}$ is fractionally integrated with parameter $d \in (-1/2, 1/2)$. With $v_{T,t}$ scaled by T^{-d}, $c_{T,s} \approx C_0 T^{-d} s^{d-1}$, so that $T^{-1}\sum_{l=LT+1}^\infty \left(T\sup_{|s|\geq l}|c_{T,s}|\right)^2 \to C_0^2 \int_L^\infty s^{2d-2}ds$, which can be made arbitrarily small by choosing L large. Further, for ϕ close to zero, $\Upsilon_T(\phi) \approx C_0^2(\phi T)^{-2d}$, so that $S(\lambda) = (2\pi)^{-1}C_0^2\lambda^{-2d}$. Under suitable assumptions about about higher frequency properties of $v_{T,t}$, (146) and (147) hold, since $T^{-1}\int_{K_T/T}^\pi (\phi T)^{-2d}\phi^{-2}d\phi = \int_{K_T}^{\pi T}\lambda^{-2d-2}d\lambda \to 0$ and $T^{-1/2}\int_{1/T}^\pi (\phi T)^{-d}\phi^{-1}d\phi = T^{-1/2}\int_1^{\pi T}\lambda^{-d-1}d\lambda = T^{-1/2}d^{-1}(1 - (\pi T)^{-d}) \to 0$. For instance, even integrable poles in Υ_T at frequencies other than zero can be accommodated.

Third, suppose $v_{T,t}$ is an AR(1) process with local-to-unity coefficient $\rho_T = 1 - c/T$ and innovation variance equal to T^{-1}. Then $c_{T,s} = T^{-1}\rho_T^s$, $s \geq 0$. Thus $T^{-1}\sum_{l=LT+1}^\infty \left(T\sup_{|s|\geq l}|c_{T,s}|\right)^2 \to \int_L^\infty e^{-2cs}ds$, which can be made arbitrarily small by choosing L large. Further, $\Upsilon_T(\phi) = (2\pi)^{-1}T^{-2}/|1 - \rho_T e^{-\mathrm{i}\phi}|^2$, which is seen to satisfy (145) with $S(\lambda) = (2\pi)^{-1}(\lambda^2 + c^2)^{-1}$. Conditions (146) and (147) also hold in this example, since $\Upsilon_T(\phi) \leq (2\pi)^{-1}$.

Corresponding central limit theorems for a vector of weighted averages of one or multiple time series follow readily from Theorem 2 by invoking the Cramér-Wold device. If the object

of interest are weighted averages of the non-stationary process $\mathbf{u}_{T,t} = T^{-1}\sum_{s=1}^{t}\mathbf{v}_{T,s}$, then one can invoke Theorem 2 after rewriting the weighted average as in (142) above. This approach can also handle the case where the vector $\mathbf{u}_t$ has components that are of different integration order.

References

Andersen, T. G. and T. Bollerslev (1997). Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *Journal of Finance 52*, 975–1005.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modelling and forecasting realized volatility. *Econometrica 71*, 529–626.

Anderson, T. W. (1984). An introduction to multivariate statistics. *Wiley, New York*.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*, 817–858.

Andrews, D. W. K. and W. Ploberger (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica 62*, 1383–1414.

Bai, J. (1994). Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis 15*, 453–470.

Bai, J., R. L. Lumsdaine, and J. H. Stock (1998). Testing for and dating common breaks in multivariate time series. *Review of Economic Studies 65*, 395–432.

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica 66*, 47–78.

Baillie, R. T. (1996). Long memory processes and frational integration in econometrics. *Journal of Econometrics 73*, 5–59.

Baxter, M. and R. G. King (1999). Measuring business cycles: approximate band-pass filters for economic time series. *Review of economics and statistics 81*(4), 575–593.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (Second ed.). New York: Springer-Verlag.

Bierens, H. J. (1997). Nonparametric cointegration analysis. *Journal of Econometrics 77*, 379–404.

Bobkoski, M. J. (1983). Hypothesis testing in nonstationary time series. *unpublished Ph.D. thesis, Department of Statistics, University of Wisconsin*.

Brillinger, D. R. (2001). *Time series: data analysis and theory*. SIAM.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods* (Second ed.). New York: Springer.

Cavanagh, C. L. (1985). Roots local to unity. *Working Paper, Harvard University*.

Chan, N. H. and C. Z. Wei (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *The Annals of Statistics 15*, 1050–1063.

Dickey, D. A. and W. A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association 74*, 427–431.

Ding, Z., C. W. J. Granger, and R. F. Engle (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance 1*, 83–116.

Domowitz, I. and H. White (1982). Misspecified models with dependent observations. *Journal of Econometrics 20*, 35–50.

Dou, L. (2019). Optimal har inference. *Working Paper, Princeton University*.

Dufour, J.-M. and M. L. King (1991). Optimal invariant tests for the autocorrelation coefficient in linear regressions with stationary or nonstationary AR(1) errors. *Journal of Econometrics 47*, 115–143.

Elliott, G. (1999). Efficient tests for a unit root when the initial observation is drawn from its unconditional distribution. *International Economic Review 40*, 767–783.

Elliott, G. and U. K. Müller (2006). Efficient tests for general persistent time variation in regression coefficients. *Review of Economic Studies 73*, 907–940.

Elliott, G., U. K. Müller, and M. W. Watson (2015). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica 83*, 771–811.

Elliott, G., T. J. Rothenberg, and J. H. Stock (1996). Efficient tests for an autoregressive unit root. *Econometrica 64*, 813–836.

Engle, R. F. (1974). Band spectrum regression. *International Economic Review 15*, 1–11.

Engle, R. F. and C. W. J. Granger (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica 55*, 251–276.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis* (2nd ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association 99*, 799–804.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. Hoboken, New Jersey: John Wiley & Sons.

Geweke, J. and S. Porter-Hudak (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis 4*, 221–238.

Gordon, R. J. (2016). *The Rise and Fall of American Growth: The U.S. Standard of Livining Since the Civil War*. Princeton University Press.

Granger, C. W. J. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis 1*, 15–29.

Grenander, U. and M. Rosenblatt (1957). *Statistical Analysis of Stationary Time Series*. New York: John Wiley and Sons.

Hannan, E. (1970). *Multiple Time Series*. Wiley.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics 2*, 360–378.

Jansson, M. (2004). The error in rejection probability of simple autocorrelation robust tests. *Econometrica 72*, 937–946.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control 12*, 231–254.

Kariya, T. (1980). Locally robust test for serial correlation in least squares regression. *Annals of Statistics 8*, 1065–1070.

Kempthorne, P. J. (1987). Numerical specification of discrete least favorable prior distributions. *SIAM Journal on Scientific and Statistical Computing 8*, 171–184.

Kiefer, N. and T. J. Vogelsang (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory 21*, 1130–1164.

Kiefer, N. M., T. J. Vogelsang, and H. Bunzel (2000). Simple robust testing of regression hypotheses. *Econometrica 68*, 695–714.

King, M. L. (1980). Robust tests for spherical symmetry and their application to least squares regression. *The Annals of Statistics 8*, 1265–1271.

King, M. L. (1987). Towards a theory of point optimal testing. *Econometric Reviews 6*, 169–218.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics 54*, 159–178.

Lazarus, E., D. J. Lewis, and J. H. Stock (2017). The size-power tradeoff in har inference. *manuscript, Harvard University*.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018). Har inference: Recommendations for practice. *Journal of Business and Economic Statistics 36*(4), 541–559.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. New York: Springer.

Moreira, H. and M. Moreira (2013). Contributions to the theory of optimal tests. *Working Paper, FGV/EPGE*.

Müller, U. K. (2004). A theory of robust long-run variance estimation. *Working paper, Princeton University*.

Müller, U. K. (2011). Efficient tests under a weak convergence assumption. *Econometrica 79*, 395–435.

Müller, U. K. (2014). Hac corrections for strongly autocorrelated time series. *Journal of Business and Economic Statistics 32*, 311–322.

Müller, U. K. and L. Dou (2018). Generalized local-to-unity models. *Working paper, Princeton University*.

Müller, U. K. and A. Norets (2016). Credibility of confidence sets in nonstandard econometric problems. *Econometrica 84*, 2183–2213.

Müller, U. K., J. H. Stock, and M. W. Watson (2019). An econometric model of international long-run grwoth dynamics.

Müller, U. K. and M. Watson (2018, May). Long-run covariability. *Econometrica 86*(3), 775–804.

Müller, U. K. and M. W. Watson (2008). Testing models of low-frequency variability. *Econometrica 76*, 979–1016.

Müller, U. K. and M. W. Watson (2013). Low-frequency robust cointegration testing. *Journal of Econometrics 174*, 66–81.

Müller, U. K. and M. W. Watson (2016). Measuring uncertainty about long-run predictions. *Review of Economic Studies 83*.

Müller, U. K. and M. W. Watson (2017). Low-frequency econometrics. In B. Honoré and L. Samuelson (Eds.), *Advances in Economics: Eleventh World Congress of the Econometric Society*, Volume II, pp. 63–94. Cambridge University Press.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Nerlove, M., D. Grether, and J. Carvalho (1979). *Analysis of Economic Time Series*. Academic Press.

Newey, W. K. and K. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–708.

Ng, S. and P. Perron (1995). Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Society 90*, 268–281.

Nordhaus, W. D. (1972). The recent productivity slowdown. *Brookings Papers on Economic Activity*.

Nyblom, J. (1989). Testing for the constancy of parameters over time. *Journal of the American Statistical Association 84*, 223–230.

Perron, P. (2006). Dealing with structural breaks. In K. Patterson and T. C. Mills (Eds.), *Palgrave Handbook of Econometrics, Vol. 1: Econometric Theory*, pp. 278–352. Palgrave Macmillan.

Phillips, P. C. B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika 74*, 535–547.

Phillips, P. C. B. (1998). New tools for understanding spurious regression. *Econometrica 66*, 1299–1325.

Phillips, P. C. B. (2005). Hac estimation by automated regression. *Econometric Theory 21*, 116–142.

Pratt, J. W. (1961). Length of confidence intervals. *Journal of the American Statistical Association 56*, 549–567.

Priestley, M. (1981). *Spectral Analysis and Time Series*. Academic Press.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.

Robert, C. P. (2007). *The Bayesian choice* (second ed.). Springer. Springer, New York.

Robinson, P. M. (2003). Long-memory time series. In P. M. Robinson (Ed.), *Time Series with Long Memory*, pp. 4–32. Oxford: Oxford University Press.

Rogoff, K. (1996). The purchasing power parity puzzle. *Journal of Economic Literature 34*, 647–668.

Stock, J. H. (1991). Confidence intervals for the largest autoregressive root in u.s. macroeconomic time series. *Journal of Monetary Economics 28*, 435–459.

Stock, J. H. (1994). Unit roots, structural breaks and trends. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2740–2841. New York: North Holland.

Sun, Y. (2013). Heteroscedasticity and autocorrelation robust f test using orthonormal series variance estimator. *The Econometrics Journal 16*, 1–26.

Sun, Y., P. C. B. Phillips, and S. Jin (2008). Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. *Econometrica 76*, 175–794.