Efficient Estimation of the Parameter Path in Unstable Time Series Models

ULRICH K. MÜLLER and

PHILIPPE-EMMANUEL PETALAS
Princeton University

The paper investigates inference in non-linear and non-Gaussian models with moderately time-varying parameters. We show that for many decision problems, the sample information about the parameter path can be summarized by an artificial linear and Gaussian model, at least asymptotically. The approximation allows for computationally convenient path estimators and parameter stability tests. Also, in contrast to standard Bayesian techniques, the artificial model can be robustified so that in misspecified models, decisions about the path of the (pseudo-true) parameter remain as good as in a corresponding correctly specified model.

1. INTRODUCTION

One of the central concerns in time series modelling is the stability of parameters through time. A large body of econometric work has developed around testing the hypothesis that parameters are time invariant; see Stock (1994) and Dufour and Ghysels (1996) for surveys and references. Empirically, there is substantial evidence of instabilities in the parameters of finance and macroeconomic models as documented in Stock and Watson (1996), Ghysels (1998), Primiceri (2005) and Cogley and Sargent (2005), just to name a few.

Once instabilities are suspected, a natural next step is to document their form. Knowledge of the parameter path is useful for a number of purposes. First, the estimated path is an interesting descriptive tool, as it helps to understand potential sources of the instability. Second, the end point of the parameter path is useful for forecasting purposes. Third, economic theory might imply certain features of parameter paths (think, for instance, of convergence models with time-varying mean growth of GDP), for which one might want to test in econometric models. Finally, the time-varying value of the parameter can sometimes be given a useful structural interpretation, such as a time-dependent marginal effect in a regression model.

There are several approaches to estimating the parameter path. One strand develops frequentist inference for the break date in models where the parameters are known *a priori* to be subject to a small number of sudden shifts, such as Bai (1997), Bai and Perron (1998), and Elliott and Müller (2007). A Bayesian literature (e.g. Hamilton, 1989; Chib, 1998; Sims and Zha, 2006) posits a finite number of regimes for the parameter values and obtains posterior probabilities for each regime through time. Priestley and Chao (1972), Robinson (1989, 1991), Wu and Zhao (2007), and Cai (2007), among others, develop non-parametric kernel estimators

of the time-varying parameter. Finally, a large frequentist and Bayesian literature estimates models under the assumption of a smooth stochastic evolution of the parameter. When the parameters enter the model linearly and disturbances are assumed Gaussian, then these models can be estimated by variants of Kalman filtering and smoothing—see Harvey (1989) for a review. This is not possible for models with time-varying parameters that affect, say, variances and covariances, and considerably more involved numerical techniques have been developed to deal with such models: see, for instance, Harvey, Ruiz, and Shephard (1994), Jacquier, Polson, and Rossi (1994), Durbin and Koopman (1997), Shephard and Pitt (1997), Kim, Shephard, and Chib (1998), and Primiceri (2005) for the estimation of models with time-varying second moments. In general, the estimation of time-varying parameter models outside the Gaussian state space framework requires fairly complicated and model-specific numerical techniques.

This paper is closely related to this last strand. We consider a general parametric model with local time variation, in the sense that good tests would detect the instability with probability smaller than 1 even in the limit. We analyse estimators and tests that minimize weighted average risk (WAR) and maximize weighted average power over the set of possible parameter paths, where the weighting function is proportional to the distribution function of a Gaussian process, and focuses on such local parameter variability. The main contribution is an asymptotically accurate approximation to the sample information about the parameter path. This approximation turns the problem of inference about the parameter path in the general likelihood model into the problem of inference about the parameter path in a linear Gaussian pseudo model, with the sequence of scores (evaluated at the usual maximum likelihood estimator) as the observations. Asymptotically efficient parameter path estimators and test statistics thus become straightforward to compute, and the estimation and testing problem are unified in one coherent asymptotic framework. In the special case of an underlying parametric model that is stationary for stable parameters, and a weighting that corresponds to the distribution of a Gaussian random walk, the approximate pseudo model can be chosen as a local level model in the sense of Harvey (1989), and optimal path estimators are obtained by an exponential smoothing of the sequence of score vectors. From a Bayesian perspective with the weighting function interpreted as the prior, our results provide an asymptotically accurate multivariate Gaussian approximation to the posterior distribution of the parameter path.

When the likelihood is misspecified, exact Bayesian inference no longer minimizes WAR by construction, even for losses about the pseudo-true parameter value in the sense of White (1982). We extend the ideas in Müller (2009) to construct a robustified pseudo model around the "sandwich" covariance matrix which yields as good asymptotic inference about the parameter path as one would obtain from a correctly specified model with Fisher information equal to the inverse of the sandwich covariance matrix. This robustness property further strengthens the appeal of the suggested approximation over the computationally intensive Bayesian solution, which cannot be easily robustified in the same fashion. Even if the original model is Gaussian and linear, so that the pseudo model approximation can be chosen to be exact in the correctly specified model, inference becomes more reliable in large samples by replacing the original likelihood by the robustified pseudo model.

The asymptotics considered in this paper are such that the magnitude of the instability decreases as the sample size increases. Even asymptotically, there is only limited information about the form of the instability (in contrast to the set-up underlying the non-parametric kernel estimators). We stress that parameter variations that are "small" in the statistical sense of being non-trivial to detect need not be small in an economic sense. For instance, in a stylized model, a sudden shift of 1.2 percentage points in yearly GDP mean growth in the middle of a sample of 180 quarterly observations is detected less than half the time by 5% level efficient stability tests (Elliott and Müller, 2007), yet such a shift is arguably of major economic

(and policy) relevance. Many instabilities that economists care about, such as those arising from Lucas-critique arguments (for instance Linde, 2001), the stability of monetary policy (for instance Bernanke and Mihov, 1998), or reduced-form bivariate econometric relationships between macroeconomic variables in general (Stock and Watson, 1996) have been difficult (or at least non-trivial) to determine empirically and are hence "small" in the statistical sense. In these instances, accurate approximations might well be generated by a modelling strategy in which, correspondingly, there is only limited statistical information about the instability asymptotically.

Our results are driven by a quadratic approximation to the log-likelihood of the general model. Such approximations of the likelihood for models with a finite dimensional parameter have a long history in statistics and econometrics and allow the substitution of a complex decision problem by a simpler one—see, for instance, LeCam (1986). Recent applications in time-series econometrics include Andrews and Ploberger (1994), Phillips and Ploberger (1996), and Ploberger (2004). The sample information about the parameter path is more difficult to approximate, as the path is not finite dimensional. Some numerical methods for time series models with latent variables, such as those developed by Durbin and Koopman (1997) and Shephard and Pitt (1997)—see Durbin and Koopman (2001) for an overview—employ similar quadratic expansions of the log-likelihood at some stage. Brown and Low (1996) and Nussbaum (1996) prove the asymptotic equivalence of some specific infinite dimensional decision problems with the continuous time problem of observing Gaussian white noise with some unknown drift. These papers (essentially) establish the asymptotic equivalence of the frequentist risk for any bounded loss function. Compared to this literature, our results are more specific, as we only show equivalence with respect to WAR, where the weighting functions correspond to the distribution of a (finite mixture of) Gaussian processes. At the same time, our results are substantially more general, as they apply to a wide class of parametric time series models.

The remainder of the paper is organized as follows. The next section heuristically derives the approximating pseudo model, provides a simple algorithm for the path estimator and parameter stability test statistics for a random walk weighting function, and numerically illustrates the ideas with the problem of estimating time-varying variances. Section 3 contains the formal discussion of our results, and Section 4 concludes. All proofs are collected in the Appendix.

2. MOTIVATION AND DEFINITION OF EFFICIENT PARAMETER PATH ESTIMATORS AND STABILITY TESTS

2.1. *Heuristic derivation of approximating pseudo model*

Consider a stationary and stable time series model with known log-likelihood function of the form $\sum_{t=1}^{T} l_t(\theta)$, with parameter $\theta \in \Theta \subset \mathbb{R}^k$. The corresponding unstable model has the same likelihood with time-varying parameter $\{\theta_t\}_{t=1}^{T} = \{\theta + \delta_t\}_{t=1}^{T}$. Suppose the researcher is interested in obtaining path estimators of low expected loss for some given loss function, that is, low risk. Risk depends on the true parameter path $\{\theta + \delta_t\}_{t=1}^{T}$, and no estimator achieves uniformly low risk over all such paths. A reasonable frequentist criterion for the quality of a path estimator thus is WAR, where the weighting is over alternative true parameter paths. In particular, in this paper we derive asymptotically WAR minimizing path estimators for a diffuse weighting of baseline value θ, and a weighting function for the deviations $\{\delta_t\}_{t=1}^{T}$ that correspond to the distribution of a Gaussian process of magnitude $T^{-1/2}$.

The sample information about the path $\{\theta + \delta_t\}_{t=1}^{T}$ is fully contained in the function $\sum l_t(\theta + \delta_t)$, where "$\sum$" denotes a sum over $t = 1, \ldots, T$. Let $\hat{\theta}$ be the maximum likelihood estimator of θ ignoring parameter instability, i.e. $\hat{\theta}$ maximizes $\sum l_t(\theta)$. Denote by

$s_t(\theta) = \partial l_t(\theta)/\partial\theta$ the sequence $t = 1, \ldots, T$ of $k \times 1$ score vectors, and by $h_t(\theta) = -\partial s_t(\theta)/\partial\theta'$ the sequence of $k \times k$ Hessians. By T second-order Taylor expansions of l_t around $\hat{\theta}$

$$\sum(l_t(\theta + \delta_t) - l_t(\hat{\theta})) = \sum[s_t(\hat{\theta})'(\theta + \delta_t - \hat{\theta}) - \tfrac{1}{2}(\theta + \delta_t - \hat{\theta})'h_t(\tilde{\theta}_t)(\theta + \delta_t - \hat{\theta})] \quad (1)$$

where $\tilde{\theta}_t$ lies on the line segment between $\theta + \delta_t$ and $\hat{\theta}$. Suppose the likelihood model is regular enough to ensure a "Local Law of Large Numbers" for the Hessians, such that for sequences $\{\theta_t\}_{t=1}^T$ with θ_t close to $\hat{\theta}$ for $t = 1, \ldots, T$, $T^{-1}\sum h_t(\theta_t) - \hat{H} \xrightarrow{P} 0$, where the matrix $\hat{H}$ is defined as $\hat{H} = T^{-1}\sum h_t(\hat{\theta})$. Since the deviations $\{\delta_t\}_{t=1}^T$ are persistent and of order $T^{-1/2}$, and the maximum likelihood estimator $\hat{\theta}$ is a $\sqrt{T}$-consistent estimator of the baseline value θ, the sequence $\{\theta + \delta_t - \hat{\theta}\}_{t=1}^T$ is persistent and of order $T^{-1/2}$. Also, because the stable model is assumed stationary, smooth averages of $h_t(\tilde{\theta}_t)$ are close to $\hat{H}$ in all parts of the sample, so that

$$\sum(\theta + \delta_t - \hat{\theta})'h_t(\tilde{\theta}_t)(\theta + \delta_t - \hat{\theta}) \simeq \sum(\theta + \delta_t - \hat{\theta})'\hat{H}(\theta + \delta_t - \hat{\theta}). \quad (2)$$

One might think that a more accurate approximation of $h_t(\tilde{\theta}_t)$ is given by $h_t(\hat{\theta})$ rather than by $\hat{H}$ as in (2). But this is not necessarily the case. A (local) average of $h_t(\hat{\theta})$ might well be a good approximation to the (local) average of $h_t(\tilde{\theta}_t)$ even if the approximation $h_t(\tilde{\theta}_t) \simeq h_t(\hat{\theta})$ is poor, and given that δ_t is persistent, only the (local) average of $h_t(\tilde{\theta}_t)$ matters.

Using (2), we obtain

$$\sum(l_t(\theta + \delta_t) - l_t(\hat{\theta}) - \tfrac{1}{2}s_t(\hat{\theta})'\hat{H}^{-1}s_t(\hat{\theta}))$$
$$\simeq -\tfrac{1}{2}\sum(s_t(\hat{\theta}) - \hat{H}(\theta + \delta_t - \hat{\theta}))'\hat{H}^{-1}(s_t(\hat{\theta}) - \hat{H}(\theta + \delta_t - \hat{\theta})). \quad (3)$$

Neither $\sum l_t(\hat{\theta})$ nor $\sum s_t(\hat{\theta})'\hat{H}^{-1}s_t(\hat{\theta})$ depend on $\{\theta + \delta_t\}_{t=1}^T$, so that ignoring these constants, the log-likelihood of the path $\{\theta + \delta_t\}_{t=1}^T$ is well approximated by a quadratic form.[1] In fact, the right-hand side of (3) is recognized as the log-likelihood function of the Gaussian random variable $s_t(\hat{\theta}) + \hat{H}\hat{\theta}$ with mean $\hat{H}(\theta + \delta_t)$ and covariance matrix $\hat{H}$. The information in the sample about $\theta + \delta_t$ can therefore be approximately summarized by the pseudo model

$$s_t(\hat{\theta}) + \hat{H}\hat{\theta} = \hat{H}(\theta + \delta_t) + v_t, \; t = 1, \ldots, T \quad (4)$$

with $v_t \sim i.i.d.\,\mathcal{N}(0, \hat{H})$. The pseudo model (4) links the observed variables on the left-hand side with the object of interest $\{\theta + \delta_t\}_{t=1}^T$ in a particularly straightforward manner, as the matrix multiplying $\theta + \delta_t$ does not depend on t.

For a weighting function for the baseline value θ that is diffuse, the weighting on the mean $T^{-1}\sum\delta_t$ in (4) has no bearing on the analysis. For convenience, one might thus assume a weighting function for $\{\delta_t\}_{t=1}^T$ that corresponds to the distribution of a *demeaned* Gaussian process (so that $\sum\delta_t = 0$ and δ_t is the deviation at date t from the average parameter value θ). Under that assumption, we trivially have $\sum\delta_t'\hat{H}(\theta - \hat{\theta}) = 0$, and also $\sum s_t(\hat{\theta}) = 0$ from the first-order condition of the maximum likelihood estimator. Thus, the right-hand side of (3) becomes

$$-\tfrac{1}{2}\sum(s_t(\hat{\theta}) - \hat{H}\delta_t)'\hat{H}^{-1}(s_t(\hat{\theta}) - \hat{H}\delta_t) - \tfrac{1}{2}T(\theta - \hat{\theta})'\hat{H}(\theta - \hat{\theta})$$

---

1. Shephard and Pitt (1997) and Durbin and Koopman (1997) employ second-order Taylor expansion of the log-likelihood as in (1) to derive proposal densities for their simulation based analysis of non-Gaussian state space models, but they do not consider the additional simplification of the approximating model provided by (2).

and the sample information about θ and $\{\delta_t\}_{t=1}^T$ is approximately independent and described by the pseudo model

$$\hat{\theta} = \theta + T^{-1/2}\hat{H}^{-1}v_0 \tag{5}$$

$$s_t(\hat{\theta}) = \hat{H}\delta_t + v_t, \quad t = 1, \ldots, T \tag{6}$$

with $v_t \sim i.i.d.\,\mathcal{N}(0,\hat{H})$. The approximation in (5) is the standard Bernstein–von Mises result that in large samples, the likelihood about a parameter converges to that of a Gaussian random variable with mean θ and covariance matrix $T^{-1}\hat{H}^{-1}$. The focus and contribution of this paper is to argue for the Gaussian "local level" model (6) (or, equivalently, for (4)) as an asymptotically efficient summary of the sample information about the deviations $\{\delta_t\}_{t=1}^T$. For weighting functions for $\{\delta_t\}_{t=1}^T$ that are Markovian, the information about the parameter path can then be extracted by variants of the Kalman smoother. Also, asymptotically efficient tests of parameter instability in the general likelihood model can be obtained by performing an optimal test in the pseudo model.

Now suppose that the likelihood is misspecified. As demonstrated by White (1982), $\hat{\theta}$ then consistently estimates the pseudo-true parameter θ_0 in a stable model, and $\hat{\theta}$ has an asymptotically Gaussian sampling distribution with the "sandwich" covariance matrix S, $\sqrt{T}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0,S)$. This sandwich matrix is typically consistently estimated by $\hat{S} = \hat{H}^{-1}\hat{V}\hat{H}^{-1}$, where $\hat{V}$ is a consistent estimator of the long-run variance of $s_t(\theta_0)$, such as $\hat{V} = T^{-1}\sum s_t(\hat{\theta})s_t(\hat{\theta})'$ if the scores remain uncorrelated under misspecification, or a Newey and West (1987)-type estimator if not. At the same time, the Taylor expansions leading to (5) and (6) are heuristically valid even if $\sum l_t(\theta)$ does not describe the true likelihood. There is a mismatch between the pseudo model (5), $\hat{\theta} \sim \mathcal{N}(\theta, \hat{H}^{-1}/T)$, and the approximate sampling distribution $\hat{\theta} \sim \mathcal{N}(\theta, S/T)$ under misspecification. Müller (2009) shows that due to this discrepancy, lower risk decisions about the pseudo-true parameter value are obtained by using the "correct" model $\hat{\theta} \sim \mathcal{N}(\theta, \hat{S}/T)$ under parameter stability. The analogous robustified pseudo model in the context of inference about the pseudo-true parameter path is given by

$$\hat{\theta} = \theta + T^{-1/2}\hat{S}v_0 \tag{7}$$

$$\hat{H}\hat{V}^{-1}s_t(\hat{\theta}) = \hat{S}^{-1}\delta_t + v_t, \quad t = 1, \ldots, T \tag{8}$$

with $v_t \sim i.i.d.\,\mathcal{N}(0,\hat{S}^{-1})$. Note that the long-run variance of the robustified scores $\hat{H}\hat{V}^{-1}s_t(\theta_0)$ equals the robustified average Hessian S^{-1}, so that (7) and (8) behave like a correctly specified model with Fisher Information S^{-1}. Also, if the model is correctly specified, $\hat{V} - \hat{H} \xrightarrow{p} 0$ by the information matrix equality and $\hat{S}^{-1} - \hat{H} \xrightarrow{p} 0$, so that the robustified pseudo model is large sample equivalent to the pseudo model (5) and (6).

2.2. *Parameter path estimator and test statistic for random walk parameter evolution*

We now turn to an explicit description of the optimal parameter path estimator and test statistics assuming an approximately stationary model and a weighting function for δ_t that is a (demeaned) multivariate Gaussian random walk. We allow for a potential misspecification of the likelihood, and assume that under misspecification, the object of interest is the evolution of the pseudo-true parameter, so that inference is based on the robust local level pseudo model (7) and (8).

A random walk weighting function (or prior in a Bayesian context) has been used extensively in econometric applications: see, for instance, Harvey (1989), Stock and Watson (1996),

Stock and Watson (1998), Stock and Watson (2002), Boivin (2003), Primiceri (2005), and Cogley and Sargent (2005). Without loss of generality, let the first $p \leq k$ parameters of θ, denoted β, be those whose path is to be estimated (so that the last $k - p$ elements of δ_t are zero). A theoretically attractive choice for the covariance matrix of the Gaussian random walk is to let the first p elements of $\{\delta_t\}_{t=1}^T$ to be proportional to the corresponding elements of $\hat{S}$, the inverse of the information. This choice equates the degree of uncertainty about the time variation of β in any given direction with the average sample information about that direction, and hence leads to equal signal-to-noise ratios in all unstable directions. Also, under this choice, asymptotic results remain identical under re-parameterizations of β. For the factor of proportionality c^2/T^2, we suggest a default choice of minimizing WAR relative to an equal-probability mixture of $n_G = 11$ values $c \in \{0, 5, 10, \ldots, 50\}$. The value c is interpreted as the standard deviation of the end point of the random walk weighting function, measured in multiples of the standard deviation of the full sample parameter estimator. The suggested list of values for c thus cover a wide range of magnitudes for the time variation. An approximately WAR minimizing path estimator under truncated quadratic loss with large truncation point, and large sample weighted average power maximizing parameter stability test, are obtained as follows:

1. For $t = 1, \ldots, T$, let x_t and $\tilde{y}_t$ be the first p elements of $\hat{H}^{-1}s_t(\hat{\theta})$ and $\hat{H}\hat{V}^{-1}s_t(\hat{\theta})$, respectively.
2. For $c_i \in C = \{0, 5, 10, \ldots, 50\}$, $i = 0, \ldots, 10$, compute

   (a) $r_i = 1 - c_i/T$, $z_{i,1} = x_1$ and $z_{i,t} = r_i z_{i,t-1} + x_t - x_{t-1}$, $t = 2, \ldots, T$;
   (b) the residuals $\{\tilde{z}_{i,t}\}_{t=1}^T$ of a linear regression of $\{z_{i,t}\}_{t=1}^T$ on $\{r_i^{t-1}I_p\}_{t=1}^T$;
   (c) $\overline{z}_{i,T} = \tilde{z}_{i,T}$, and $\overline{z}_{i,t} = r_i \overline{z}_{i,t+1} + \tilde{z}_{i,t} - \tilde{z}_{i,t+1}$, $t = 1, \ldots, T - 1$;
   (d) $\{\hat{\beta}_{i,t}\}_{t=1}^T = \{\hat{\theta} + x_t - r_i \overline{z}_{i,t}\}_{t=1}^T$;
   (e) $\text{qLL}(c_i) = \sum_{t=1}^T (r_i \overline{z}_{i,t} - x_t)' \tilde{y}_t$ and $\tilde{w}_i = \sqrt{T(1 - r_i^2)r_i^{T-1}/(1 - r_i^{2T})} \exp[-\frac{1}{2}\text{qLL}(c_i)]$ (set $\tilde{w}_0 = 1$).

3. Compute $w_i = \tilde{w}_i / \sum_{j=0}^{10} \tilde{w}_j$.
4. The parameter path estimator is given by $\{\hat{\beta}_t\}_{t=1}^T = \{\sum_{i=0}^{10} w_i \hat{\beta}_{i,t}\}_{t=1}^T$.
5. The statistic qLL(10) tests the null hypothesis of stability of β and rejects for small values. Critical values depend on p and are tabulated in Table 1 of Elliott and Müller (2006).

In many applications, it will be of interest to get some sense of the accuracy of the path estimator $\{\hat{\beta}_t\}_{t=1}^T$. One such measure is given by the variances

$$\Omega_t = \sum_{i=0}^{10} w_i (T^{-1}\hat{S}_\beta \kappa_t(c_i) + (\hat{\beta}_{i,t} - \hat{\beta}_t)(\hat{\beta}_{i,t} - \hat{\beta}_t)'), \quad \kappa_t(c) = \frac{c(1 + e^{2c} + e^{2ct/T} + e^{2c(1-t/T)})}{2e^{2c} - 2}$$

where $\hat{S}_\beta$ is the upper left $p \times p$ block of $\hat{S} = \hat{H}^{-1}\hat{V}\hat{H}^{-1}$ and $\kappa_t(0) = 1$. From a Bayesian perspective with the weighting function for $\{\delta_t\}_{t=1}^T$ and θ interpreted as priors, Ω_t is the covariance matrix of the approximate posterior for β_t. This approximate posterior distribution is a mixture of multivariate normals $\mathcal{N}(\hat{\beta}_{i,t}, T^{-1}\hat{S}_\beta \kappa_t(c_i))$, $i = 0, \ldots, 10$, with mixing probabilities w_i. The interval $[\hat{\beta}_{t,j} - 1.96\sqrt{\Omega_{t,jj}}, \hat{\beta}_{t,j} + 1.96\sqrt{\Omega_{t,jj}}]$ with $\hat{\beta}_{t,j}$ the j-th element of $\hat{\beta}_t$ and $\Omega_{t,jj}$ the (j,j) element of Ω_t is thus approximately the 95% equal-tailed posterior probability interval for $\beta_{t,j}$, the j-th component of β at time t (one could, of course, also determine the exact 95% interval for the given mixture of normals posterior, with typically very similar results). This interval is not a confidence interval in the frequentist sense, but it can be justified without explicit Bayesian reasoning as a WAR minimizing interval estimator—see Chapter 5.2.5 of Schervish (1995) and the example below.

TABLE 1

Weighted average risks in time varying variances model

df	c = 4			c = 8			c = 12		
	∞	12	6	∞	12	6	∞	12	6
	Average square loss, $T = 160$								
Known c, local level	1.02	1.02	1.02	1.17	1.17	1.13	1.58	1.55	1.43
Known c, Kalman	1.01	1.01	1.01	1.05	1.06	1.05	1.17	1.19	1.16
Unknown c, Bayesian	1.25	1.55	2.10	1.13	1.26	1.53	1.10	1.16	1.33
Unknown c, local level	1.17	1.22	1.21	1.29	1.30	1.25	1.73	1.70	1.57
Unknown c, Kalman	1.15	1.18	1.17	1.12	1.13	1.11	1.24	1.25	1.21
	Average square loss, $T = 480$								
Known c, local level	1.01	1.00	0.99	1.07	1.06	1.03	1.25	1.20	1.15
Known c, Kalman	1.00	1.00	0.99	1.02	1.01	0.99	1.04	1.04	1.02
Unknown c, Bayesian	1.26	1.67	2.89	1.13	1.32	1.93	1.09	1.20	1.58
Unknown c, local level	1.19	1.21	1.26	1.20	1.18	1.18	1.37	1.32	1.29
Unknown c, Kalman	1.17	1.19	1.21	1.11	1.10	1.09	1.12	1.11	1.09
	End point interval estimation loss, $T = 160$								
Known c, local level	1.01	1.01	0.93	1.09	1.10	0.99	1.39	1.36	1.21
Known c, Kalman	1.01	1.00	0.93	1.03	1.04	0.96	1.10	1.09	1.02
Unknown c, Bayesian	1.22	1.31	1.37	1.14	1.17	1.17	1.13	1.12	1.11
Unknown c, local level	1.20	1.22	1.15	1.26	1.26	1.15	1.56	1.53	1.36
Unknown c, Kalman	1.20	1.22	1.16	1.20	1.20	1.12	1.28	1.27	1.21
	End point interval estimation loss, $T = 480$								
Known c, local level	1.01	0.98	0.91	1.03	1.01	0.92	1.15	1.10	1.01
Known c, Kalman	1.01	0.98	0.91	1.01	1.00	0.91	1.03	1.02	0.95
Unknown c, Bayesian	1.24	1.35	1.65	1.13	1.16	1.30	1.11	1.13	1.18
Unknown c, local level	1.21	1.21	1.21	1.20	1.17	1.11	1.29	1.26	1.17
Unknown c, Kalman	1.21	1.21	1.20	1.17	1.15	1.09	1.18	1.18	1.11

Notes: Data generating process parameters are in columns, estimation procedures in rows. Entries are weighted average risk relative to Bayesian inference in model (9) with c known based on 3200 data draws. "Unknown c, local level" inference is as described in Section 2.2, and "Known c, local level" inference is based on the pseudo model (7) and (8) and the column weighting function. "Kalman" inference is based on the pseudo model (23) below with $s_t^r(\hat{\theta}) = \hat{H}\hat{V}^{-1}s_t(\hat{\theta})$ and $\bar{h}_t^r = \hat{H}\hat{V}^{-1}\sum_{s=1}^{T}\phi(T^{-4/5}(s-t))h_t(\hat{\theta})/\sum_{s=1}^{T}\phi(T^{-4/5}(s-t))$, where ϕ is the density of $\mathcal{N}(0,1)$, combined with the column weighting function in the "Known c" rows, and with an equal probability mixture of random walks weighting function with variances $c^2\hat{S}/T^2$, $c \in C$, in the "Unknown c" rows as described in Theorems 4 and 5 below, implemented using the algorithm in the Appendix. "Unknown c, Bayesian" inference is based on a uniform discrete prior on $\{0, 1, 2, \ldots, 50\}$ for c. Posteriors are estimated by a combination of importance sampling (with a "Kalman"-type approximation as proposal) and Gibbs sampling using the algorithm described in Kim, Shephard, and Chib (1998).

2.3. *Time varying variances example*

We now turn to a numerical illustration of these ideas. Specifically, consider the problem of estimating the path of the log-standard deviations of a univariate time series,

$$y_t = \exp[\theta_t]\varepsilon_t, \quad \varepsilon_t \sim i.i.d.\ \mathcal{N}(0,1), \quad t = 1, \ldots, T \tag{9}$$

so that, up to a constant, $l_t(\theta) = -\theta - \frac{1}{2}\exp[-2\theta]y_t^2$, $s_t(\theta) = -1 + \exp[-2\theta]y_t^2$, $h_t(\theta) = 2\exp[-2\theta]y_t^2$, $\hat{\theta} = \frac{1}{2}\ln[T^{-1}\sum y_t^2]$ and $\hat{H} = H = 2$. The specification (9) has been used as a building block to model time-varying variances in macroeconomics and finance—see, for instance, Jacquier, Polson, and Rossi (1994), Durbin and Koopman (1997), Kim, Shephard, and Chib (1998), Shephard and Pitt (1997), Stock and Watson (2002), Primiceri (2005) and Cogley and Sargent (2005). These papers develop and apply Bayesian methods for estimating the parameter path θ_t. By minimizing posterior expected loss for each observed data set, Bayesian methods also minimize WAR with weights equal to the prior in a correctly specified

model. It thus makes sense to use Bayesian inference as a benchmark for the WAR of the path estimator described above. While the estimation of all models is based on the likelihood of (9), we also compute risk for data that is drawn from

$$y_t = \exp[\theta_t]\sqrt{\frac{\text{df}-2}{\text{df}}}\varepsilon_t, \quad \varepsilon_t \sim i.i.d.\text{student-t(df)}, \text{ df} > 2, \quad t = 1, \ldots, T, \quad (10)$$

so that the maintained model (9) is misspecified for df $< \infty$. Note that θ_t, the log-standard deviation of y_t, remains the pseudo-true parameter in (10) when estimating (9).

In addition to the path estimator based on the local level pseudo model (7) and (8) of Section 2.2, we also consider inference based on a robustified pseudo model that does not replace $h_t(\tilde{\theta}_t)$ by the constant $\hat{H}$ in (2), but by a kernel smoothed average of $h_t(\hat{\theta})$. This pseudo model leads to a somewhat more involved Kalman-smoother-based algorithm for obtaining an optimal parameter path estimate described in the Appendix. The potential advantage is higher approximation accuracy, as the smoother takes into account some low-frequency movements in $h_t(\tilde{\theta}_t)$.

We compare WAR in two decision problems: (i) estimation of the parameter path under mean square error loss, so that for a path estimate $\{a_t\}_{t=1}^T$, loss is given by $T^{-1}\sum(\theta_t - a_t)^2$ (and risk becomes mean squared error averaged over t); (ii) estimation of an interval $[a_l, a_h]$ for the end point of the parameter path θ_T, with loss equal to $a_h - a_l + 40 \cdot \mathbf{1}[\theta_T < a_l](a_l - \theta_T) + 40 \cdot \mathbf{1}[\theta_T > a_u](\theta_T - a_u)$ (so that a 10% increase in risk is equivalent to systematically reporting 10% longer intervals with the same coverage probability, and with end points that are no closer to θ_T when θ_T falls outside the interval).[2] Under the approximation discussed in Section 2.2 (with $p = k = 1$ and $\{\hat{\beta}_t\}_{t=1}^T = \{\hat{\theta}_t\}_{t=1}^T$), the best decisions are given by $\{a_t\}_{t=1}^T = \{\hat{\theta}_t\}_{t=1}^T$ and, using Proposition 5.78 of Schervish (1995), $[\hat{a}_l, \hat{a}_h] = [\hat{\theta}_T - 1.96\sqrt{\Omega_{t,jj}}, \hat{\theta}_T + 1.96\sqrt{\Omega_{t,jj}}]$, respectively.

WAR equals expected loss for data that is generated with parameters randomly drawn from the weighting function. Table 1 reports relative WAR estimated in this way for the weighting function $\theta_0 \sim \mathcal{N}(0, 100)$ and $\theta_t - \theta_{t-1} \sim i.i.d.\mathcal{N}(0, c^2/HT^2)$ for $c = 4, 8, 12$ and $T = 160, 480$ (think of 40 years of quarterly and monthly data, respectively). Under this weighting function, the median range of $\{\theta_t\}_{t=1}^T$ is approximately $1.1c/\sqrt{T}$, and $1.1c/\sqrt{T} \simeq 0.70$ for $c = 8$ and $T = 160$, which compares with the estimated range of the log-standard deviation of the US four-quarter growth rate of about 0.59 (cf. Table 1, Stock and Watson, 2002). By inverting the QLR test statistic, Stock and Watson (1996) obtain median unbiased estimates for c for the parameters of 76 univariate AR(6) models of US post-war macroeconomic monthly time series, and never find an estimate larger than 12. Cogley and Sargent (2005) estimate time-varying coefficients and volatility of a monetary VAR and report that this time variation would be detected by a 5% level parameter stability test about 25% of the time, which roughly corresponds to the columns with $c = 4$ in Table 1. This evidence suggests that the degree of instability implied by the weighting functions considered in Table 1 are moderate to large by an empirical standard.

Except for Monte Carlo error, the entries under "df $= \infty$, known c" must be larger than unity by the small sample optimality of Bayesian inference in the correctly specified model. At the same time, the results of Section 3 show that these entries are approximately equal to 1 for large sample sizes. For "unknown c", the moderately lower risk of pseudo-model-based inference relative to Bayesian inference in the correctly specified model with $c = 4$ seems to

---

2. The theoretical development in Section 3 assumes loss to be bounded. We computed WARs with truncated loss functions and truncation point 40 times larger than the median loss, and found results very similar to those reported in Table 1.

stem from a moderate downward bias in the estimated c induced by the robustification. Under misspecification, Bayesian inference is no longer optimal by construction, and inference based on robust pseudo models does relatively better. This effect is especially pronounced when c is unknown, since Bayesian inference rationalizes outliers generated by the student-t disturbances by variation in θ_t, leading to an upward biased posterior for c, and a corresponding under-smoothing of the parameter path. For very large instabilities $c = 12$, the simple algorithm of Section 2.2 has substantially larger WAR relative to Bayesian inference, but the more complicated Kalman pseudo model continues to provide quite accurate approximations.

In the Supplementary Materials,[3] we report additional computations for true paths that are either a linear trend or a step function with known or unknown break date. As expected, path estimators that impose the correct parametric restriction have lower risk relative to the Local Level and Kalman smoother path estimators, at least if the magnitude of the instability is large. At the same time, in the single break model with small or moderate break (less than 6 standard deviations of the full sample estimator $\hat{\theta}$), the Local Level and Kalman estimators outperform estimators of the path that rely on a break date estimated by the least squares mean shift of $|y_t|$ or y_t^2.

3. ASYMPTOTICALLY EFFICIENT INFERENCE IN UNSTABLE TIME SERIES MODELS

We begin by introducing some additional notation and definitions. Consider a standard parametric model for data $\mathbf{Y}_T = (y_{T,1}, \ldots, y_{T,T}) \in \mathbb{R}^{mT}$ in a sample of size T, a random vector defined on the complete probability space $(\mathcal{F}, \mathfrak{F}, P)$, with parameter $\theta \in \Theta \subset \mathbb{R}^k$, and known density $\prod_{t=1}^{T} f_{T,t}(\theta)$ with respect to some σ-finite measure μ_T. This form of likelihood arises naturally in the "forecasting error decomposition" of models, where $f_{T,t}(\theta)$ is the conditional likelihood of $y_{T,t}$ given $\mathfrak{F}_{T,t-1}$, where $\mathfrak{F}_{T,t} \subset \mathfrak{F}$ is the σ-field generated by $\{y_{T,s}\}_{s=1}^{t}$. In models with weakly exogenous components, $f_{T,t}(\theta)$ can be decomposed into two pieces $f_{T,t}(\theta) = f_{T,t}^1(\theta) f_{T,t}^2$, where $f_{T,t}^2$ captures the contribution of the evolution of weakly exogenous components and does not depend on θ. If this is the case, only $f_{T,t}^1(\theta)$ needs to be specified. Define $l_{T,t}(\theta) = \ln f_{T,t}^1(\theta)$, $s_{T,t}(\theta) = \partial l_{T,t}(\theta)/\partial \theta$, and $h_{T,t}(\theta) = -\partial s_{T,t}(\theta)/\partial \theta'$. In the following definitions and conditions, we omit the dependence on T of $\mathfrak{F}_{T,t}$, $l_{T,t}$, $s_{T,t}$, $h_{T,t}$, and so forth, to enhance readability. Let $[\cdot]$ indicate the largest lesser integer function, let $||\cdot||$ denote the spectral norm, let "$\otimes$" be the Kronecker product, and let "$\xrightarrow{p}$" and "$\Rightarrow$" denote convergence in probability and convergence in distribution as $T \to \infty$, respectively. Convergences of cadlag functions on the unit interval are relative to the usual Billingsley (1968)-metric.

We assume the following condition on this model with true and stable parameter θ_0.

Condition 1 (MEAS). *The functions $f_{T,t}^1 : \mathbb{R}^m \times \Theta \mapsto \mathbb{R}$ are jointly measurable for $t = 1, \ldots, T$.*

(DIFF) θ_0 is an interior point of Θ, and in some neighbourhood $\Theta_0 \subseteq \Theta$ of θ_0, l_t is twice continuously differentiable a.s. for $t = 1, \ldots, T$.

(ID) There exists $\eta > 0$ such that for all $\epsilon > 0$ there exists $K(\epsilon) > 0$ for which $P(\sup_{||\theta-\theta_0|| \geq \epsilon} T^{-1} \sum \sup_{||v|| < T^{-1/2+\eta}, \theta+v \in \Theta} (l_t(\theta + v) - l_t(\theta_0)) < -K(\epsilon)) \to 1$

(LLLN) (i) For any decreasing ball $\mathcal{B}_T$ around θ_0, i.e. $\mathcal{B}_T = \{\theta : ||\theta - \theta_0|| < b_T\}$ for a sequence of real numbers $b_T \to 0$, $T^{-1} \sum_{t=1}^{T} \sup_{\theta \in \mathcal{B}_T} ||h_t(\theta) - h_t(\theta_0)|| \xrightarrow{p} 0$, (ii) $T^{-1} \sum_{t=1}^{T} ||h_t(\theta_0)|| = O_p(1)$ and (iii) $\sup_{\lambda \in [0,1]} \left\| T^{-1} \sum_{t=1}^{[\lambda T]} h_t(\theta_0) - \int_0^\lambda \Gamma(l) dl \right\| \xrightarrow{p} 0$ for some nonstochastic matrix function Γ (possibly indexed by θ_0), with $\Gamma(\lambda)$ positive definite for all $\lambda \in [0, 1]$.

---

3. Supplementary Materials are located on the ReStud website. See http://www.restud.org.uk/supplementary.asp

(MDA) $\{s_t(\theta_0), \mathfrak{F}_t\}$ *is a martingale difference array, there exists* $\epsilon > 0$ *such that* $T^{-1} \sum_{t=1}^{T}$
$E[||s_t(\theta_0)||^{2+\epsilon} | \mathfrak{F}_{t-1}] = O_p(1)$ *and* $\sup_{\lambda \in [0,1]} ||T^{-1} \sum_{t=1}^{[\lambda T]} E[s_t(\theta_0) s_t(\theta_0)' | \mathfrak{F}_{t-1}] - \int_0^\lambda \Gamma(l) dl ||$
$\xrightarrow{p} 0$.

Condition 1 is a set of fairly standard high-level assumptions on the "forecast error decomposition" part of the likelihood. (DIFF) assumes existence of two derivatives. (ID) is similar to the global identification condition assumed in Schervish (1995, p. 436), somewhat strengthened to ensure that even a slightly perturbed evaluation of the likelihood at parameter values different from θ_0 still yields a lower likelihood with high probability. (LLLN) is a Local Law of Large Numbers for the second derivatives h_t. Part (i) controls the average variability of the second derivative h_t as a function of the parameter. Part (iii) allows the information accrual to vary over the sample, and $\Gamma(\lambda)$ describes the average information at time $t = [\lambda T]$. This allows, for instance, accommodating regression models with a time trend t/T as regressor (the scaling by $1/T$ ensures that the probability limit of $T^{-1} \sum_{t=1}^{[\lambda T]} h_t(\theta_0)$ remains $O_p(1)$ and positive definite). If $h_t(\theta_0)$, $t = 1, \ldots, T$ is positive semi-definite almost surely, part (ii) of (LLLN) is implied by part (iii). (MDA) assumes the sequence of scores to constitute a martingale difference array with slightly more than two conditional moments, with an average conditional variance of $\Gamma(\lambda)$ at time $t = [\lambda T]$. Whenever the relevant conditional moments exist, $\{s_t(\theta_0), \mathfrak{F}_t\}$ and $\{s_t(\theta_0) s_t(\theta_0)' - h_t(\theta_0), \mathfrak{F}_t\}$ are martingale difference arrays by construction—see Chapter 6.2 of Hall and Heyde (1980). Phillips and Ploberger (1996) and Li and Müller (2009) make very similar assumptions to (LLLN) and (MDA). Models with asymptotically stochastic information, such as unit root models, are not covered by Condition 1.

Now consider an unstable version of this parametric model, with time-varying parameter $\theta_t = \theta + \delta_t$, $t = 1, \ldots, T$, so that the density of the data $\mathbf{Y}_T$ becomes

$$f_T(\theta, \boldsymbol{\delta}) = \prod_{t=1}^{T} f_{T,t}(\theta + \delta_t), \quad \theta + \delta_t \in \Theta \text{ for } t = 1, \ldots, T \tag{11}$$

where θ and δ_t are $k \times 1$ and $\boldsymbol{\delta} = (\delta_1', \ldots, \delta_T')' \in \mathbb{R}^{Tk}$.[4] Alternative estimators of $\{\theta + \delta_t\}_{t=1}^{T}$, or generally actions, are evaluated via a loss function $L_T : \mathbb{R}^k \times \mathbb{R}^{Tk} \times \mathbb{A}_T \mapsto [0, \overline{L}] \subset \mathbb{R}$, where the action space $\mathbb{A}_T$ is a topological space and L_T is assumed Borel-measurable with respect to the product sigma algebra on $\mathbb{R}^k \times \mathbb{R}^{Tk} \times \mathbb{A}_T$. (For reasons that will become apparent below, loss is also defined for parameter values outside Θ.) The bound $\overline{L}$ is finite and does not depend on T; this assumption of bounded loss greatly facilitates the subsequent analysis. When the true parameter evolution is $\{\theta + \delta_t\}_{t=1}^{T}$ and action $a \in \mathbb{A}_T$ is taken, the incurred loss is $L_T(\theta, \boldsymbol{\delta}, a)$. A typical action could be an estimate of the entire parameter path, so that $\mathbb{A}_T = \Theta^T$, or an estimate of the parameter at a specific point in time, in which case $\mathbb{A}_T = \Theta$. Decisions $\hat{a}$ are measurable functions from the data to $\mathbb{A}_T$. The risk of decision $\hat{a}$ given parameter evolution $\{\theta + \delta_t\}_{t=1}^{T}$ is hence given as $r(\theta, \boldsymbol{\delta}, \hat{a}) = \int L_T(\theta, \boldsymbol{\delta}, \hat{a}) f_T(\theta, \boldsymbol{\delta}) d\mu_T$, which in general depends on $\boldsymbol{\delta}$ and θ.

Let Q_T be a measure on $\mathbb{R}^{Tk}$, and let $w : \Theta \mapsto [0, \infty)$ be a Lebesgue probability density. For each $\theta \in \Theta$, let $\mathcal{V}_T(\theta) = \{\boldsymbol{\delta} : \delta_t + \theta \in \Theta \forall t\} \subseteq \mathbb{R}^{Tk}$. The WAR of decision $\hat{a}$ is then given by

$$WAR(\hat{a}) = \int_\Theta w(\theta) \int_{\mathcal{V}_T(\theta)} r(\theta, \boldsymbol{\delta}, \hat{a}) dQ_T(\boldsymbol{\delta}) d\theta. \tag{12}$$

---

4. In rational expectation models, the presence of time variation in θ potentially affects the model's solution, and thus complicates the derivation of an appropriate likelihood compared to the corresponding model with stable parameters; see Fernandez-Villaverde and Rubio-Ramirez (2007) for one possible computationally intensive approach.

The weighting functions w and Q_T describe the importance attached to alternative true parameter paths in the overall risk calculations; the weight function w attaches different weights to the baseline value θ, and Q_T describes the weight on deviations from this baseline value. In the parameterization $\{\theta_t\}_{t=1}^T = \{\theta + \delta_t\}_{t=1}^T$, the average $T^{-1}\sum\delta_t$ and θ are obviously not uniquely identified. The same WAR criterion may thus be expressed by different choices of w and Q_T. The parameterization is useful because the weighting schemes analysed in this paper assume different asymptotic properties of Q_T and w as follows.

Condition 2 (GS). *The weight function Q_T is the distribution of $\{T^{-1/2}G(t/T)\}_{t=1}^T$, where G is a $k \times 1$ zero mean Gaussian semi-martingale on the unit interval with covariance kernel $E[G(r)G(s)'] = \kappa_G(r,s)$. There exists a finite set of numbers $\tau = \{0, \tau_1, \ldots, \tau_q\} \subset [0,1]$ such that $||\partial^2\kappa_G(r,s)/\partial r\partial s||$ and $||\partial^2\kappa_G(r,s)/\partial r^2||$ are bounded when $r, s \notin \tau$ and $r \neq s$, κ_G admits bounded left and right derivatives with respect to r for all $r = s \in [0,1]\setminus\tau$, and $||\partial\kappa_G(r,s)/\partial r||$ is bounded for $r \in [0,s)\setminus\tau$ and $s \in \tau$.*

(CNT) The weight function w does not depend on T and w is continuous at θ_0.

Under Condition 2 (GS), the weight function Q_T focuses on persistent paths of relatively small variability, because Gaussian processes that satisfy the differentiability assumptions on their kernel are almost surely continuous for all $s \in [0,1]\setminus\tau$ by Kolmogorov's continuity theorem. This concentration on persistent parameter paths drives the derivation of the asymptotic equivalence results below, and it is appealing in many applications, as parameter instability is typically thought of as a low-frequency phenomenon. As discussed in Section 2 above, a popular choice in applied work has been the assumption that parameters vary as a Gaussian random walk, which may be achieved by setting G equal to $G(\cdot) = \Upsilon^{1/2}W(\cdot)$, where W is a $k \times 1$ standard Wiener process. Random walk parameter variability that only occurs in, say, the first half of the sample is achieved by letting $G(s) = \mathbf{1}[s \leq 1/2]\Upsilon^{1/2}W(s) + \mathbf{1}[s > 1/2]\Upsilon^{1/2}W(1/2)$. An assumption of slowly mean reverting parameters can be expressed by letting G be a stationary Ornstein–Uhlenbeck process, more weight on smoother paths by letting G be an integrated Brownian motion $G(s) = \int_0^s W(r)dr$, etc. Condition 2 also accommodates piece-wise constant paths with finitely many jumps, as in the multiple breaks literature, although specification of Q_T requires knowledge of the break dates.

Under Condition 2 (GS), the WAR criterion (12) focuses on parameter paths whose variability is of order of magnitude $T^{-1/2}$. This choice is motivated by a desire to develop procedures that work well when there is relatively little information about the parameter path. For parameter paths of fixed magnitude and persistence, larger samples naturally contain more information, as more adjacent observations can be used to pinpoint the value of the slowly varying parameter at a given date. The sample-size-dependent choice of the magnitude of $\{\delta_t\}$ under Q_T counteracts this effect, making the estimation of the form of the scaled parameter variation $\{T^{1/2}\delta_t\}$ difficult even asymptotically. In this way, the asymptotic arguments derived below based on the sequence of weights as described in Condition 2 (GS) become hopefully relevant to the small sample problem where there is in fact little information about the parameter evolution. At the same time, Condition 2 (CNT) assumes w not to depend on the sample size, reflecting a "global" uncertainty about the baseline level of the time-varying parameter path. With the continuity at θ_0, the weight function becomes asymptotically flat in the $T^{-1/2}$ local neighbourhood around θ_0, so that sample information dominates inference about the baseline value.

The order of magnitude $T^{-1/2}$ for δ_t under Condition 2 (GS) corresponds to the local neighbourhood in which efficient stability tests have non-trivial asymptotic power. The null

hypothesis of a stability test is that the parameter path $\{\theta_t\}_{t=1}^T = \{\theta + \delta_t\}_{t=1}^T$ is constant, i.e.

$$H_0 : \delta_t = 0 \quad \text{for } t = 1, \ldots, T \tag{13}$$

against the alternative that the parameter is time varying. For the development of optimal parameter stability tests, it makes sense to restrict the parameter paths under the alternative such that the difference to the corresponding stable model is the time variability of the path, rather than a different average value of the path. The appropriate restriction is achieved by the multivariate Gaussian measure Q_T^* of $\{T^{-1/2}(G(t/T) - (\sum_{s=1}^T \Gamma(s/T))^{-1} \sum_{s=1}^T \Gamma(s/T)G(s/T))\}_{t=1}^T$. When information accrual is constant, that is $\Gamma(s) = H$ for all $s \in [0, 1]$, then the restriction amounts to a demeaning of δ_t, such that $\sum \delta_t = 0$ a.s. under Q_T^*. In the general case, the restriction forces $\sum \Gamma(t/T)\delta_t = 0$, so that the information weighed parameter path deviations sum to zero, just as in the efficient tests derived by Andrews and Ploberger (1994). Intuitively, a model with time-varying parameter is closest to the stable model with a parameter that is the information-weighted average of the parameter path.

Possibly randomized parameter stability tests φ_T are measurable functions from the data to the interval $[0, 1]$, where $\varphi_T(\mathbf{y}_T)$ indicates the probability of rejecting the null hypothesis of parameter stability when observing $\mathbf{Y}_T = \mathbf{y}_T$. Tests of the same size can then usefully be compared by considering their weighted average power

$$WAP(\varphi_T) = \int_{\mathcal{V}_T(\theta_0)} \int f_T(\theta_0, \boldsymbol{\delta})\varphi_T d\mu_T dQ_T^*(\boldsymbol{\delta}) \tag{14}$$

similar to Andrews and Ploberger (1994). While θ_0 is typically unknown, we show below that there exists a feasible test φ_T^* that asymptotically maximizes this weighted average power.

With the weighting of parameter paths specified as the distribution of a Gaussian process, the problem of finding WAR-minimizing actions essentially becomes a non-linear smoothing exercise. The WAR-minimizing decision is to choose the action a that minimizes

$$\frac{\int_\Theta w(\theta) \int_{\mathcal{V}_T(\theta)} f_T(\theta, \boldsymbol{\delta})L_T(\theta, \boldsymbol{\delta}, a)dQ_T(\boldsymbol{\delta})d\theta}{\int_\Theta w(\theta) \int_{\mathcal{V}_T(\theta)} f_T(\theta, \boldsymbol{\delta})dQ_T(\boldsymbol{\delta})d\theta} \tag{15}$$

for each data realization $\mathbf{Y}_T = \mathbf{y}_T$. With the weighting functions normalized to integrate to unity, this is simply Bayes Rule for minimizing Bayes risk (15), which can be interpreted as finding the action that minimizes the posterior expected loss, i.e. loss integrated with respect to the posterior distributions of $(\theta, \boldsymbol{\delta})$ under a prior for $(\theta, \boldsymbol{\delta})$ that corresponds to the weights in Condition 2.

A large literature has developed around numerically finding exact posterior distributions in non-linear filtering/smoothing problems, often by Monte Carlo simulation techniques, as reviewed in the introduction. This paper complements this research by an asymptotic analysis. First, this yields a deeper theoretical understanding of the link between the estimation testing problems. Second, the asymptotic analysis suggests a computationally simple and asymptotically efficient procedure for choosing the risk-minimizing action. Third, unlike the Bayes Rule computed numerically from (15), the approximately risk-minimizing action can easily be appropriately modified for potentially misspecified models.

Note that Condition 1 makes assumptions about the stable model only, that is, on its behaviour when the parameter path is constant. Clearly, with a focus on the problem of estimating the parameter path, we need to argue for the accuracy of approximations also when the true data-generating process has time-varying parameters. In general, most models with

time-varying parameters generate non-stationary data, to which standard asymptotic results are not easily applicable. In a vector autoregressive regression model, for instance, parameter instabilities lead to highly complicated interactions between the evolution of the lagged variables and the unstable parameters. Our approach is thus to derive asymptotic results for unstable models as an implication of the *contiguity* of models with time-varying parameters of order $T^{-1/2}$ to the corresponding stable model, similar to Andrews and Ploberger (1994), Phillips and Ploberger (1996), Elliott and Müller (2006), and Li and Müller (2009). The following Lemma follows from Lemma 1 of Li and Müller (2009) and the additional discussion in their appendix.

Lemma 1.   *Let $\pi_0 : [0,1] \mapsto \mathbb{R}^k$ be a piece-wise continuous function with at most a finite number of discontinuities and left and right limits everywhere. Under Condition 1, the sequence of densities $\prod_{t=1}^{T} f_{T,t}(\theta_0, T^{-1/2}\pi_0(t/T))$ is contiguous to the sequence $f_T(\theta_0, 0)$. Furthermore, under Conditions 1 and 2, the two sequences of densities $\int_{\mathcal{V}_T(\theta_0)} f_T(\theta_0, \delta) dQ_T(\delta) / \int_{\mathcal{V}_T(\theta_0)} dQ_T(\delta)$ and $\int_{\mathcal{V}_T(\theta_0)} f_T(\theta_0, \delta) dQ_T^*(\delta) / \int_{\mathcal{V}_T(\theta_0)} dQ_T^*(\delta)$ are contiguous to the sequence $f_T(\theta_0, 0)$.*

The main result of the paper is the following Theorem.

Theorem 1.   *Let the sequence of positive definite matrices $\{\tilde{h}_t\}_{t=1}^{T} = \{\tilde{h}_{T,t}\}_{t=1}^{T}$ satisfy*

$$\sup_{\lambda \in [0,1]} \left\| T^{-1} \sum_{t=1}^{[\lambda T]} \tilde{h}_t - \int_0^{\lambda} \Gamma(s) ds \right\| \xrightarrow{p} 0 \tag{16}$$

in the stable model with parameter θ_0.

(i) Consider WAR (12) of alternative decisions $\hat{a}$ under Condition 2. If Condition 1 and (16) hold for almost all θ_0 in the support of w, and for each $\mathbf{Y}_T = \mathbf{y}_T$, the decision $\hat{a}^*$ minimizes WAR with weights as in Condition 2 or a flat weighting of θ and the weight function Q_T on δ in the pseudo model

$$s_t(\hat{\theta}) + \tilde{h}_t \hat{\theta} = \tilde{h}_t(\delta_t + \theta) + v_t, \quad v_t \sim independent \; \mathcal{N}(0, \tilde{h}_t), \quad t = 1, \ldots, T, \tag{17}$$

then for all $\hat{a}$, $\liminf_{T \to \infty}[WAR(\hat{a}) - WAR(\hat{a}^*)] \geq 0$.

(ii) For any $\mathbf{Y}_T = \mathbf{y}_T$, let $\tilde{Q}_T^*$ be the distribution of $\{T^{-1/2}G(t/T) - T^{-1/2}(\sum_{s=1}^{T} \tilde{h}_s)^{-1} \sum_{s=1}^{T} \tilde{h}_s G(s/T)\}_{t=1}^{T}$ (induced by G), and let φ_T^* be a test of (13) of asymptotic level α that maximizes weighted average power with respect to the weighting function $\tilde{Q}_T^*$ in the pseudo model

$$s_t(\hat{\theta}) = \tilde{h}_t \delta_t + v_t, \quad v_t \sim independent \; \mathcal{N}(0, \tilde{h}_t), \quad t = 1, \ldots, T. \tag{18}$$

Then under Conditions 1 and 2, for any other test φ_T of (13) of asymptotic level α, $\liminf_{T \to \infty} [WAP(\varphi_T^*) - WAP(\varphi_T)] \geq 0$.

(iii) Under Condition 1, the total variation distance between the posterior distribution of (θ, δ) in model (11) with priors as in Condition 2 and the posterior distribution of (θ, δ) in the pseudo model (17) with either the same priors or with a flat prior on θ and prior Q_T on δ converges in probability to zero in both the stable model with parameter θ_0 and any unstable model that satisfies the condition of Lemma 1.

Theorem 1 asserts that asymptotically efficient decisions and tests are obtained from combining the sample information from pseudo models (17) and (18), respectively, with the weighting of Condition 2. Since both of these are Gaussian, the resulting distribution can be computed explicitly. Let $\mathbf{e}$ be the $Tk \times k$ matrix $\mathbf{e} = (I_k, \ldots, I_k)'$, $D_{\tilde{h}} = \text{diag}(\tilde{h}_1, \ldots, \tilde{h}_T)$,

$\Sigma_\delta = E_\delta[\delta\delta']$, where E_δ denotes integration of $\delta \sim Q_T$ of Condition 2, $K = \Sigma_\delta(D_{\tilde{h}}\Sigma_\delta + I_{Tk})^{-1}$, $\hat{\mathbf{s}} = (s_1(\hat{\theta})', \ldots, s_T(\hat{\theta})')'$ and

$$\Sigma = K + (I_{Tk} - KD_{\tilde{h}})\mathbf{e}(\mathbf{e}'D_{\tilde{h}}\mathbf{e} - \mathbf{e}'D_{\tilde{h}}KD_{\tilde{h}}\mathbf{e})^{-1}\mathbf{e}'(I_{Tk} - D_{\tilde{h}}K). \tag{19}$$

Note that with $\delta \sim \mathcal{N}(0, \Sigma_\delta)$ and the measurements $X_t = \tilde{h}_t\delta_t + \nu_t$, $\nu_t \sim$ independent $\mathcal{N}(0, \tilde{h}_t)$, $t = 1, \ldots, T$, the distribution of δ conditional on the measurements $\mathbf{X} = (X_1', \ldots, X_T')$ and $D_{\tilde{h}}$ is $\delta|(\mathbf{X}, D_{\tilde{h}}) \sim \mathcal{N}(K\mathbf{X}, K)$. The second term in the definition of Σ results from the uncertainty concerning the baseline value θ. The matrix Σ remains the same if Σ_δ is substituted by the covariance matrix of δ under $\tilde{Q}_T^*$, as defined in Theorem 1 (ii).[5]

Theorem 2. *Let Π be the distribution $\mathcal{N}(\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}, \Sigma)$.*

(i) The decision $\hat{a}^*$ that minimizes expected risk relative to the distribution $\mathbf{e}\theta + \delta \sim \Pi$ for each $\mathbf{Y}_T = \mathbf{y}_T$ minimizes WAR in the pseudo model (17) with a flat weighting on θ.

(ii) A test that rejects for large values of $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}$ is the optimal stability test in the pseudo model (18), and under Conditions 1 and 2

$$\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}} \Rightarrow 2\ln\left(\frac{E_G \exp[\int_0^1 G^*(s)'\Gamma(s)^{1/2}dW^*(s) - \frac{1}{2}\int_0^1 G^*(s)'\Gamma(s)G^*(s)ds]}{E_G \exp[-\frac{1}{2}\int_0^1 G^*(s)'\Gamma(s)G^*(s)ds]}\right)$$

under the null hypothesis, where $G^*(s) = G(s) - (\int_0^1 \Gamma(\lambda)d\lambda)^{-1}\int_0^1 \Gamma(\lambda)G(\lambda)d\lambda$, the standard $k \times 1$ Wiener process W^* is independent of G and E_G denotes integration with respect to the probability measure of G.

(iii) The posterior distribution of $\mathbf{e}\theta + \delta$ under a flat prior on θ in the pseudo model (17) is given by Π.

Comments:

1. Part (i) of Theorem 1 establishes that for arbitrary bounded loss functions, the decision that minimizes WAR in the Gaussian pseudo model (17) is also asymptotically optimal in the true model. As shown in part (i) of Theorem 2, this amounts to finding the risk-minimizing action relative to a multivariate Gaussian distribution for the parameter path. Note that loss may be defined arbitrarily (subject to the bound $\bar{L}$) for parameter values outside Θ, allowing the problem in the pseudo model to be made entirely spherical. For the wide range of bounded bowl-shaped loss functions for which one would choose the posterior mean in a Gaussian model, an asymptotically efficient parameter path estimator is hence given by $\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}$ by Anderson's (1955) Lemma. Note that such loss functions include those that consider a weighted average of symmetric losses incurred by estimation errors in the parameter value, such as

$$L_T(\theta, \delta, a) = \sum_{t=1}^T q_{T,t}L_0(T(\theta + \delta_t - a_t)'W_L(\theta + \delta_t - a_t)) \tag{20}$$

where $a = (a_1', \ldots, a_T')' \in \mathbb{R}^{Tk}$, $\inf_{t \leq T} q_{T,t} \geq 0$, $\sum_{t=1}^T q_{T,t} = 1$, W_L is a non-negative definite $k \times k$ matrix and $L_0 : [0, \infty) \mapsto [0, \bar{L}]$ is a monotonically increasing function with $L_0(0) = 0$. The scaling by T in (20) ensures that the loss does not become trivial as $T \to \infty$ even for good path estimators, although Theorems 1 and 2 remain true without this scaling. This class of loss

---

5. This follows from Theorem 2 (i): combined with the flat weighting on θ, all weighting functions for δ_t that imply the same weighting for $\{\delta_t - T^{-1}\sum_{s=1}^T \delta_s\}_{t=1}^T$ yield the same overall weighting function for $\{\theta + \delta_t\}_{t=1}^T$.

functions (20) contains the special case where one only cares about the parameter at time T, i.e. $q_{T,T} = 1$ and $q_{T,t} = 0$ for all $t < T$, which arises naturally in a forecasting problem.

For more general losses and decision problems, the asymptotically efficient decision can still be obtained by implementing the efficient decision in the Gaussian pseudo model. This typically represents a very substantial computational simplification.

2. Part (ii) of Theorems 1 and 2 spell out the implications of the approximation for efficient tests of the null hypothesis of parameter stability (13). Part (i) of Theorem 2 shows that under symmetric loss, the asymptotically efficient parameter path estimator is $\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}$ with an asymptotic uncertainty described by a zero mean multivariate normal with covariance matrix Σ. The asymptotically efficient test statistic $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}} = (\Sigma\hat{\mathbf{s}})'\Sigma^{+}(\Sigma\hat{\mathbf{s}})$, where Σ^{+} denotes a general inverse, is recognized to be of the usual Wald form: efficient instability tests are based on a quadratic form in the efficient estimator of the instability. Efficient estimation and testing in (potentially) unstable models are hence unified in one framework. This ensures coherence between the stability test and the path estimator, as $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}$ can be large only if the path estimator $\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}$ shows substantial variation.

3. Part (iii) of Theorems 1 and 2 describe the approximation result in Bayesian terms: the posterior distribution of the parameter path $\mathbf{e}\theta + \boldsymbol{\delta}$ comes arbitrarily close to the Tk-dimensional multivariate normal distribution $\mathcal{N}(\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}, \Sigma)$. This is a considerably stronger statement than a convergence in distribution of, say, the posterior of $T^{1/2}\delta_{[\cdot T]}$ viewed as an element of the space of cadlag functions on the unit interval. With $G(s) = 0$, so that $\Sigma_{\delta} = K = 0$, Σ becomes $\Sigma = \mathbf{e}(\mathbf{e}'D_{\tilde{h}}\mathbf{e})^{-1}\mathbf{e}'$, and one recovers the standard result that the posterior distribution of θ converges to $\mathcal{N}(\hat{\theta}, T^{-1}\tilde{H}^{-1})$ where $\tilde{H} = T^{-1}\sum \tilde{h}_t \xrightarrow{p} \int \Gamma(\lambda)d\lambda$, the average information.

In practice, part (iii) of Theorem 1 is useful for Bayesian analyses, as it provides a simple way to compute approximation to the posterior of the unstable parameter path. Even if the exact small sample posterior is required, the approximation of Theorem 1 might be accurate enough for a simple importance sampling algorithm to succeed.

4. The asymptotic distribution of the test statistic $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}$ is provided in Theorem 2 (ii). This distribution is non-standard and depends on the weighting function G and the evolution of the information Γ. Even with Γ known, a simulation based on this expression is quite cumbersome due to the integration over the measure of G. Theorem 2 (ii) is still useful as it shows the existence of an asymptotic distribution. It thus suffices to consider a computationally convenient stable model that has the same asymptotic distribution, such as the stable Gaussian location model $X_t = \tilde{h}_t\theta + Z_t$, $t = 1, \ldots, T$ with Z_t independent and distributed $\mathcal{N}(0, \tilde{h}_t)$. The limiting distribution of $\hat{\mathbf{Z}}'\Sigma\hat{\mathbf{Z}}$ with $\hat{\mathbf{Z}} = (\hat{Z}'_1, \ldots, \hat{Z}'_T)'$ and $\hat{Z}_t = Z_t - \tilde{h}_t(\sum_{s=1}^{T}\tilde{h}_s)^{-1}\sum_{s=1}^{T}\tilde{h}_s Z_s$ is therefore the same as the asymptotic null distribution of $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}$, for $\{\tilde{h}_t\}$ drawn both from the stable model and under all local alternatives for which Lemma 1 implies (16) to also hold.[6] Asymptotically justified critical values of the test statistic $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}$ might hence be obtained by considering the empirical distribution of sufficiently many draws from the distribution of $\hat{\mathbf{Z}}'\Sigma\hat{\mathbf{Z}}$, similar to the approach of Hansen (1996).

5. The approximation results in Theorems 1 and 2 hold for any choice of positive definite sequences $\{\tilde{h}_t\}$ that satisfy (16) in the stable model: in the limit, it is only the average behaviour of $\tilde{h}_t$ that determines the properties of the pseudo models (17) and (18). In particular, with $\Gamma(s) = H$ a constant function, this result allows us to choose $\tilde{h}_t$ time invariant $\tilde{h}_t = \hat{H}$ for any consistent estimator $\hat{H}$ of H, as exploited by the algorithm of Section 2. Without the assumption of a constant Γ, it makes sense to set $\tilde{h}_t$ equal to a standard non-parametric estimator of $\Gamma(s)$,

---

6. Formally, this follows from replacing $\hat{\mathbf{s}}$ by $\hat{\mathbf{Z}}$ in the derivation of the asymptotic null distribution in Theorem 2 (ii).

such as a kernel-smoothed average of $h_t(\hat{\theta})$ as studied in Robinson (1989). Lemma 3 (vi) in the Appendix shows that $\tilde{h}_t = h_t(\hat{\theta})$, and thus for continuous $\Gamma(s)$, also kernel-smoothed averages with bandwidth of order $T^{-1/5}$, satisfy (16) under Condition 1. As explained in Section 2.2, the smoothing extracts the pertinent low-frequency properties of $\Gamma(s)$ without imposing an accurate quadratic approximation of the log-likelihood for each t. Pronounced instabilities can also lead to effectively time-varying $\Gamma(s)$, as in the example of Section 2.3; so choosing $\tilde{h}_t$ time varying in this fashion might improve approximation accuracy even in models where formally $\Gamma(s) = H$.

6. For certain applications, it makes sense to make the scale of the weighting function in the estimation (12) and testing problems (14) a function of the information Γ. In a testing context, for instance, it is often attractive to choose G such that alternatives that are equally difficult to detect receive a similar weight, as in Wald (1943) and, conditional on the break date, in Andrews and Ploberger (1994). Typically, of course, Γ is unknown, and needs to be estimated from the data. Optimal decisions and tests from the pseudo models (17) and (18) with respect to an estimated weighting function generally continue to be asymptotically optimal decisions in terms of (12) and (14), i.e. with respect to the data-independent weighting functions described in Condition 2.

Theorem 3. *Suppose $\{\hat{\Lambda}_{T,t}\}_{t=1}^T$ are non-singular $k \times k$ statistics such that $\sup_{t \leq T} ||\hat{\Lambda}_{T,t} - I_k|| \overset{p}{\to} 0$ and $\sum_{t=2}^T ||\hat{\Lambda}_{T,t} - \hat{\Lambda}_{T,t-1}|| \overset{p}{\to} 0$ in the stable model with parameter θ_0. Then part (ii) of Theorem 1 also holds for $\tilde{Q}_T^*$ replaced by the distribution of $\{T^{-1/2}\hat{\Lambda}_{T,t}G(t/T) - T^{1/2} (\sum_{s=1}^T \tilde{h}_s)^{-1} \sum_{s=1}^T \tilde{h}_s \hat{\Lambda}_{T,s}G(s/T)\}_{t=1}^T$ (induced by G). Furthermore, if $\sup_{\theta \in \Theta, \delta \in \mathbb{R}^{Tk}, a \in \mathbb{A}_T} |L_T(\theta, \text{diag}(\Lambda_{T,1}, \ldots, \Lambda_{T,T})\delta, a) - L_T(\theta, \delta, a)| \to 0$ for all sequences $\{\Lambda_{T,t}\}_{t=1}^T$ satisfying $\sup_{t \leq T} ||\Lambda_{T,t} - I_k|| \to 0$ and $\sum_{t=2}^T ||\Lambda_{T,t} - \Lambda_{T,t-1}|| \to 0$ as $T \to \infty$, then also part (i) of Theorem 1 holds for Q_T replaced by the distribution of $\{T^{-1/2}\hat{\Lambda}_{T,t}G(t/T)\}_{t=1}^T$ (induced by G).*

In a typical application of Theorem 3, suppose one aims at computing the asymptotically efficient test for a Condition 2 weighting function with $G(\cdot) = c\overline{\Gamma}^{-1/2}W(\cdot)$, where c is a known scalar constant, but the average information $\overline{\Gamma} = \int_0^1 \Gamma(\lambda)d\lambda$ is not known. Then Theorem 3 shows that this test may be computed from the pseudo model (18) with an estimated weighting function that corresponds to the distribution of $c\widehat{\overline{\Gamma}}^{-1/2}W(\cdot) = c\widehat{\overline{\Gamma}}^{-1/2}\overline{\Gamma}^{1/2}G(\cdot)$, i.e. based on the statistic $\hat{s}'\Sigma\hat{s}$ where Σ_δ in the definition (19) of Σ has i, j-th $k \times k$ block equal to $T^{-2}c^2 \sum_{t=1}^{\min(i,j)} \widehat{\overline{\Gamma}}^{-1}$, as long as $\widehat{\overline{\Gamma}} \overset{p}{\to} \overline{\Gamma}$ under θ_0 stable. In the more general case where $G(\cdot) = R(\cdot)G_0(\cdot)$ with G_0 a known Gaussian process and $R : [0, 1] \mapsto \mathbb{R}^{k \times k}$ an unknown fixed and nonsingular matrix function, Theorem 3 requires beyond consistency that the scaled estimation error $\hat{\Lambda}_{T,t} = \hat{R}_{T,t}R(t/T)^{-1}$ is smooth by imposing $\sum_{t=2}^T ||\hat{\Lambda}_{T,t} - \hat{\Lambda}_{T,t-1}|| \overset{p}{\to} 0$. This condition is typically satisfied for parametric estimators of R when R is of bounded variation, such as, for example, when R is a linear trend of unknown slope or when R is a step function with known step locations.

Moreover, optimal decisions from the pseudo model typically retain their WAR (12) optimality under such estimated weights, such as the path estimator $\mathbf{e}\hat{\theta} + \Sigma\hat{s}$ under the class of loss functions (20) when L_0 is continuous. The restriction of the loss functions in the second claim of Theorem 3 is necessary to rule out a somewhat pathological focus of L_T on the scale of the weighting function for δ.[7]

---

7. For example, with $G(s) = W(s)$ and $\Lambda_{T,t} = (1 + T^{-1/4})I_k$, $L_T(\theta, \delta, a) = (T^{1/2}\text{tr}(T \sum (\Delta \delta_t)(\Delta \delta_t)' - I_k))^2 \wedge 1$, $\lim_{T \to \infty} E_\delta L_T(\theta, \delta, a) \neq \lim_{T \to \infty} E_\delta L_T(\theta, (1 + T^{-1/4})\delta, a)$.

7. For some purposes, it makes sense to consider weighting functions that are more agnostic about the magnitude and/or form of the parameter instability than is possible under Condition 2. One way to achieve this without foregoing the computational advantages of a Gaussian weighting function is to consider weighting functions (or priors) for δ that are a weighted average of distributions of different Gaussian processes. The following Theorem shows how parts (i) and (iii) of Theorems 1 and 2 need to be adapted in the case of such a finite mixture.

Theorem 4.   *Let G_i, $i = 1, \ldots, n_G$ be processes satisfying Condition 2 (GS). If Q_T is the distribution of the mixture of $\{T^{-1/2} G_i(t/T)\}$ with mixing probabilities p_i, then parts (i) and (iii) of Theorems 1 and 2 hold with Π replaced by the mixture of n_G multivariate normal distributions $\mathcal{N}(\mathbf{e}\hat{\theta} + \Sigma_i \hat{\mathbf{s}}, \Sigma_i)$ with mixing probabilities proportional to*

$$\tilde{w}_i = p_i |D_{\tilde{h}} \Sigma_{\delta(i)} + I_{Tk}|^{-1/2} |\mathbf{e}' D_{\tilde{h}} \mathbf{e} - \mathbf{e}' D_{\tilde{h}} K_i D_{\tilde{h}} \mathbf{e}|^{-1/2} \exp[\tfrac{1}{2} \hat{\mathbf{s}}' \Sigma_i \hat{\mathbf{s}}], \quad i = 1, \ldots, n_G, \qquad (21)$$

where K_i, $\Sigma_{\delta(i)}$, and Σ_i are defined as K, Σ_δ, and Σ in (19) with Σ_δ replaced by $\Sigma_{\delta(i)}$, the covariance matrix of $T^{-1/2}(G_i(1/T)', G_i(2/T)', \ldots, G_i(1)')'$ for $i = 1, \ldots, n_G$.

Theorem 4 is a simple consequence of the fact that the Gaussian pseudo model (17) remains an accurate approximation of the sample information for each of the n_G weighting functions, such that the likelihood ratios can be explicitly computed. The WAR-minimizing parameter path estimator under mixture weightings generally depends much more on the loss function than in the single Gaussian process case, as mixtures of normal distributions are not generally symmetric around their mean. Under truncated quadratic loss (20) with $L_0(x) = \min(x, \overline{L})$, the WAR-minimizing path estimator converges to $\sum_{i=1}^{n_G} \tilde{w}_i \Sigma_i \hat{\mathbf{s}} / \sum_{i=1}^{n_G} \tilde{w}_i$ as $\overline{L} \to \infty$.

8. So far we assumed that the model in Condition 1 is correctly specified. As demonstrated by Huber (1967) and White (1982), in misspecified models maximum likelihood estimators are consistent for the pseudo-true parameter value that minimizes the Kullback–Leibler divergence of the true model from the maintained model. This pseudo-true parameter value sometimes remains the natural object of interest as, for instance, in exponential models with correctly specified mean (cf. Gourieroux, Monfort, and Trognon, 1984). We now discuss a modification of the pseudo model (17) such that, under some conditions, the best decision about the pseudo-true parameter path in a misspecified model yields the same asymptotic risk as the best decision in a corresponding correctly specified model.

Suppose the evolution of the pseudo-true parameter through time is $\theta_t = \theta_0 + T^{-1/2} \pi_0(t/T)$, and let $\hat{s}_t = s_t(\hat{\theta})$ and $\hat{h}_t = h_t(\hat{\theta})$ be the scores and Hessians in the misspecified model. Under standard primitive assumptions, by the usual Taylor approximations, one would typically find that

$$T^{-1/2} \sum_{t=1}^{[\cdot T]} \hat{s}_t \Rightarrow J(\cdot) - \int_0^{\cdot} \Xi(\lambda) d\lambda \left(\int_0^1 \Xi(\lambda) d\lambda \right)^{-1} J(1),$$

$$T^{1/2}(\hat{\theta} - \theta_0) \Rightarrow \left(\int_0^1 \Xi(\lambda) d\lambda \right)^{-1} J(1)$$

and

$$T^{-1} \sum_{t=1}^{[\cdot T]} \hat{h}_t \xrightarrow{p} \int_0^{\cdot} \Xi(\lambda) d\lambda,$$

where

$$J(s) = \int_0^s V(\lambda)^{1/2} dW(\lambda) + \int_0^s \Xi(\lambda)\pi_0(\lambda)d\lambda \tag{22}$$

with $V : [0,1] \mapsto \mathbb{R}^{k \times k}$ and $\Xi : [0,1] \mapsto \mathbb{R}^{k \times k}$ positive definite, non-stochastic matrix functions. See, for instance, Andrews (1993) and Li and Müller (2009) for primitive conditions that induce such convergences. The matrix $V(s)$ is the average long-run variance of the scores $s_t(\theta_0)$ at the time $t = [sT]$, which in a misspecified model is not in general equal to the average of the Hessians $\Xi(s)$. The pseudo model (17) based on $\hat{s}_t$, $\hat{\theta}$ and $\hat{h}_t$ directly thus behaves differently than the pseudo model of any correctly specified model, even asymptotically. Note, however, that if we premultiply the scores and Hessians by $\Xi(t/T)V(t/T)^{-1}$, we obtain a long-run variance for $\Xi(t/T)V(t/T)^{-1}s_t(\theta_0)$ of $\Xi(s)V(s)^{-1}\Xi(s)$ at time $t = [sT]$, which coincides with the local average of $\Xi(t/T)V(t/T)^{-1}\hat{h}_t$. This adjustment is the time-varying parameter analogue to the sandwich pseudo model suggested in Müller (2009) for Bayesian inference in stable misspecified models.

Condition 3. *In the misspecified model with parameter path equal to $\theta_t = \theta_0 + T^{-1/2}\pi_0(t/T)$, there exist sequences of invertible $k \times k$ matrices $\{\hat{\Xi}_t\}_{t=1}^T$ and $\{\hat{V}_t\}_{t=1}^T$ such that with $\hat{s}_t^r = \hat{\Xi}_t \hat{V}_t^{-1} \hat{s}_t - T^{-1} \sum_{s=1}^T \hat{\Xi}_s \hat{V}_s^{-1} \hat{s}_s$ and $\hat{\theta}^r = \hat{\theta} + (\sum_{t=1}^T \hat{\Xi}_t \hat{V}_t^{-1} \hat{h}_t)^{-1} \sum_{t=1}^T \hat{\Xi}_t \hat{V}_t^{-1} \hat{s}_t$, we have*

$$T^{-1/2} \sum_{t=1}^{[\cdot T]} \hat{s}_t^r \Rightarrow J^r(\cdot) - \int_0^{\cdot} \Gamma(\lambda)d\lambda \left(\int_0^1 \Gamma(\lambda)d\lambda \right)^{-1} J^r(1)$$

and

$$T^{1/2}(\hat{\theta}^r - \theta^0) \Rightarrow \left(\int_0^1 \Gamma(\lambda)d\lambda \right)^{-1} J^r(1),$$

where

$$J^r(s) = \int_0^s \Gamma(\lambda)^{1/2} dW(\lambda) + \int_0^s \Gamma(\lambda)\pi_0(\lambda)d\lambda$$

and $\Gamma(s) = \Xi(s)V(s)^{-1}\Xi(s)$. Further, there exist matrices $\{\tilde{h}_t^r\}_{t=1}^T$ such that

$$\sup_{\lambda \in [0,1]} \left\| T^{-1} \sum_{t=1}^{[\lambda T]} \tilde{h}_t^r - \int_0^{\lambda} \Gamma(s)ds \right\| \xrightarrow{p} 0.$$

The "robustified" estimator $\hat{\theta}^r$ and partial sums of the scores $\hat{s}_t^r$ and Hessians $\tilde{h}_t^r$ of Condition 3 behave just like the maximum likelihood estimator and partial sums of the scores and Hessians would in a correctly specified model with average Fisher Information at time $t = [sT]$ equal to $\Gamma(s) = \Xi(s)V(s)^{-1}\Xi(s)$. This suggests that asymptotically, best decisions in the robustified pseudo model

$$\hat{s}_t^r + \tilde{h}_t^r \hat{\theta}^r = \tilde{h}_t^r (\delta_t + \theta) + \nu_t, \quad \nu_t \sim \text{independent } \mathcal{N}(0, \tilde{h}_t^r), \quad t = 1, \ldots, T, \tag{23}$$

have the same risk as best decisions in a correctly specified model with this Fisher Information. At the same time, if the model ends up being correctly specified, then $\Xi(s) = V(s)$ by the information equality, and $\hat{\theta}^r$ and partial sums of $\hat{s}_t^r$ and $\tilde{h}_t^r$ have the same asymptotic properties as the original $\hat{\theta}$, $\hat{s}_t$ and $\hat{h}_t$.

In general, for Condition 3 to follow from the weak convergences mentioned above, $\hat{\Xi}_t$ and $\hat{V}_t$ must be sufficiently accurate estimators of $\Xi(t/T)$ and $V(t/T)$. Clearly, the construction of appropriate estimators is the more difficult the less is known about the variability of $\Xi(\cdot)$ and $V(\cdot)$. In the important special case of constant $\Xi(s) = H$ and $V(\cdot) = V$, it suffices to set $\hat{\Xi}_t = \hat{H}$ and $\hat{V}_t = \hat{V}$ for any consistent estimators $(\hat{H}, \hat{V})$ of (H, V) (and in this case, $\tilde{\theta}^r = \hat{\theta}$ and one possible choice for $\tilde{h}_t^r$ is $\tilde{h}_t^r = \hat{S}^{-1} = \hat{H}\hat{V}^{-1}\hat{H}$). Typically, the estimators $\hat{H} = T^{-1}\sum_{t=1}^{T}\hat{h}_t$ and, as long as $s_t(\theta_0)$ is serially uncorrelated in the stable misspecified model, $\hat{V} = T^{-1}\sum_{t=1}^{T}\hat{s}_t\hat{s}_t'$ are consistent, whereas if the misspecification leads to potentially autocorrelated $s_t(\theta_0)$, one needs to apply a standard long-run variance estimator to the scores $\{\hat{s}_t\}_{t=1}^{T}$. See Li and Müller (2009) for a discussion of possible primitive conditions for Condition 3.

Theorem 5. *Consider a correctly specified model satisfying Condition 1 and parameter path equal to $\theta_t = \theta_0 + T^{-1/2}\pi_0(t/T)$, where π_0 satisfies the conditions of Lemma 1, and let $\hat{a}^*$ be the decision that, for each draw, minimizes expected risk relative to the distribution of $\Pi \sim \mathcal{N}(\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}, \Sigma)$. Similarly, consider a potentially misspecified model satisfying Condition 3, and let $\hat{a}^{r*}$ be the decision that, for each draw, minimizes expected risk relative to the distribution of $\Pi^r \sim \mathcal{N}(\mathbf{e}\hat{\theta}^r + \Sigma^r\hat{\mathbf{s}}^r, \Sigma^r)$, where $\hat{\mathbf{s}}^r$ is the $Tk \times 1$ vector of stacked scores $\{\hat{s}_t^r\}_{t=1}^{T}$, and Σ^r is constructed as Σ in (19) with $\tilde{h}_t$ replaced by $\tilde{h}_t^r$.*

(i) Let $a^*(\Pi_T)$ be the action that minimizes expected risk relative to the $Tk \times 1$ multivariate distribution Π_T, and define $\overline{\boldsymbol{\pi}}_0 = (\pi_0(1/T)', \ldots, \pi_0(T/T)')'$. If the common loss function L_T is such that

$$L_T(\theta_0, \overline{\boldsymbol{\pi}}_0, a^*(\Pi_{1T})) - L_T(\theta_0, \overline{\boldsymbol{\pi}}_0, a^*(\Pi_{2T})) \to 0 \qquad (24)$$

whenever the total variation distance between the two $Tk \times 1$ normal distributions Π_{1T} and Π_{2T} converges to zero, then the difference between the sampling distributions of $L_T(\theta_0, \overline{\boldsymbol{\pi}}_0, \hat{a}^*)$ in the correctly specified model and $L_T(\theta_0, \overline{\boldsymbol{\pi}}_0, \hat{a}^{r*})$ in the potentially misspecified model converges to zero in the Prohorov metric.

(ii) If (24) is strengthened to hold whenever the total variation distance between the two mixtures of n_G normal distributions Π_{1T} and Π_{2T} converges to zero, then the conclusion of part (i) also applies with Π replaced by the mixture distribution of Theorem 4, and Π^r defined as the analogous mixture based on $\hat{\theta}^r$, $\hat{s}_t^r$ and $\tilde{h}_t^r$ in the place of $\hat{\theta}$, $s_t(\hat{\theta})$ and $\tilde{h}_t$.

(iii) The test statistics $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}$ and $\hat{\mathbf{s}}^{r'}\Sigma^r\hat{\mathbf{s}}^r$ have the same asymptotic distribution.

Theorem 5 formalizes the link between best inference based on the pseudo model in a misspecified model using the robustified statistics of Condition 3, and best inference based on the pseudo model in a corresponding correctly specified model. The key assumption (24) for part (i) is that similar multivariate normal distributions Π_{1T} and Π_{2T} induce best actions of similar loss. This holds, for instance, for the loss function (20) as long as L_0 is continuous. Under this assumption, the sampling distribution of the losses in these two models is asymptotically identical, which—given that the loss is assumed bounded—also implies identical (frequentist) risk of the best decisions $\hat{a}^*$ and $\hat{a}^{r*}$. Thus, from a decision theoretic perspective, ignoring the misspecification is not costly in the sense that one still obtains inference of the same asymptotic quality as in the corresponding correctly specified model. Similarly, part (iii) shows that the test statistic $\hat{\mathbf{s}}^{r'}\Sigma^r\hat{\mathbf{s}}^r$ has the same asymptotic distribution as $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}$ does in the corresponding correctly specified model, and thus the same local power. In particular, if Condition 3 holds for $\pi_0(\cdot) = 0$, one may simulate the null distribution by many draws of $\hat{\mathbf{Z}}^{r'}\Sigma^r\hat{\mathbf{Z}}^r$ with $\hat{\mathbf{Z}}^r = (\hat{Z}_1^{r'}, \ldots, \hat{Z}_T^{r'})'$, $\hat{Z}_t^r = Z_t^r - \tilde{h}_t^r(\sum_{s=1}^{T}\tilde{h}_s^r)^{-1}\sum_{s=1}^{T}\tilde{h}_s^r Z_s^r$ and Z_t independent $\mathcal{N}(0, \tilde{h}_t^r)$ pseudo random variables, as discussed in Comment 4.

It is easy to see that these correspondences do not hold in general without the robustification detailed in Condition 3 and, in analogy to the formal results in Müller (2009), one would expect that asymptotic risk is generally smaller in the robustified pseudo model. The reason is that under misspecification, both the original likelihood and uncorrected pseudo model (17) convey a misleading account of the sample information about the pseudo-true parameter, as the variance of partial sums of $\hat{s}_t$ is different from the partial sum behaviour of the Hessians $\hat{h}_t$. Whenever the optimal action depends on the variance Σ^r, one therefore should obtain lower asymptotic risk decisions from the robustified model. What is more, if the weighting function (or prior) averages over alternative Gaussian processes as in Theorem 4, then a Bayesian or uncorrected pseudo model (17) will in general lead to biased estimation of the magnitude of the parameter time variation, as the variability of $\hat{s}_t$ is mistakenly judged relative to $\tilde{h}_t$. Inference based on the robustified pseudo model (23) is thus not only more convenient computationally compared to a fully fledged Bayesian analysis, but also provides more reliable inference about the pseudo-true parameter path in misspecified models.

9. Much applied work is based on the special case where the prior or weighting function of a time-varying parameter is a Gaussian random walk, such that $G(\cdot) = \Upsilon^{1/2} W(\cdot)$ for some positive semidefinite matrix Υ and standard Wiener process W; see the citations in Section 2. The Markovian structure of the Wiener process enables the application of an iterative Kalman smoothing algorithm. We provide such an algorithm in the Appendix, which also takes care of the impact of the flat weighting of θ in the smoothing, along similar lines as Jong (1991), and enables computation of the statistics appearing in Theorems 2, 4, and 5 without matrix computations of dimension $Tk \times Tk$.

The algorithm described in Section 2 of this paper exploits the additional computational simplifications when $\tilde{h}_t^r = \hat{S}^{-1}$, $t = 1, \ldots, T$ and $\Upsilon = c^2 \hat{S}_\beta$, where $\hat{S}_\beta$ is the upper left $p \times p$ block of $\hat{S}$.[8] By Theorem 3, this choice of Υ minimizes asymptotic WAR for the weighting function induced by $\Upsilon = c^2 H_\beta^{-1}$ for most loss functions in a correctly specified model with $\Gamma(s) = H$ as long as $\hat{S}_\beta \xrightarrow{p} H_\beta^{-1}$, where H_β^{-1} is the upper left $p \times p$ block of H^{-1}.

10. A number of previous papers have considered parameter stability tests against random walk-type alternatives: Nyblom (1989) derives locally best tests against general martingale variability in the parameters for general likelihood models, Shively (1988) considers small sample tests in a linear regression model, and Elliott and Müller (2006) derive asymptotic results for point optimal parameter instability tests in linear regression models for a class of weighting functions that includes the Gaussian random walk case. The contribution of Theorems 1, 2, and 5 with respect to this literature is the generalization of the point optimal tests to general, potentially misspecified likelihood models, including non-stationary models with, say, a time trend. Under the assumption of correct misspecification, the degree of generality of the results here concerning parameter stability tests is similar to those of Andrews and Ploberger (1994), but the focus there is on parameters that shift at unknown dates, which leads Andrews and Ploberger (1994) to consider weighting functions that are a continuous mixture of piece-wise constant parameter paths with Gaussian shifts.

Elliott and Müller (2006) show that efficient tests for a Gaussian random walk in the parameters and efficient tests for a single break at unknown date have asymptotic power that is roughly comparable no matter what the true alternative is; the efficient tests for the Gaussian random walk have the advantage that they avoid the need for trimming the break dates

---

8. The algorithm in Section 2.2 stems from combining our results with those of Elliott and Müller (2006): applying the matrix identity (A5) in the Appendix, Σ^r becomes $(I_T - G_c) \otimes \hat{S}_\beta$ with G_c defined in Elliott and Müller (2006), and the expressions for $qLL(c) = -\hat{s}^{r\prime} \Sigma^r \hat{s}^r$, $\tilde{w}_i$ and κ_t in Section 2.2 follow.

away from the beginning and end of the sample, and their computational convenience, at least compared to efficient tests for more than one potential break.

4. CONCLUSIONS

Most economic relationships are potentially unstable over time. In empirical work, this translates into time-varying parameters of estimated models. It is often of interest to keep track of this potential instability. Going beyond time variation in the coefficients of Gaussian linear regression models, however, typically leads to substantial numerical and computational complications.

This paper considers a general likelihood model and focuses on parameter instabilities of a magnitude that are non-trivial to detect, which seems a relevant part of the parameter space for many instabilities economists care about. The main contribution is an asymptotically justified approximation to the sample information about the time-varying parameter, so that under a Gaussian weighting, WAR-minimizing path estimators and weighted average power maximizing parameter stability tests become straightforward to compute. In addition to this computational advantage, an appropriately robustified version of the approximating model yields decisions of the same asymptotic risk as in a corresponding correctly specified model.

APPENDIX

A.1. *Iterative formulas for the path estimator and related statistics when* $G(\cdot) = \Upsilon^{1/2} W(\cdot)$:

For notational convenience, we describe the algorithm for the pseudo models (17) and (18). For statistics based on the robustified pseudo model (23), replace $(\hat{\theta}, s_t(\hat{\theta}), \tilde{h}_t, \hat{Z}_t)$ by $(\hat{\theta}^r, \hat{s}_t^r, \tilde{h}_t^r, \hat{Z}_t^r)$ throughout.

With $\hat{s}_t = s_t(\hat{\theta})$, compute

$$\hat{a}_t = \hat{a}_{t-1} + P_{t-1}(\tilde{h}_t P_{t-1} + I_k)^{-1}(\hat{s}_t - \tilde{h}_t \hat{a}_{t-1})$$

$$\hat{A}_t = \hat{A}_{t-1} + P_{t-1}(\tilde{h}_t P_{t-1} + I_k)^{-1}(\tilde{h}_t - \tilde{h}_t \hat{A}_{t-1})$$

$$P_t = P_{t-1} + T^{-2}\Upsilon - P_{t-1}(\tilde{h}_t P_{t-1} + I_k)^{-1}\tilde{h}_t P_{t-1}$$

for $t = 1, \ldots, T$ with $\hat{a}_0 = 0$, $\hat{A}_0 = 0$ and $P_0 = T^{-2}\Upsilon$. Further, compute

$$\hat{b}_t = \hat{a}_t + (I_k - T^{-2}\Upsilon P_t^{-1})(\hat{b}_{t+1} - \hat{a}_t)$$

$$\hat{B}_t = \hat{A}_t + (I_k - T^{-2}\Upsilon P_t^{-1})(\hat{B}_{t+1} - \hat{A}_t)$$

$$R_t = P_t - \Upsilon + (I_k - T^{-2}\Upsilon P_t^{-1})(R_{t+1} - P_t)(I_k - T^{-2}\Upsilon P_t^{-1})'$$

for $t = T - 1, \ldots, 1$ with $\hat{b}_T = \hat{a}_T$, $\hat{B}_T = \hat{A}_T$ and $R_T = P_T - T^{-2}\Upsilon$. Let $\hat{d} = \left(\sum_{t=1}^T \tilde{h}_t(I_k - \hat{B}_t)\right)^{-1} \sum_{t=1}^T (\hat{s}_t - \tilde{h}_t \hat{b}_t)$. The tth $k \times 1$ block of $\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}$ is then given by $\hat{\theta} + \hat{b}_t + (I_k - \hat{B}_t)\hat{d}$, and the t, tth $k \times k$ block of Σ is given by $R_t + (I_k - \hat{B}_t)\left(\sum_s \tilde{h}_s(I_k - \hat{B}_s)\right)^{-1}(I_k - \hat{B}_t')$. Also, $\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}} = \sum_{t=1}^T \hat{s}_t'\hat{b}_t + \left(\sum_{t=1}^T (\hat{s}_t - \tilde{h}_t \hat{b}_t)\right)'\hat{d}$, $|D_{\tilde{h}}\Sigma_\delta + I_{Tk}| = \prod_{t=1}^T |\tilde{h}_t P_{t-1} + I_k|$ and $|\mathbf{e}'D_{\tilde{h}}\mathbf{e} - \mathbf{e}'D_{\tilde{h}}KD_{\tilde{h}}\mathbf{e}| = |\sum_{t=1}^T \tilde{h}_t(I_k - \hat{B}_t)|$. To compute $\hat{\mathbf{Z}}'\Sigma\hat{\mathbf{Z}}$, replace $\hat{s}_t$ by $\hat{Z}_t$ throughout.

To generate a draw from $\mathcal{N}(\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}, \Sigma)$, one may proceed as follows. Draw $\tilde{b}_T \sim \mathcal{N}(\hat{a}_T, P_T - T^{-2}\Upsilon)$, and then draw iteratively for $t = T - 1, \ldots, 1$

$$\tilde{b}_t \sim \mathcal{N}(\tilde{b}_{t+1} - T^{-2}\Upsilon P_t^{-1}(\tilde{b}_{t+1} - \hat{a}_t), T^{-2}\Upsilon - T^{-4}\Upsilon P_t^{-1}\Upsilon).$$

Draw $\tilde{d} \sim \mathcal{N}(\hat{d}, \left(\sum_{t=1}^T \tilde{h}_t(I_k - \hat{B}_t)\right)^{-1})$ independent of $\{\tilde{b}_t\}_{t=1}^T$. Then $\{\hat{\theta} + \tilde{b}_t + (I_k - \hat{B}_t)\tilde{d}\}_{t=1}^T$ constitutes a draw from $\mathcal{N}(\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}, \Sigma)$.

If Υ is singular, then P_t^{-1} is to be replaced by the Moore–Penrose generalized inverse of P_t.

A.2. *Proofs*

A.2.1. Notation.   For notational ease, extend the domain of f_T by letting $f_T(\theta) = 0$ for $\theta \notin \Theta$, and let $s_t(\theta) = 0$ for $\theta \notin \Theta_0$, $t = 1, \ldots, T$.

The following notation is used in the following Lemmas and proofs:

- the $Tk \times k$ vector $\mathbf{e} = (I_k, \ldots, I_k)'$
- the $k \times k$ matrices $\Gamma_t = \Gamma(t/T)$, $\tilde{H} = T^{-1} \sum \tilde{h}_t$ and $\hat{\Gamma} = T^{-1} \sum \Gamma_t$
- the $Tk \times Tk$ matrices $D_\Gamma = \operatorname{diag}(\Gamma_1, \ldots, \Gamma_T)$, $D_{\tilde{h}} = \operatorname{diag}(\tilde{h}_1, \ldots, \tilde{h}_T)$ and $F = T^{-1/2}F_0 \otimes I_k$, where F_0 is a $T \times T$ matrix with zeros above the main diagonal and 1's elsewhere
- the $k \times 1$ vectors $u = T^{1/2}(\theta - \theta_0)$, $\hat{u} = T^{1/2}(\hat{\theta} - \theta_0)$, $\hat{s}_t = s_t(\hat{\theta})$, $t = 1, \ldots, T$ and $\overline{\delta} = \hat{\Gamma}^{-1} T^{-1} \sum_{t=1}^{T} \Gamma_t \delta_t$
- the $Tk \times 1$ vectors $\hat{\mathbf{s}} = (\hat{s}_1', \ldots \hat{s}_T')'$ and $\mathbf{s}_0 = (s_1(\theta_0)', \ldots, s_1(\theta_0)')'$
- the indicator functions $\mathcal{S}_T(\boldsymbol{\delta}) = \mathbf{1}[T^{1/2} \sup_{t \leq T} ||\delta_t|| < T^\eta]$, where η is defined in Condition 1 (ID) and we assume $\eta < 1/2$ without loss of generality and $\mathcal{A}_T(u) = \mathbf{1}[||u|| < \alpha_T]$ with $\alpha_T \to \infty$ defined in Lemma 4 below
- the real valued functions $LR_T(u, \boldsymbol{\delta}) = \frac{f_T(\theta_0 + T^{-1/2}u, \boldsymbol{\delta})}{f_T(\theta_0, 0)}$, $\widehat{LR}_T(u, \boldsymbol{\delta}) = \exp[\sum \hat{s}_t'\delta_t - \frac{1}{2}\sum \delta_t'\tilde{h}_t\delta_t + T^{-1/2}(\hat{u} - u)' \sum \tilde{h}_t\delta_t - \frac{1}{2}u'\tilde{H}u + \hat{u}'\tilde{H}u]$ and $\overline{LR}_T(\boldsymbol{\delta}) = \exp[\sum \hat{s}_t'\delta_t - \frac{1}{2}\sum \delta_t'\tilde{h}_t\delta_t + \frac{1}{2}(T^{-1/2}\sum \delta_t'\tilde{h}_t)\tilde{H}^{-1}T^{-1/2}\sum \tilde{h}_t\delta_t]$
- the scalars $m_T = \int E_\delta w(\theta_0 + T^{-1/2}u)LR_T(u, \boldsymbol{\delta})du$, $\hat{m}_T = w(\theta_0) \int E_\delta \widehat{LR}_T(u, \boldsymbol{\delta})du$ and $M_T = E_\delta \prod_{t=1}^{T} \mathbf{1}[(\theta_0 + \delta_t) \in \Theta]$

A.2.2. Proofs of theorems in the main text.   The general strategy for the proof of Theorem 1 is as follows. Given Lemma 1, it suffices to prove convergences in probability for data generated under the stable model. All following probability calculations are thus made under the stable Condition 1 model, if not explicitly noted otherwise. We first establish part (iii) of Theorem 1, from which part (i) follows relatively easily. The main thrust of the proof of part (iii) is the argument that $\int E_\delta \left| w(\theta_0 + T^{-1/2}u)LR_T(u, \boldsymbol{\delta}) - w(\theta_0)\widehat{LR}_T(u, \boldsymbol{\delta}) \right| du$ converges in probability to zero. Lemma 4 (i) below shows that replacing $LR_T(u, \boldsymbol{\delta})$ by $\mathcal{S}_T(\boldsymbol{\delta})\mathcal{A}_T(u)LR_T(u, \boldsymbol{\delta})$ in this expression induces a negligible approximation error. The "main" approximation via Taylor series expansions is performed in Lemma 2, whose statement and proof is below. The proofs of the additional Lemmas of Section A.2.3 may be found in the Supplementary Materials.

Proof of Theorem 1.   (iii) We focus on the claim for a flat weighting on θ; the claim for a weighting w on θ follows very similarly.

Let $\hat{f}_T(\theta, \boldsymbol{\delta})$ be the density of the observations in the pseudo model (17), so that $\widehat{LR}_T(u, \boldsymbol{\delta}) = \hat{f}_T(\theta_0 + T^{-1/2}u, \boldsymbol{\delta})/\hat{f}_T(\theta_0, 0)$. The total variation distance between the posterior distributions computed from the true model density f_T and the pseudo model density $\hat{f}_T$ is then given by

$$\int E_\delta \left| \frac{w(\theta_0 + T^{-1/2}u)LR_T(u, \boldsymbol{\delta})}{m_T} - \frac{w(\theta_0)\widehat{LR}_T(u, \boldsymbol{\delta})}{\hat{m}_T} \right| du$$

$$\leq \hat{m}_T^{-1} \int E_\delta \left| w(\theta_0 + T^{-1/2}u)LR_T(u, \boldsymbol{\delta}) - w(\theta_0)\widehat{LR}_T(u, \boldsymbol{\delta}) \right| du + \hat{m}_T^{-1}|m_T - \hat{m}_T|$$

where $m_T = \int E_\delta w(\theta_0 + T^{-1/2}u)LR_T(u, \boldsymbol{\delta})du > 0$ a.s. and $\hat{m}_T = w(\theta_0) \int E_\delta \widehat{LR}_T(u, \boldsymbol{\delta})du > 0$ a.s. Since

$$|\hat{m}_T - m_T| \leq \int E_\delta \left| w(\theta_0 + T^{-1/2}u)LR_T(u, \boldsymbol{\delta}) - w(\theta_0)\widehat{LR}_T(u, \boldsymbol{\delta}) \right| du \tag{A1}$$

it suffices to show that $\int E_\delta \left| w(\theta_0 + T^{-1/2}u)LR_T(u, \boldsymbol{\delta}) - w(\theta_0)\widehat{LR}_T(u, \boldsymbol{\delta}) \right| du \xrightarrow{p} 0$ and $\hat{m}_T^{-1} = O_p(1)$.

Now by Fubini's theorem and a direct calculation,

$$\int E_\delta \widehat{LR}_T(u, \boldsymbol{\delta})du = \exp\left[\frac{1}{2}\hat{u}'\tilde{H}\hat{u}\right] E_\delta \int \exp\left[\hat{\mathbf{s}}'\boldsymbol{\delta} - \frac{1}{2}\boldsymbol{\delta}'D_{\tilde{h}}\boldsymbol{\delta} + T^{-1/2}(\hat{u} - u)'\mathbf{e}'D_{\tilde{h}}\boldsymbol{\delta} - \frac{1}{2}(u - \hat{u})'\tilde{H}(u - \hat{u})\right] du$$

$$= (2\pi)^{k/2}|\tilde{H}|^{-1/2}\exp\left[\frac{1}{2}\hat{u}'\tilde{H}\hat{u}\right] E_\delta \overline{LR}_T(\boldsymbol{\delta}). \tag{A2}$$

Lemma 3 (iii) shows $\hat{u} = O_p(1)$, so that also $\exp[-\frac{1}{2}\hat{u}'\tilde{H}\hat{u}] = O_p(1)$. By Lemma 7 and the continuous mapping theorem (CMT), also $(E_\delta \overline{LR}_T(\boldsymbol{\delta}))^{-1} = O_p(1)$, and $\hat{m}_T^{-1} = O_p(1)$ follows.

Furthermore, with $\mathcal{S}_T(\delta)$ and $\mathcal{A}_T(u)$ as defined in Lemma 4,

$$\int E_\delta \left| w(\theta_0 + T^{-1/2}u)LR_T(u,\delta) - w(\theta_0)\widehat{LR}_T(u,\delta) \right| du$$

$$\leq \int E_\delta \left| \mathcal{A}_T(u)\mathcal{S}_T(\delta)w(\theta_0 + T^{-1/2}u)LR_T(u,\delta) - w(\theta_0)\widehat{LR}_T(u,\delta) \right| du$$

$$+ \int E_\delta(1 - \mathcal{A}_T(u)\mathcal{S}_T(\delta))w(\theta_0 + T^{-1/2}u)LR_T(u,\delta)du.$$

The last term converges in probability to zero by Lemma 4, part (i). Also

$$\int E_\delta \left| \mathcal{A}_T(u)\mathcal{S}_T(\delta)w(\theta_0 + T^{-1/2}u)LR_T(u,\delta) - w(\theta_0)\widehat{LR}_T(u,\delta) \right| du$$

$$\leq \int |w(\theta_0 + T^{-1/2}u) - w(\theta_0)|E_\delta \mathcal{A}_T(u)\mathcal{S}_T(\delta)LR_T(u,\delta)du$$

$$+ w(\theta_0) \int E_\delta \left| \mathcal{A}_T(u)\mathcal{S}_T(\delta)LR_T(u,\delta) - \widehat{LR}_T(u,\delta) \right| du.$$

The last term converges in probability to zero by Lemma 2 (iii). For the first term after the inequality, we compute

$$\int |w(\theta_0 + T^{-1/2}u) - w(\theta_0)|E_\delta \mathcal{A}_T(u)\mathcal{S}_T(\delta)LR_T(u,\delta)du$$

$$\leq \sup_{||u||<\alpha_T} |w(\theta_0 + T^{-1/2}u) - w(\theta_0)| \left(\int E_\delta |\mathcal{A}_T(u)\mathcal{S}_T(\delta)LR_T(u,\delta) - \widehat{LR}_T(u,\delta)|du + w(\theta_0)^{-1}\hat{m}_T \right).$$

But $T^{-1/2}\alpha_T \to 0$ and the continuity of w at θ_0 imply $\sup_{||u||<\alpha_T} |w(\theta_0 + T^{-1/2}u) - w(\theta_0)| \to 0$. Furthermore, as shown above, $\hat{m}_T = O_p(1)$, and the result follows from Lemma 2 (iii).

The convergence in probability under the unstable model follows from Lemma 1.

(i) For brevity, we again focus on the case of a flat weighting on θ only.

By definition of the WAR and Fubini's Theorem

$$WAR(\hat{a}) = \int w(\theta_0)E_\delta \int L_T(\theta_0,\delta,\hat{a})f_T(\theta_0,\delta)d\mu_T d\theta_0$$

$$= \int \frac{\int E_\delta L_T(\theta,\delta,\hat{a})f_T(\theta,\delta)w(\theta)d\theta}{\int E_\delta f_T(\theta,\delta)w(\theta)d\theta} \int E_\delta f_T(\theta_0,\delta)w(\theta_0)d\theta_0 d\mu_T$$

$$= \int w(\theta_0) \int \frac{\int E_\delta L_T(\theta_0 + T^{-1/2}u,\delta,\hat{a})LR_T(u,\delta)w(\theta_0 + T^{-1/2}u)du}{m_T} E_\delta f_T(\theta_0,\delta)d\mu_T d\theta_0.$$

Similarly, define

$$\widehat{WAR}(\hat{a}) = \int w(\theta_0) \int \frac{\int E_\delta L_T(\theta_0 + T^{-1/2}u,\delta,\hat{a})\widehat{LR}_T(u,\delta)w(\theta_0)du}{\hat{m}_T} E_\delta f_T(\theta_0,\delta)d\mu_T d\theta_0. \tag{A3}$$

Note that

$$\sup_{a\in\mathbb{A}_T} |WAR(a) - \widehat{WAR}(a)|$$

$$\leq \overline{L} \int w(\theta_0) \iint E_\delta \left| \frac{LR_T(u,\delta)w(\theta_0 + T^{-1/2}u)}{m_T} - \frac{w(\theta_0)\widehat{LR}_T(u,\delta)}{\hat{m}_T} \right| du E_\delta f_T(\theta_0,\delta)d\mu_T d\theta_0. \tag{A4}$$

Now since $m_T > 0$ and $\hat{m}_T > 0$ a.s., we have

$$\int E_\delta \left| \frac{LR_T(u,\delta)w(\theta_0 + T^{-1/2}u)}{m_T} - \frac{w(\theta_0)\widehat{LR}_T(u,\delta)}{\hat{m}_T} \right| du$$

$$\leq \int E_\delta \left(m_T^{-1}LR_T(u,\delta)w(\theta_0 + T^{-1/2}u) + \hat{m}_T^{-1}w(\theta_0)\widehat{LR}_T(u,\delta) \right) du = 2$$

almost surely. Let $M_T = E_\delta \prod_{t=1}^{T} \mathbf{1}[(\theta_0 + \delta_t) \in \Theta] > 0$. Since Θ contains an open ball around θ_0 and $\sup_{\lambda \in [0,1]} ||G(\lambda)||$ is bounded almost surely, $M_T \to 1$. Note that for all T, $M_T^{-1} E_\delta f_T(\theta_0, \delta)$ is a probability density with respect to μ_T, so that the convergence in probability

$$\int E_\delta \left| \frac{LR_T(u, \delta) w(\theta_0 + T^{-1/2} u)}{m_T} - \frac{w(\theta_0) \widehat{LR}_T(u, \delta)}{\hat{m}_T} \right| du \xrightarrow{p} 0$$

established in part (iii) of the Theorem under the unstable model with density $M_T^{-1} E_\delta f_T(\theta_0, \delta)$ implies via dominated convergence that

$$M_T \iint E_\delta \left| \frac{LR_T(u, \delta) w(\theta_0 + T^{-1/2} u)}{m_T} - \frac{w(\theta_0) \widehat{LR}_T(u, \delta)}{\hat{m}_T} \right| du M_T^{-1} E_\delta f_T(\theta_0, \delta) d\mu_T \to 0$$

for almost all θ_0. Since this is also bounded by 2, by another application of the dominated convergence theorem, we have

$$\int w(\theta_0) \iint E_\delta \left| \frac{LR_T(u, \delta) w(\theta_0 + T^{-1/2} u)}{m_T} - \frac{w(\theta_0) \widehat{LR}_T(u, \delta)}{\hat{m}_T} \right| du E_\delta f_T(\theta_0, \delta) d\mu_T d\theta_0 \to 0$$

so that (A4) converges to zero.

Since for any $\hat{a}$, $\widehat{WAR}(\hat{a}) - \widehat{WAR}(\hat{a}^*) \geq 0$ by the definition of $\hat{a}^*$ and $\widehat{WAR}(\hat{a})$,

$$WAR(\hat{a}) - WAR(\hat{a}^*) = \left(\widehat{WAR}(\hat{a}) - \widehat{WAR}(\hat{a}^*) \right) + \left(WAR(\hat{a}) - \widehat{WAR}(\hat{a}) \right) + \left(\widehat{WAR}(\hat{a}^*) - WAR(\hat{a}^*) \right)$$

$$\geq \left(WAR(\hat{a}) - \widehat{WAR}(\hat{a}) \right) + \left(\widehat{WAR}(\hat{a}^*) - WAR(\hat{a}^*) \right) \to 0.$$

(ii) By the Neyman Pearson Lemma, Fubini's Theorem, and a direct calculation, the weighted average power maximizing test of (13) under a Q_T^* weighting rejects for large values of $E_\delta LR_T(0, \delta - \mathbf{e}\bar{\delta})$, and the weighted average power maximizing test in the pseudo model (18) under a $\tilde{Q}_T^*$ weighting rejects for large values of $E_\delta \overline{LR}_T(\delta)$ (where E_δ continues to denote integration with respect to Q_T as defined in Condition 1). We have

$$|E_\delta LR_T(0, \delta - \mathbf{e}\bar{\delta}) - E_\delta \overline{LR}_T(\delta)| \leq E_\delta |\mathcal{S}_T(\delta) LR_T(0, \delta - \mathbf{e}\bar{\delta}) - \overline{LR}_T(\delta)| + E_\delta (1 - \mathcal{S}_T(\delta)) LR_T(0, \delta - \mathbf{e}\bar{\delta}) \xrightarrow{p} 0$$

by applying Lemmas 4 (ii) and 2 (iv). Furthermore, the asymptotic distribution of $E_\delta \overline{LR}_T(\delta)$ under the null hypothesis is absolutely continuous by Lemma 7, so that the result follows from the second claim in Lemma 1 by the same arguments as employed in Andrews and Ploberger (1994) in the proof of their Theorem 2.

Proof of Theorem 2. (iii) In matrix form, the pseudo model (17) is $\hat{\mathbf{s}} + D_{\tilde{h}} \mathbf{e}\hat{\theta} | (D_{\tilde{h}}, \delta, \theta) \sim \mathcal{N}(D_{\tilde{h}}(\delta + \mathbf{e}\theta), D_{\tilde{h}})$, so that conditionally on $D_{\tilde{h}}$ and θ only,

$$\begin{pmatrix} \hat{\mathbf{s}} + D_{\tilde{h}} \mathbf{e}\hat{\theta} \\ \delta \end{pmatrix} | (D_{\tilde{h}}, \theta) \sim \mathcal{N}\left(\begin{pmatrix} D_{\tilde{h}} \mathbf{e}\theta \\ 0 \end{pmatrix}, \begin{pmatrix} D_{\tilde{h}} + D_{\tilde{h}} \Sigma_\delta D_{\tilde{h}} & D_{\tilde{h}} \Sigma_\delta \\ \Sigma_\delta D_{\tilde{h}} & \Sigma_\delta \end{pmatrix} \right).$$

Using the identity

$$(I_{Tk} + D_{\tilde{h}} \Sigma_\delta)^{-1} = I_{Tk} - (I_{Tk} + D_{\tilde{h}} \Sigma_\delta)^{-1} D_{\tilde{h}} \Sigma_\delta \tag{A5}$$

we find with $K = \Sigma_\delta D_{\tilde{h}} (D_{\tilde{h}} + D_{\tilde{h}} \Sigma_\delta D_{\tilde{h}})^{-1} = \Sigma_\delta - \Sigma_\delta D_{\tilde{h}} (D_{\tilde{h}} + D_{\tilde{h}} \Sigma_\delta D_{\tilde{h}})^{-1} D_{\tilde{h}} \Sigma_\delta$ that

$$\delta | (\hat{\mathbf{s}} + D_{\tilde{h}} \mathbf{e}\hat{\theta}, D_{\tilde{h}}, \theta) \sim \mathcal{N}(K(\hat{\mathbf{s}} + D_{\tilde{h}} \mathbf{e}(\hat{\theta} - \theta)), K).$$

Furthermore, with a flat prior, the posterior for θ is proportional to the likelihood, so that $(\hat{\mathbf{s}} + D_{\tilde{h}} \mathbf{e}\hat{\theta}) | (D_{\tilde{h}}, \theta) \sim \mathcal{N}(D_{\tilde{h}} \mathbf{e}\theta, D_{\tilde{h}} + D_{\tilde{h}} \Sigma_\delta D_{\tilde{h}})$ implies $\theta | (\hat{\mathbf{s}} + D_{\tilde{h}} \mathbf{e}\hat{\theta}, D_{\tilde{h}}) \sim \mathcal{N}((\mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} \mathbf{e})^{-1} \mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} D_{\tilde{h}}^{-1} \hat{\mathbf{s}} + \hat{\theta}, (\mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} \mathbf{e})^{-1})$. Thus

$$\begin{pmatrix} \delta \\ \theta \end{pmatrix} | (\hat{\mathbf{s}} + D_{\tilde{h}} \mathbf{e}\hat{\theta}, D_{\tilde{h}}) \sim \mathcal{N}\left(\begin{pmatrix} K(\hat{\mathbf{s}} - D_{\tilde{h}} \mathbf{e}(\mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} \mathbf{e})^{-1} \mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} D_{\tilde{h}}^{-1} \hat{\mathbf{s}}) \\ (\mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} \mathbf{e})^{-1} \mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} D_{\tilde{h}}^{-1} \hat{\mathbf{s}} + \hat{\theta} \end{pmatrix}, V_{\delta\theta} \right)$$

$$\text{where } V_{\delta\theta} = \begin{pmatrix} K + K D_{\tilde{h}} \mathbf{e}(\mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} \mathbf{e})^{-1} \mathbf{e}' D_{\tilde{h}} K & -K D_{\tilde{h}} \mathbf{e}(\mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} \mathbf{e})^{-1} \\ -(\mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} \mathbf{e})^{-1} \mathbf{e}' D_{\tilde{h}} K & (\mathbf{e}'(D_{\tilde{h}}^{-1} + \Sigma_\delta)^{-1} \mathbf{e})^{-1} \end{pmatrix}$$

and employing once more (A5), we conclude $\delta + \mathbf{e}\theta|(\hat{\mathbf{s}} + D_{\tilde{h}}\mathbf{e}\hat{\theta}, D_{\tilde{h}}) \sim \mathcal{N}(\mathbf{e}\hat{\theta} + \Sigma\hat{\mathbf{s}}, \Sigma)$.

(i) Immediate from Theorem 1 (i) and the proof of part (i).

(ii) Note that with $v = T^{-1/2}(\hat{u} - u)$, by Fubini's Theorem and (A2),

$$(2\pi)^{k/2}|\mathbf{e}'D_{\tilde{h}}\mathbf{e}|^{-1/2}E_\delta\overline{LR}_T(\delta) = \int E_\delta \exp\left[\hat{\mathbf{s}}'\delta - \tfrac{1}{2}\delta'D_{\tilde{h}}\delta + v'\mathbf{e}'D_{\tilde{h}}\delta - \tfrac{1}{2}v'\mathbf{e}'D_{\tilde{h}}\mathbf{e}v\right]dv$$

$$= |D_{\tilde{h}}\Sigma_\delta + I_{Tk}|^{-1/2}\int \exp\left[\tfrac{1}{2}(\hat{\mathbf{s}} + D_{\tilde{h}}\mathbf{e}v)'K(\hat{\mathbf{s}} + D_{\tilde{h}}\mathbf{e}v) - \tfrac{1}{2}v'\mathbf{e}'D_{\tilde{h}}\mathbf{e}v\right]dv$$

$$= (2\pi)^{k/2}|D_{\tilde{h}}\Sigma_\delta + I_{Tk}|^{-1/2}|\mathbf{e}'D_{\tilde{h}}\mathbf{e} - \mathbf{e}'D_{\tilde{h}}KD_{\tilde{h}}\mathbf{e}|^{-1/2}\exp\left[\tfrac{1}{2}\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}\right]. \quad (A6)$$

Now let $\overline{R}(\delta) = \exp[-\tfrac{1}{2}\delta'D_{\tilde{h}}\delta + \tfrac{1}{2}\delta'D_{\tilde{h}}\mathbf{e}(\mathbf{e}'D_{\tilde{h}}\mathbf{e})^{-1}\mathbf{e}'D_{\tilde{h}}\delta]$, so that

$$\frac{E_\delta\overline{LR}_T(\delta)}{E_\delta\overline{R}(\delta)} = \exp\left[\tfrac{1}{2}\hat{\mathbf{s}}'\Sigma\hat{\mathbf{s}}\right].$$

By Lemma 2 (ii), $E_\delta\overline{R}(\delta) - E_\delta\exp[-\tfrac{1}{2}\delta'D_\Gamma(\delta - \mathbf{e}\bar{\delta})] \xrightarrow{p} 0$. By the CMT, $\exp[-\tfrac{1}{2}\delta'D_\Gamma(\delta - \mathbf{e}\bar{\delta})] \Rightarrow \exp[-\tfrac{1}{2}\int_0^1 G^*(s)'$ $\Gamma(s)G^*(s)ds]$, and since $\overline{R}(\delta) < 1$ a.s., also $E_\delta\overline{R}(\delta) \to E_G\exp[-\tfrac{1}{2}\int_0^1 G^*(s)'\Gamma(s)G^*(s)ds]$. The result now follows from Lemma 7.

Proof of Theorem 3. We write $\hat{\Lambda}_t$ for $\hat{\Lambda}_{T,t}$ to enhance readability, and let $D_{\hat{\Lambda}} = \text{diag}(\hat{\Lambda}_1, \ldots, \hat{\Lambda}_T)$.

For the first claim, note that if $\{\tilde{h}_t\}_{t=1}^T$ satisfies (16), so does $\hat{\Lambda}_t'\tilde{h}_t$ and $\hat{\Lambda}_t'\tilde{h}_t\hat{\Lambda}_t$. Also, by summation by parts with $\hat{\Lambda}_0 = \hat{\Lambda}_1$

$$T^{-1/2}\sum_{j=1}^t \hat{\Lambda}_j's_j(\hat{\theta}) = \hat{\Lambda}_t'T^{-1/2}\sum_{j=1}^t s_j(\hat{\theta}) - \sum_{j=1}^t(\hat{\Lambda}_j' - \hat{\Lambda}_{j-1}')(T^{-1/2}\sum_{l=1}^{j-1}s_l(\hat{\theta}))$$

so that $T^{-1/2}\sum_{j=1}^{[\cdot T]}\hat{s}_j \Rightarrow S_L(\cdot)$ implies $T^{-1/2}\sum_{j=1}^{[\cdot T]}\hat{\Lambda}_j'\hat{s}_j \Rightarrow S_L(\cdot)$. The result thus follows from Theorem 2 (ii).

For the second claim, proceed as in the proof of part (i) of Theorem 1, but with $\widehat{WAR}(\hat{a})$ in (A3) substituted by

$$\widehat{WAR}_\Lambda(\hat{a}) = \int w(\theta_0) \int \frac{\int E_\delta w(\theta_0)\widehat{LR}_T(u, D_{\hat{\Lambda}}\delta)L(\theta_0 + T^{-1/2}u, D_{\hat{\Lambda}}\delta, \hat{a})du}{\hat{m}_{\Lambda,T}}E_\delta f_T(\theta_0, \delta)d\mu_T d\theta_0$$

where $\hat{m}_{\Lambda,T} = \int E_\delta w(\theta_0)\widehat{LR}_T(u, D_{\hat{\Lambda}}\delta)du$. We have

$$|\widehat{WAR}_\Lambda(\hat{a}) - \int w(\theta_0) \int \frac{\int E_\delta w(\theta_0)\widehat{LR}_T(u, D_{\hat{\Lambda}}\delta)L(\theta_0 + T^{-1/2}u, \delta, \hat{a})du}{\hat{m}_{\Lambda,T}}E_\delta f_T(\theta_0, \delta)d\mu_T d\theta_0|$$

$$\leq \int w(\theta_0)\left(\sup_{\theta\in\Theta, \delta\in\mathbb{R}^{Tk}, a\in\mathbb{A}_T}|L_T(\theta, D_{\hat{\Lambda}}\delta, a) - L_T(\theta, \delta, a)|\right)E_\delta f_T(\theta_0, \delta)d\mu_T d\theta_0 \to 0$$

where the convergence follows from $\sup_{\theta\in\Theta, \delta\in\mathbb{R}^{Tk}, a\in\mathbb{A}_T}|L_T(\theta, D_{\hat{\Lambda}}\delta, a) - L_T(\theta, \delta, a)| \xrightarrow{p} 0$ in the stable model, Lemma 1 and the dominated convergence theorem as $\sup_{\theta\in\Theta, \delta\in\mathbb{R}^{Tk}, a\in\mathbb{A}_T}|L_T(\theta, D_{\hat{\Lambda}}\delta, a) - L_T(\theta, \delta, a)| \leq 2\overline{L}$. It thus suffices to proceed as in the proof of Theorem 1 with $\widehat{LR}_T(u, \delta)$ replaced by $\widehat{LR}_T(u, D_{\hat{\Lambda}}\delta)$, and the result follows from Theorem 2 (i), as in the proof of the first claim.

Proof of Theorem 4. For the claim regarding the analogous statement of Theorem 1 (iii), proceed up to equation (A1) as in the proof of Theorem 1 (iii) with E_δ now denoting integration with respect to the mixture. With $E_{\delta(i)}$ denoting integration with respect to the measure of $\{T^{-1/2}G_i(t/T)\}$, it then suffices to show that $\int E_{\delta(i)}|w(\theta_0 + T^{-1/2}u)LR_T(u, \delta_{(i)}) - w(\theta_0)\widehat{LR}_T(u, \delta_{(i)})|du \xrightarrow{p} 0$ for $i = 1, \ldots, n_G$ and $\hat{m}_T^{*-1} = (\sum_i p_i\hat{m}_{(i),T})^{-1} = O_p(1)$, where $\hat{m}_{(i),T} = w(\theta_0)\int E_{\delta(i)}\widehat{LR}_T(u, \delta_{(i)})du$. From the same reasoning as in the proof of Theorem 1 (iii), $\hat{m}_{(i),T}^{-1} = O_p(1)$, so that also $\hat{m}_T^{*-1} = O_p(1)$. The result now follows by proceeding as in the remainder of the proof of Theorem 1 (iii) and by invoking Lemmas 4 (i) and 2 (iii) for each of the n_G components in the measure of δ.

The conditionally normal distribution $\mathcal{N}(\mathbf{e}\hat{\theta} + \Sigma_i\hat{\mathbf{s}}, \Sigma_i)$ follows as in the proof of Theorem 2. For the mixing probabilities, note that the Bayes factor between pseudo models i and j is given by $\hat{m}_{(i),T}/\hat{m}_{(j),T}$. Using (A2) and (A6), we find

$$\frac{\hat{m}_{(i),T}}{\hat{m}_{(j),T}} = \frac{|D_{\tilde{h}}\Sigma_{\delta(i)} + I_{Tk}|^{-1/2}|\mathbf{e}'D_{\tilde{h}}\mathbf{e} - \mathbf{e}'D_{\tilde{h}}K_iD_{\tilde{h}}\mathbf{e}|^{-1/2}\exp[\tfrac{1}{2}\hat{\mathbf{s}}'\Sigma_i\hat{\mathbf{s}}]}{|D_{\tilde{h}}\Sigma_{\delta(j)} + I_{Tk}|^{-1/2}|\mathbf{e}'D_{\tilde{h}}\mathbf{e} - \mathbf{e}'D_{\tilde{h}}K_jD_{\tilde{h}}\mathbf{e}|^{-1/2}\exp[\tfrac{1}{2}\hat{\mathbf{s}}'\Sigma_j\hat{\mathbf{s}}]}$$

so that the posterior odds of model i and model j in the pseudo model are as stated.

Proof of Theorem 5. Let $\mathrm{LR}_T^{\pi_0}$ be the likelihood ratio statistic between the correctly specified model with parameter evolution $\theta_t = \theta_0 + T^{-1/2}\pi_0(t/T)$, and $\hat{u}_T = T^{1/2}(\hat{\theta} - \theta_0)$. Proceed as in Lemma 1 of Li and Müller (2009) and Lemma 3 to show that in the stable model,

$$\begin{pmatrix} \mathrm{LR}_T^{\pi_0} \\ T^{-1/2}\sum_{t=1}^{[\cdot T]}\hat{s}_t \\ \hat{u}_T \end{pmatrix} \Rightarrow \begin{pmatrix} \exp[\int_0^1 \pi_0(l)'\Gamma(l)^{1/2}dW(l) - \frac{1}{2}\int_0^1 \pi_0(l)'\Gamma(l)\pi_0(l)dl] \\ \int_0^{\cdot}\Gamma(l)^{1/2}dW(l) - \int_0^{\cdot}\Gamma(l)dl(\int_0^1\Gamma(l)dl)^{-1}\int_0^1\Gamma(l)^{1/2}dW(l) \\ (\int_0^1\Gamma(l)dl)^{-1}\int_0^1\Gamma(l)^{1/2}dW(l) \end{pmatrix}.$$

Thus, by a general version of LeCam's Third Lemma (see, for instance, Pollard, 2001), we have that in the unstable model with parameter evolution $\theta_t = \theta_0 + T^{-1/2}\pi_0(t/T)$, $(S_T(\cdot) = T^{-1/2}\sum_{t=1}^{[\cdot T]}\hat{s}_t \Rightarrow S_{\pi_0}(\cdot), \hat{u}_T \Rightarrow U_{\pi_0})$, where $S_{\pi_0}(\cdot)$ and U_{π_0} are the weak limits of $T^{-1/2}\sum_{t=1}^{[\cdot T]}\hat{s}_t^r$ and $\hat{u}_T^r = T^{1/2}(\hat{\theta}^r - \theta_0)$ in Condition 3, respectively.

(i) By Theorems 11.7.1 and 11.7.2 of Dudley (2002), there exist a probability space $(\tilde{\mathcal{F}}, \tilde{\mathfrak{z}}, \tilde{P})$ with associated random elements $\tilde{U}_{\pi_0}, (\tilde{u}_T, \tilde{u}_T^r), \tilde{S}_{\pi_0}(\cdot), \tilde{S}_T(\cdot), \tilde{S}_T^r(\cdot), \tilde{\Upsilon}_T(\cdot)$ and $\tilde{\Upsilon}_T^r(\cdot)$ such that (i) for all $T \geq 1$, $(\tilde{u}_T, \tilde{u}_T^r, \tilde{S}_T(\cdot), \tilde{S}_T^r(\cdot), \tilde{\Upsilon}_T(\cdot), \tilde{\Upsilon}_T^r(\cdot))$ has the same distribution as $(\hat{u}_T, \hat{u}_T^r, S_T(\cdot), T^{-1}\sum_{t=1}^{[\cdot T]}\hat{s}_t^r, T^{-1}\sum_{t=1}^{[\cdot T]}\tilde{h}_t, T^{-1}\sum_{t=1}^{[\cdot T]}\tilde{h}_t^r)$ and $(\tilde{U}_{\pi_0}, \tilde{S}_{\pi_0}(\cdot))$ has the same distribution as $(U_{\pi_0}, S_{\pi_0}(\cdot))$; and (ii) $(\tilde{u}_T, \tilde{u}_T^r, \tilde{S}_T(\cdot), \tilde{S}_T^r(\cdot), \tilde{\Upsilon}_T(\cdot), \tilde{\Upsilon}_T^r(\cdot)) \to (\tilde{U}_{\pi_0}, \tilde{U}_{\pi_0}, \tilde{S}_{\pi_0}(\cdot), \tilde{S}_{\pi_0}(\cdot), \int_0^{\cdot}\Gamma dl, \int_0^{\cdot}\Gamma dl)$ $\tilde{P}$-almost surely. Since $\tilde{S}_{\pi_0}(\cdot)$ and $\int_0^{\cdot}\Gamma dl$ are continuous almost surely, note that this also implies $\sup_{\lambda \in [0,1]}||\tilde{S}_T(\lambda) - \tilde{S}_T^r(\lambda)|| \to 0$ and $\sup_{\lambda \in [0,1]}||\tilde{\Upsilon}_T(\lambda) - \tilde{\Upsilon}_T^r(\lambda)|| \to 0$ $\tilde{P}$-almost surely. Define $\tilde{\Pi}$ and $\tilde{\Pi}^r$ just as Π in Theorem 2 with $(\hat{\theta}, \{\hat{s}_t\}, \{\tilde{h}_t\})$ replaced by $(\theta_0 + T^{1/2}\tilde{u}_T, \{T^{1/2}(\tilde{S}_T(t/T) - \tilde{S}_T((t-1)/T))\}, \{T(\tilde{\Upsilon}_T(t/T) - \tilde{\Upsilon}_T((t-1)/T))\})$ and $(\theta_0 + T^{1/2}\tilde{u}_T^r, \{T^{1/2}(\tilde{S}_T^r(t/T) - \tilde{S}_T^r((t-1)/T))\}, \{T(\tilde{\Upsilon}_T^r(t/T) - \tilde{\Upsilon}_T^r((t-1)/T))\})$, respectively. Now proceeding as in the proof of Theorem 1 (iii) shows via Lemma 2 (i) that the total variation distance between $\tilde{\Pi}$ and $\tilde{\Pi}^r$ converges to zero $\tilde{P}$-almost surely. By assumption about L_T, this implies the corresponding convergence $L_T(\theta_0, \overline{\pi}_0, a^*(\tilde{\Pi})) - L_T(\theta_0, \overline{\pi}_0, a^*(\tilde{\Pi}^r)) \to 0$ $\tilde{P}$-almost surely, which implies the result by another application of Theorem 11.7.1 of Dudley (2002).

(ii) and (iii) Immediate from the proof of part (i), the proof of part (ii) of Theorem 2, and Lemma 2 (ii). ∎

Lemma 2. *Define* $\widehat{LR}_T^a(u, \delta) = \exp[\sum \hat{s}_{at}'\delta_t - \frac{1}{2}\sum \delta_t' h_{a1t}\delta_t + T^{-1/2}(\hat{u}_a - u)'\sum h_{a2t}\delta_t - \frac{1}{2}u'H_a u + \hat{u}_a'H_a u]$ *and* $\overline{LR}_T^a(\delta) = \exp[\sum s_{at}'\delta_t - \frac{1}{2}\sum\delta_t'h_{a1t}\delta_t + \frac{1}{2}(T^{-1/2}\sum\delta_t'h_{a2t})H_a^{-1}T^{-1/2}\sum h_{a2t}\delta_t]$ *for* $i = 1, 2$. *Suppose* $T^{-1/2}\sum_{t=1}^{[\cdot T]}\hat{s}_t \Rightarrow S_L(\cdot)$, $\sup_{t \leq T}||T^{-1/2}\sum_{s=1}^t(\hat{s}_s - \hat{s}_{as})|| \xrightarrow{p} 0$, $(\hat{u}, \hat{u}_a) \Rightarrow (U_L, U_L)$, $H_a \xrightarrow{p} \int_0^1\Gamma(\lambda)d\lambda$ *and* $\sup_{t \leq T}||T^{-1}\sum_{s=1}^t(h_{a1s} - \tilde{h}_s, h_{a2s} - \tilde{h}_s)|| \xrightarrow{p} (0,0)$. *Then under Condition 2,*

(i) $\int E_\delta\left|\widehat{LR}_T(u, \delta) - \widehat{LR}_T^a(u, \delta)\right| du \xrightarrow{p} 0$

(ii) $E_\delta\overline{LR}_T(\delta) - E_\delta\overline{LR}_T^a(\delta) \xrightarrow{p} 0$

Furthermore, if in addition Condition 1 holds, then also

(iii) $\int E_\delta\left|\widehat{LR}_T(u, \delta) - \mathcal{A}_T(u)\mathcal{S}_T(\delta)LR_T(u, \delta)\right| du \xrightarrow{p} 0$

(iv) $E_\delta\overline{LR}_T(\delta) - E_\delta\mathcal{S}_T(\delta)LR_T(0, \delta - \mathbf{e}\overline{\delta}) \xrightarrow{p} 0$

Proof. (i) By the Cauchy–Schwarz inequality, we find

$$\int E_\delta\left|\widehat{LR}_T(u, \delta) - \widehat{LR}_T^a(u, \delta)\right| du \leq \int (E_\delta\widehat{LR}_T(u, \delta)^2)^{1/2} \cdot (E_\delta(1 - \exp[\varsigma_T(u, \delta)])^2)^{1/2}du \tag{A7}$$

where

$$\varsigma_T(u, \delta) = \sum(\hat{s}_{at} - \hat{s}_t + T^{-1/2}\hat{u}_a h_{at} - T^{-1/2}\hat{u}'\tilde{h}_t)'\delta_t - \frac{1}{2}\sum\delta_t'(h_{a1t} - \tilde{h}_t)\delta_t$$
$$- T^{-1/2}u'\sum(h_{a2t} - \tilde{h}_t)\delta_t - \frac{1}{2}u'(H_a - \tilde{H})u + \hat{u}'(H_a - \tilde{H})u.$$

We have

$$\widehat{LR}_T(u, \delta)^2 = \exp[2\hat{s}'\delta - (\delta - T^{-1/2}\mathbf{e}(\hat{u} - u))'D_{\tilde{h}}(\delta - T^{-1/2}\mathbf{e}(\hat{u} - u)) + 2\hat{u}'\tilde{H}\hat{u}]$$

and by another application of the Cauchy–Schwarz inequality

$$E_\delta\widehat{LR}_T(u, \delta)^2 \leq \exp[2\hat{u}'\tilde{H}\hat{u}](E_\delta\exp[4\hat{s}'\delta])^{1/2}(E_\delta\exp[-2(\delta - T^{-1/2}\mathbf{e}(\hat{u} - u))'D_{\tilde{h}}(\delta - T^{-1/2}\mathbf{e}(\hat{u} - u))])^{1/2}.$$

By Lemma 6 (iii), $E_\delta\exp[4\hat{s}'\delta] = O_p(1)$, and $\exp[2\hat{u}'\tilde{H}\hat{u}] = O_p(1)$ by assumption. By Lemma 6 (i),

$$E_\delta\exp[-2(\delta - T^{-1/2}\mathbf{e}(\hat{u} - u))'D_{\tilde{h}}(\delta - T^{-1/2}\mathbf{e}(\hat{u} - u))] \leq \exp\left[-\frac{1}{2}\tilde{C}_T||\hat{u} - u||^2\right]$$

where $\tilde{C}_T = O_p(1)$ and $\tilde{C}_T^{-1} = O_p(1)$ and does not depend on u. Therefore,

$$E_\delta \widehat{LR}_T(u,\delta)^2 \leq O_p(1)(2\pi)^{-k/2}\tilde{C}_T^{k/2}\exp\left[-\tfrac{1}{2}\tilde{C}_T||\hat{u}-u||^2\right] \tag{A8}$$

so that with $\Phi(u)$ the c.d.f. of $u \sim \mathcal{N}(\hat{u}, \tilde{C}_T^{-1}I_k)$, from (A7) and Jensen's inequality

$$\left(\int E_\delta\left|\widehat{LR}_T(u,\delta) - \widehat{LR}_T^a(u,\delta)\right|du\right)^2 \leq K_{1T}\int E_\delta(1 - \exp[\varsigma_T(u,\delta)])^2 d\Phi(u)$$

where $K_{1T} = O_p(1)$, so that is suffices to show that $\int E_\delta \exp[\varsigma_T(u,\delta)]d\Phi(u)$ is bounded below by a random variable that converges to one in probability, and $\int E_\delta \exp[2\varsigma_T(u,\delta)]d\Phi(u)$ is bounded above by a random variable that converges to one in probability. By Lemma 6 (ii), there exist random variables $\underline{\kappa}_T \xrightarrow{p} 1$, $\overline{\kappa}_T \xrightarrow{p} 1$, $\underline{\Delta}_T \xrightarrow{p} 0$ and $\overline{\Delta}_T \xrightarrow{p} 0$ such that

$$\int E_\delta \exp[\varsigma_T(u,\delta)]d\Phi(u) \geq \underline{\kappa}_T \int \exp\left[\underline{\Delta}_T||u||^2 - \tfrac{1}{2}u'(H_a - \tilde{H})u + \hat{u}'(H_a - \tilde{H})u\right]d\Phi(u)]d\Phi(u)$$

$$\int E_\delta \exp[2\varsigma_T(u,\delta)]d\Phi(u) \leq \overline{\kappa}_T \int \exp\left[\overline{\Delta}_T||u||^2 - \tfrac{1}{2}u'(H_a - \tilde{H})u + \hat{u}'(H_a - \tilde{H})u\right]d\Phi(u)]d\Phi(u)$$

and the result follows.

(ii) By a direct calculation,

$$(2\pi)^{-k/2}|\tilde{H}|^{-1/2}\exp\left[\tfrac{1}{2}\hat{u}'\tilde{H}\hat{u}\right]|E_\delta\overline{LR}_T(\delta) - E_\delta\overline{LR}_T^a(\delta)|$$

$$= \left|E_\delta\int\widehat{LR}_T(u,\delta)du - E_\delta\left|H_a\right|^{1/2}|\tilde{H}|^{-1/2}\exp\left[-\tfrac{1}{2}\hat{u}'(H_a - \tilde{H})\hat{u}\right]\int\widehat{LR}_T^a(u,\delta)du\right|$$

$$\leq \int E_\delta\left|\widehat{LR}_T(u,\delta) - \widehat{LR}_T^a(u,\delta)\right|du + |1 - |H_a|^{1/2}|\tilde{H}|^{-1/2}\exp\left[-\tfrac{1}{2}\hat{u}'(H_a - \tilde{H})\hat{u}\right]|E_\delta\int\widehat{LR}_T^a(u,\delta)du$$

and the result follows from part (i) and $|H_a|^{1/2}|\tilde{H}|^{-1/2}\exp[-\tfrac{1}{2}\hat{u}'(H_a - \tilde{H})\hat{u}] \xrightarrow{p} 1$.

(iii) Let $\mathcal{U}_T$ be the indicator of the event that $||\hat{u}|| \leq \alpha_T$. By Lemma 3 (iii), $\hat{u} = O_p(1)$, so that $E\mathcal{U}_T \to 1$. Note that if $\mathcal{U}_T\psi_T \xrightarrow{p} 0$ for some sequence of random variables ψ_T, then also $\psi_T \xrightarrow{p} 0$. Let T be large enough such that $\Theta_T = \{\theta : ||\theta - \theta_0|| < 2T^{-1/2}\alpha_T + T^{\eta-1/2} + T^{\eta-1/2}(\sup_{\lambda\in[0,1]}||\Gamma(\lambda)||)/(\inf_{\lambda\in[0,1]}||\Gamma(\lambda)||)\} \subset \Theta_0$, so that $\theta_0 + \mathcal{A}_T(u)\mathcal{S}_T(\delta)\mathcal{U}_T(T^{-1/2}(u - \hat{u}) + \delta_t - \bar{\delta}) \in \Theta_0$ almost surely for all $t \leq T$.

Let $g_v : [0,1] \mapsto \mathbb{R}$ with $g_v(\lambda) = l_t(\theta_0 + \lambda v) - l_t(\theta_0)$. Note that for $\theta_0 + v \in \Theta_T$, g_v is twice continuously differentiable with $g_v'(\lambda) = v's_t(\theta_0 + \lambda v)$ and $g_v''(\lambda) = -v'h_t(\theta_0 + \lambda v)v$, so that by a first order Taylor expansion in the integral remainder form, $l_t(\theta_0 + v) - l_t(\theta_0) = g_v(1) - g_v(0) = g_v'(0) + \int_0^1 \lambda g_v''(1 - \lambda)d\lambda = v's_t(\theta_0) - \tfrac{1}{2}v'(2\int_0^1 \lambda h_t(\theta_0 + (1-\lambda)v)d\lambda)v$, and similarly, $s_t(\theta_0 + v) = s_t(\theta_0) - (\int_0^1 h_t(\theta_0 + \lambda v)d\lambda)v$. Thus, for $||u|| < \alpha_T$, $T^{1/2}\sup_{t\leq T}||\delta_t|| < T^\eta$ and $||\hat{u}|| < \alpha_T$

$$l_t(\theta_0 + T^{-1/2}u + \delta_t) - l_t(\theta_0 + T^{-1/2}u) = s_t(\theta_0 + T^{-1/2}u)'\delta_t - \tfrac{1}{2}\delta_t'h_{1,t}(u,\delta)\delta_t$$

$$l_t(\theta_0 + T^{-1/2}u) - l_t(\theta_0) = T^{-1/2}u's_t(\theta_0) - \tfrac{1}{2}u'h_{2,t}(u)u$$

$$s_t(\theta_0 + T^{-1/2}u) = s_t(\theta_0 + T^{-1/2}\hat{u}) - h_{3,t}(u,\hat{u})T^{-1/2}(u - \hat{u})$$

$$s_t(\theta_0) = s_t(\theta_0 + T^{-1/2}\hat{u}) + h_{4,t}(\hat{u})T^{-1/2}\hat{u} \tag{A9}$$

almost surely, where $h_{1,t}(u,\delta) = 2\int_0^1 \lambda h_t(\theta_0 + T^{-1/2}u + (1-\lambda)\delta_t)d\lambda$, $h_{2,t}(u) = 2\int_0^1 \lambda h_t(\theta_0 + (1-\lambda)T^{-1/2}u)d\lambda$, $h_{3,t}(u,\hat{u}) = \int_0^1 h_t(\theta_0 + \lambda T^{-1/2}(\hat{u} - u))d\lambda$ and $h_{4,t}(\hat{u}) = \int_0^1 h_t(\theta_0 + \lambda T^{-1/2}\hat{u})d\lambda$, $t = 1, \ldots, T$. Define $\{h_{1,t}(u,\delta)\}_{t=1}^T = \{h_t(\theta_0)\}_{t=1}^T$ when $||u|| \geq \alpha_T$ or $T^{1/2}\sup_{t\leq T}||\delta_t|| > T^\eta$, define $\{h_{2,t}(u)\}_{t=1}^T = \{\tilde{h}_t\}_{t=1}^T$ when $||u|| \geq \alpha_T$, define $\{h_{3,t}(u,\hat{u})\}_{t=1}^T = \{\tilde{h}_t\}_{t=1}^T$ when $||u|| \geq \alpha_T$ or $||\hat{u}|| \geq \alpha_T$, and define $\{h_{4,t}(\hat{u})\}_{t=1}^T = \{\tilde{h}_t\}_{t=1}^T$ when $||\hat{u}|| > \alpha_T$. Further, let $\hat{H}_4(\hat{u}) = T^{-1}\sum h_{4,t}(\hat{u})$ and $\hat{H}_2(u) = T^{-1}\sum h_{2,t}(u)$. For notational convenience, we drop the dependence of $h_{1,t}$, $h_{2,t}$, $\hat{H}_3$ and $\hat{H}_4$ on u, $\hat{u}$ and δ. With these definitions, we have

$$\mathcal{A}_T(u)\mathcal{S}_T(\delta)\mathcal{U}_T\left|LR_T(u,\delta) - \exp\left[\sum\hat{s}_t'\delta_t + T^{-1/2}(\hat{u} - u)'\sum h_{3,t}\delta_t - \tfrac{1}{2}\sum\delta_t'h_{1,t}\delta_t + \hat{u}'\hat{H}_4u - \tfrac{1}{2}u'\hat{H}_2u\right]\right| = 0$$

almost surely, uniformly in $u \in \mathbb{R}^k$ and $\delta \in \mathbb{R}^{Tk}$.

Let

$$\varsigma_T(u,\delta) = (\hat{u} - u)' T^{-1/2} \sum (h_{3,t} - \tilde{h}_t)\delta_t - \tfrac{1}{2} \sum \delta_t'(h_{1,t} - \tilde{h}_t)\delta_t - \tfrac{1}{2} u'(\hat{H}_2 - \tilde{H})u + \hat{u}'(\hat{H}_4 - \tilde{H})u.$$

Now $\sup_{u \in \mathbb{R}^k, \delta \in \mathbb{R}^{Tk}} \mathcal{A}_T(u)\mathcal{S}_T(\delta)\mathcal{U}_T|LR_T(u,\delta) - \widehat{LR}(u,\delta)\exp \varsigma_T(u,\delta)| = 0$ a.s., and by the Cauchy–Schwarz inequality and $\mathcal{U}_T \leq 1$ a.s.

$$\mathcal{U}_T \int E_\delta \left| \widehat{LR}_T(u,\delta) - \mathcal{A}_T(u)\mathcal{S}_T(\delta)LR_T(u,\delta) \right| du$$

$$\leq \int [(E_\delta \widehat{LR}_T(u,\delta)^2)(E_\delta(1 - \mathcal{A}_T(u)\mathcal{S}_T(\delta)\exp \varsigma_T(u,\delta))^2)]^{1/2} du.$$

Proceeding as in the proof of part (i), it suffices to show that $\int E_\delta(1 - \mathcal{A}_T(u)\mathcal{S}_T(\delta)\exp \varsigma_T)^2 d\Phi(u) \xrightarrow{p} 0$. We first compute the expectation with respect to δ. This is complicated by the fact that $h_{1,t}$ depends on δ. To circumvent this problem, we bound $\varsigma_T(u,\delta)$ by $\underline{\varsigma}_T(u,\delta) \leq \varsigma_T(u,\delta) \leq \overline{\varsigma}_T(u,\delta)$, where $\underline{\varsigma}_T(u,\delta)$ and $\overline{\varsigma}_T(u,\delta)$ are defined just as $\varsigma_T(u,\delta)$, but with $h_{1,t}$ replaced by a term that does not depend on δ (or u).

Specifically, for each $t \leq T$, define $d_t = 2\sup_{\|v\| < \alpha_T + T^\eta} \left\| h_t(\theta_0 + T^{-1/2}v) - h_t(\theta_0) \right\|$, so that for any $v \in \mathbb{R}^k$ with $\|v\| = 1$,

$$|v'(h_t(\theta_0) - h_{1,t})v| \leq \|h_t(\theta_0) - h_{1,t}\| \leq d_t$$

since for $\|u\| < \alpha_T$ and $T^{1/2}\sup_{t \leq T}\|\delta_t\| < T^\eta$, $\|h_t(\theta_0) - h_{1,t}\| = \|2\int_0^1 \lambda(h_t(\theta_0 + T^{-1/2}u + (1-\lambda)\delta_t) - h_t(\theta_0))d\lambda\|$ and $h_{1,t}(u,\delta) = h_t(\theta_0)$ otherwise. Thus, for all $\delta \in \mathbb{R}^{Tk}$,

$$\sum \delta_t'(h_t(\theta_0) - d_t I_k)\delta_t \leq \sum \delta_t' h_{1,t}\delta_t \leq \sum \delta_t'(h_t(\theta_0) + d_t I_k)\delta_t.$$

Now let

$$\overline{\varsigma}_T(u,\delta) = \varsigma_T(u,\delta) + \tfrac{1}{2}\sum \delta_t'(h_{1,t} - h_t(\theta_0) + d_t I_k)\delta_t$$

$$\underline{\varsigma}_T(u,\delta) = \varsigma_T(u,\delta) + \tfrac{1}{2}\sum \delta_t'(h_{1,t} - h_t(\theta_0) - d_t I_k)\delta_t$$

so that $\underline{\varsigma}_T(u,\delta) \leq \varsigma_T(u,\delta) \leq \overline{\varsigma}_T(u,\delta)$. We obtain

$$0 \leq E_\delta(1 - \mathcal{A}_T(u)\mathcal{S}_T(\delta)\exp \varsigma_T(u,\delta))^2$$

$$\leq 1 - 2E_\delta \mathcal{A}_T(u)\mathcal{S}_T(\delta)\exp \underline{\varsigma}_T(u,\delta) + E_\delta \mathcal{A}_T(u)\mathcal{S}_T(\delta)\exp 2\overline{\varsigma}_T(u,\delta)$$

$$\leq 1 - 2E_\delta \exp \underline{\varsigma}_T(u,\delta) + E_\delta \exp 2\overline{\varsigma}_T(u,\delta) + 2E_\delta(1 - \mathcal{A}_T(u)\mathcal{S}_T(\delta))\exp \underline{\varsigma}_T(u,\delta)$$

so it suffices to show that $\int E_\delta \exp \underline{\varsigma}_T(u,\delta) d\Phi(u)$ is bounded below by random variable that converges to one in probability, that $\int E_\delta \exp 2\overline{\varsigma}_T(u,\delta) d\Phi(u)$ is bounded above by a random variable that converges to one in probability, and $\int E_\delta(1 - \mathcal{S}_T(\delta)\mathcal{A}_T(u))\exp \underline{\varsigma}_T(u,\delta) d\Phi(u) \xrightarrow{p} 0$.

With $D_{h3} = \text{diag}(h_{3,1}, \ldots, \bar{h}_{3,T})$, $D_h = \text{diag}(h_1(\theta_0), \ldots, h_T(\theta_0))$ and $D_d = \text{diag}(d_1 I_k, \ldots, d_T I_k)$ we have

$$E_\delta \exp \underline{\varsigma}_T(u,\delta) = \exp \left[-\tfrac{1}{2}u'(\hat{H}_2 - \tilde{H})u + \hat{u}'(\hat{H}_4 - \tilde{H})u \right]$$

$$\cdot E_\delta \exp \left[(\hat{u} - u)'T^{-1/2}\mathbf{e}'(D_{h3} - D_{\tilde{h}})\delta - \tfrac{1}{2}\delta'(D_h - D_{\tilde{h}} + D_d)\delta \right]$$

and

$$E_\delta \exp 2\overline{\varsigma}_T(u,\delta) = \exp[-u'(\hat{H}_2 - \tilde{H})u + 2\hat{u}'(\hat{H}_4 - \tilde{H})u]$$

$$\cdot E_\delta \exp[2(\hat{u} - u)'T^{-1/2}\mathbf{e}'(D_{h3} - D_{\tilde{h}})\delta - \delta'(D_h - D_{\tilde{h}} - D_d)\delta].$$

Since

$$\sup_{u \in \mathbb{R}^k, t \leq T} T^{-1} \left\| \sum_{s=1}^t (h_{3,s}(u,\hat{u}) - \tilde{h}_s) \right\| \leq \sup_{t \leq T, \|u\| \leq \alpha_T, \|\hat{u}\| < \alpha_T} T^{-1} \left\| \sum_{s=1}^t h_{3,s}(u,\hat{u}) - \tilde{h}_s \right\| \xrightarrow{p} 0$$

$$\sup_{t \leq T} T^{-1} \left\| \sum_{s=1}^t (h_s(\theta_0) + d_t I_k - \tilde{h}_s) \right\| \leq \sup_{t \leq T} \left\| T^{-1} \sum_{s=1}^t (h_s(\theta_0) - \tilde{h}_s) \right\| + T^{-1}\sum_{t=1}^T d_t \xrightarrow{p} 0$$

by (16), Lemma 3 (ii) and Condition 1 (LLLN), and similarly, $\sup_{t \leq T} ||T^{-1} \sum_{s=1}^{t} (h_s(\theta_0) - d_t I_k - \Gamma_s)|| \xrightarrow{p} 0$, Lemma 6 (ii) is applicable, and we obtain

$$E_\delta \exp \underline{\varsigma}_T(u, \delta) \geq \exp\left[-\tfrac{1}{2} u'(\hat{H}_2 - \tilde{H})u + \hat{u}'(\hat{H}_4 - \tilde{H})u\right] \underline{\kappa}_T \exp[\underline{\Delta}_T ||u - \hat{u}||^2]$$

$$E_\delta \exp 2\overline{\varsigma}_T(u, \delta) \leq \exp[-u'(\hat{H}_2 - \tilde{H})u + 2\hat{u}'(\hat{H}_4 - \tilde{H})u]\overline{\kappa}_T \exp[\overline{\Delta}_T ||u - \hat{u}||^2]$$

uniformly in u, where $\underline{\kappa}_T$, $\overline{\kappa}_T$, $\underline{\Delta}_T$ and $\overline{\Delta}_T$ do not depend on u and $\underline{\kappa}_T \xrightarrow{p} 1$, $\overline{\kappa}_T \xrightarrow{p} 1$, $\underline{\Delta}_T \xrightarrow{p} 0$ and $\overline{\Delta}_T \xrightarrow{p} 0$. Also

$$\sup_{u \in \mathbb{R}^k} ||\hat{H}_2(u) - \tilde{H}|| \leq \sup_{||u|| < \alpha_T} T^{-1}|| \sum h_{2,t}(u) - \Gamma_t || \xrightarrow{p} 0$$

by (16) and Lemma 3 (ii), and similarly, $\hat{H}_4 - \tilde{H} \xrightarrow{p} 0$. Thus, $\int E_\delta \exp \underline{\varsigma}_T(u, \delta) d\Phi(u) \geq \int \underline{\kappa}_T^* \exp[\underline{\Delta}_T^* ||u - \hat{u}||^2]$ $d\Phi(u) \xrightarrow{p} 1$ and $\int E_\delta \exp 2\overline{\varsigma}_T(u, \delta) d\Phi(u) \leq \int \overline{\kappa}_T^* \exp[\overline{\Delta}_T^* ||u - \hat{u}||^2] d\Phi(u) \xrightarrow{p} 1$ for suitably defined $\underline{\kappa}_T^* \xrightarrow{p} 1$, $\overline{\kappa}_T^* \xrightarrow{p} 1$, $\underline{\Delta}_T^* \xrightarrow{p} 0$ and $\overline{\Delta}_T^* \xrightarrow{p} 0$. We are left to show that $\int E_\delta(1 - \mathcal{S}_T(\delta)\mathcal{A}_T(u)) \exp \underline{\varsigma}_T d\Phi(u) \xrightarrow{p} 0$. By the Cauchy–Schwarz inequality

$$\left[\int E_\delta(1 - \mathcal{S}_T(\delta)\mathcal{A}_T(u)) \exp \underline{\varsigma}_T(u, \delta) d\Phi(u)\right]^2 \leq \left[\int E_\delta(1 - \mathcal{S}_T(\delta)\mathcal{A}_T(u))^2 d\Phi(u)\right]\left[\int E_\delta \exp 2\underline{\varsigma}_T(u, \delta) d\Phi(u)\right].$$

From the same reasoning as above, $\int E_\delta \exp 2\underline{\varsigma}_T(u, \delta) d\Phi(u) = O_p(1)$, and

$$\int E_\delta(1 - \mathcal{S}_T(\delta)\mathcal{A}_T(u)) d\Phi(u) \leq \int E_\delta(1 - \mathcal{S}_T(\delta)) d\Phi(u) + \int E_\delta(1 - \mathcal{A}_T(u)) d\Phi(u).$$

But $\int E_\delta(1 - \mathcal{S}_T(\delta)) d\Phi(u) = E_\delta(1 - \mathcal{S}_T(\delta)) = E_\delta \mathbf{1}[T^{1/2} \sup_{t \leq T} ||\delta_t|| \geq T^\eta] \to 0$, and $\int E_\delta(1 - \mathcal{A}_T(u)) d\Phi(u) \leq \int \mathbf{1}[||u|| \geq \alpha_T] d\Phi(u) \xrightarrow{p} 0$ since $||\hat{u}|| = O_p(1)$, $\tilde{C}_T^{-1/2} = O_p(1)$ and $\alpha_T \to \infty$.

(iv) Similar to the proof of part (iii) and omitted for brevity.

A.2.3. Additional lemmas.

Lemma 3. *Under Condition 1:*
(i) $T^{-1/2} \sum_{t=1}^{[\cdot T]} s_t(\theta_0) \Rightarrow \int_0^\cdot \Gamma^{1/2}(l) dW(l)$, *where* W *is a* $k \times 1$ *standard Wiener process*
(ii) $\sup_{t \leq T, \{v_t\}_{t=1}^T \in \mathcal{B}_T^T, \{\tilde{v}_t\}_{t=1}^T \in \mathcal{B}_T^T} T^{-1}|| \sum_{s=1}^t (2\int_0^1 \lambda h_s(\theta_0 + v_s + \lambda \tilde{v}_s) d\lambda - \Gamma_s)|| \xrightarrow{p} 0$ *and* $\sup_{t \leq T, \{v_t\}_{t=1}^T \in \mathcal{B}_T^T, \{\tilde{v}_t\}_{t=1}^T \in \mathcal{B}_T^T}$
$T^{-1}|| \sum_{s=1}^t (\int_0^1 h_s(\theta_0 + \lambda(v_s - \tilde{v}_s)) d\lambda - \Gamma_s)|| \xrightarrow{p} 0$, *where* $B_T = \{\theta : ||\theta - \theta_0|| < b_T\}$ *with* $b_T \to 0$, *and* $\mathcal{B}_T^T = \mathcal{B}_T$
$\times \cdots \times \mathcal{B}_T$
(iii) $\hat{u} = T^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$
(iv) $T^{-1/2} \sum_{t=1}^{[\cdot T]} s_t(\hat{\theta}) \Rightarrow \int_0^\cdot \Gamma(l)^{1/2} dW(l) - \int_0^\cdot \Gamma(l) dl (\int_0^1 \Gamma(l) dl)^{-1} \int_0^1 \Gamma(l)^{1/2} dW(l)$
(v) $\sup_{\lambda \in [0,1]} ||T^{-1} \sum_{t=1}^{[\lambda T]} s_t(\hat{\theta}) s_t(\hat{\theta})' - \int_0^\lambda \Gamma(l) dl|| \xrightarrow{p} 0$ *and* $T^{-1} \sum s_t(\theta_0) s_t(\theta_0)' = O_p(1)$
(vi) $\sup_{\lambda \in [0,1]} ||T^{-1} \sum_{t=1}^{[\lambda T]} h_t(\hat{\theta}) - \int_0^\lambda \Gamma(l) dl|| \xrightarrow{p} 0$.

Lemma 4. *Under Conditions 1 and 2, there exists a sequence of real numbers* α_T *with* $\alpha_T \to \infty$ *and* $T^{-1/2} \alpha_T \to 0$ *such that*
(i) $\int w(\theta_0 + T^{-1/2} u) E_\delta(1 - \mathcal{A}_T(u)\mathcal{S}_T(\delta)) LR_T(u, \delta) du \xrightarrow{p} 0$
(ii) $E_\delta(1 - \mathcal{S}_T(\delta)) LR_T(0, \delta - \mathbf{e}\overline{\delta}) \xrightarrow{p} 0$.

Lemma 5. *Let* $\Sigma_\Xi(u)$ *be a* $Tk \times Tk$ *matrix consisting of* $k \times k$ *blocks* $\Xi_{i,j}(u)$, $i, j = 1, \ldots, T$, *possibly dependent on* u *and define* $c_T^U = \sup_{i,j \leq T, u \in \mathbb{R}^k} ||\Xi_{i,j}(u)||$. *Under Condition 2, there exists a constant* c_G *independent of* u *and* T *such that*
(i) $|\text{tr}((F^{-1} \Sigma_\delta F'^{-1}) \Sigma_\Xi(u))| \leq c_T^U c_G$
(ii) $|\text{tr}((F^{-1} \Sigma_\delta F'^{-1}) \Sigma_\Xi(u)(F^{-1} \Sigma_\delta F'^{-1}) \Sigma_\Xi(u))| \leq (c_T^U)^2 c_G^2$.

Lemma 6. *Under Condition 1:*
(i) There exists a sequence of random variables $\tilde{C}_T = O_p(1)$ satisfying $\tilde{C}_T^{-1} = O_p(1)$ such that

$$\sup_{v \in \mathbb{R}^k, T} (E_\delta \exp[-2(\delta - T^{-1/2}\mathbf{e}v)'D_{\tilde{h}}(\delta - T^{-1/2}\mathbf{e}v)] - \exp[-\tfrac{1}{2}\tilde{C}_T||v||^2]) \leq 0.$$

(ii) Suppose the $k \times 1$ vectors ξ_t satisfy $\sup_{t \leq T} ||T^{-1/2}\sum_{s=1}^{t}\xi_s|| \xrightarrow{p} 0$, the $k \times k$ matrix functions $\zeta_t : \mathbb{R}^k \mapsto \mathbb{R}^{k \times k}$ satisfy $\sup_{t \leq T, u \in \mathbb{R}^k} ||T^{-1}\sum_{s=1}^{t}\zeta_s(u)|| \xrightarrow{p} 0$, the $k \times k$ matrices Ξ_{it} satisfy $\sup_{t \leq T} ||T^{-1}\sum_{s=1}^{t}\Xi_{is}|| \xrightarrow{p} 0$, $i = 1, 2, 3$. Then, with $\boldsymbol{\xi} = (\xi_1', \ldots, \xi_T')'$, $D_\zeta(u) = \mathrm{diag}(\zeta_1(u), \ldots, \zeta_T(u))$, $\Xi_i = (\Xi_{i1}', \ldots, \Xi_{iT}')'$, $i = 1, 2$ and $D_\Xi = \mathrm{diag}(\Xi_{31}, \ldots, \Xi_{3T})$

$$\underline{\kappa}_T \exp[\underline{\Delta}_T||v||^2] \leq E_\delta \exp\left[\boldsymbol{\xi}'\boldsymbol{\delta} + T^{-1/2}v'\mathbf{e}'D_\zeta(u)\boldsymbol{\delta} - \tfrac{1}{2}\boldsymbol{\delta}'(T^{-1}\Xi_1\Xi_2' + D_\Xi)\boldsymbol{\delta}\right] \leq \overline{\kappa}_T \exp[\overline{\Delta}_T||v||^2]$$

uniformly in v and T, where the scalar random variables $\underline{\kappa}_T$, $\underline{\Delta}_T$, $\overline{\kappa}_T$ and $\overline{\Delta}_T$ do not depend on u or v and $\underline{\kappa}_T \xrightarrow{p} 1$, $\overline{\kappa}_T \xrightarrow{p} 1$, $\underline{\Delta}_T \xrightarrow{p} 0$ and $\overline{\Delta}_T \xrightarrow{p} 0$.
(iii) If $T^{-1/2}\sum_{t=1}^{[\cdot T]}\hat{s}_t \Rightarrow S_L(\cdot)$, then $E_\delta \exp[4\hat{s}'\boldsymbol{\delta}] = O_p(1)$.
(iv) If $J_T \in \mathcal{D}$ is a nonstochastic sequence converging to $J \in \mathcal{D}$, where $\mathcal{D}$ is the set of cadlag functions on the unit interval, then

$$\sup_T E_\delta \exp[T^{1/2}J_T(1)'(\delta_T - \overline{\delta}) - T^{1/2}\sum J_T((t-1)/T)'(\delta_t - \delta_{t-1})] < \infty$$

with $\delta_0 = 0$.

Lemma 7. *Under Conditions 1 and 2,*

$$E_\delta \overline{LR}_T(\delta) \Rightarrow E_G \exp\left[\int_0^1 G^*(s)\Gamma(s)^{1/2}dW(s) - \tfrac{1}{2}\int_0^1 G^*(s)'\Gamma(s)G^*(s)ds\right]$$

where $G^*(s) = G(s) - (\int_0^1 \Gamma(\lambda)d\lambda)^{-1}\int_0^1 \Gamma(\lambda)G(\lambda)d\lambda$.

REFERENCES

ANDERSON, T. W. (1955), "The Integral of a Symmetric Convex Set and some Probability Integrals", *Proceedings of the American Mathematical Society*, **6**, 170–176.

ANDREWS, D. W. K. and PLOBERGER, W. (1994), "Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative", *Econometrica*, **62**, 1383–1414.

BAI, J. (1997), "Estimation of a Change Point in Multiple Regressions", *Review of Economics and Statistics*, **79**, 551–563.

BAI, J. and PERRON, P. (1998), "Estimating and Testing Linear Models with Multiple Structural Changes", *Econometrica*, **66**, 47–78.

BERNANKE, B. S. and MIHOV, I. (1998), "Measuring Monetary Policy", *The Quarterly Journal of Economics*, **113**, 869–902.

BILLINGSLEY, P. (1968), "Convergence of Probability Measure" (New York: Wiley).

BOIVIN, J. (2003), "Has U.S. Monetary Policy Changed? Evidence from Drifting Coefficients and Real Time Data" (Working Paper, Columbia University).

BROWN, L. D. and LOW, M. G. (1996), "Asymptotic Equivalence of Nonparametric Regression and White Noise", *Annals of Statistics*, **24**, 2384–2398.

CAI, Z. (2007), "Trending Time-Varying Coefficient Time Series Models with Serially Correlated Errors", *Journal of Econometrics*, **136**, 163–188.

CHIB, S. (1998), "Estimation and Comparison of Multiple Change-Point Models", *Journal of Econometrics*, **75**, 221–241.

COGLEY, T. and SARGENT, T. J. (2005), "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII US", *Review of Economic Dynamics*, **8**, 262–302.

DUDLEY, R. M. (2002), *Real Analysis and Probability* (Cambridge, UK: Cambridge University Press).

DUFOUR, J.-M. and GHYSELS, E. (1996), "Recent Developments in the Econometrics of Structural Change: Overview", *Journal of Econometrics*, **70**, 1–8.

DURBIN, J. and KOOPMAN, S. J. (1997), "Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State Space Models", *Biometrika*, **84**, 669–684.

DURBIN, J. and KOOPMAN, S. J. (2001), "Time Series Analysis by State Space Methods" (Oxford: Oxford University Press).

ELLIOTT, G. and MÜLLER, U. K. (2006), "Efficient Tests for General Persistent Time Variation in Regression Coefficients", *Review of Economic Studies*, **73**, 907–940.

ELLIOTT, G. and MÜLLER, U. K. (2007), "Confidence Sets for the Date of a Single Break in Linear Time Series Regressions", *Journal of Econometrics*, **141**, 1196–1218.

FERNANDEZ-VILLAVERDE, J. and RUBIO-RAMIREZ, J. (2007), "How Structural Are Structural Parameters?", *Macroeconomics Annual*, **22**, 83–137.

GHYSELS, E. (1998), "On Stable Factor Structures in the Pricing of Risk: Do Time-Varying Betas Help or Hurt?", *Journal of Finance*, **53**, 549–573.

GOURIEROUX, C., MONFORT, A. and TROGNON, A. (1984), "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, **52**, 681–700.

HALL, P. and HEYDE, C. C. (1980), *Martingale Limit Theory and its Applications* (New York: Academic Press).

HAMILTON, J. D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle", *Econometrica*, **57**, 357–384.

HANSEN, B. E. (1996), "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis", *Econometrica*, **64**, 413–430.

HARVEY, A. C. (1989), "Forecasting, Structural Time Series Models and the Kalman Filter" (Cambridge University Press).

HARVEY, A. C., RUIZ, E. and SHEPHARD, N. (1994), "Multivariate Stochastic Variance Models", *Review of Economic Studies*, **61**, 247–264.

HUBER, P. (1967), "The Behavior of the Maximum Likelihood Estimates under Nonstandard Conditions", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Berkeley: University of California Press) 221–233.

JACQUIER, E., POLSON, N. G. and ROSSI, P. E. (1994), "Bayesian Analysis of Stochastic Volatility Models", *Journal of Business and Economic Statistics*, **12**, 371–417.

JONG, P. D. (1991), "The Diffuse Kalman Filter", *The Annals of Statistics*, **19**, 1073–1083.

KIM, S., SHEPHARD, N. and CHIB, S. (1998), "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models", *Review of Economic Studies*, **65**, 361–393.

LECAM, L. (1986), "Asymptotic Methods in Statistical Decision Theory" (New York: Springer Verlag).

LI, H. and MÜLLER, U. K. (2009), "Valid Inference in Partially Unstable General Method of Moment Models", *Review of Economic Studies*, **76**, 343–365.

LINDE, J. (2001), "Testing for the Lucas-Critique: A Quantitative Investigation", *American Economic Review*, **91**, 986–1005.

MÜLLER, U. K. (2009), "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix" (Working Paper, Princeton University).

NEWEY, W. K. and WEST, K. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, **55**, 703–708.

NUSSBAUM, M. (1996), "Asymptotic Equivalence of Density Estimation and Gaussian White Noise", *Annals of Statistics*, **24**, 2399–2430.

NYBLOM, J. (1989), "Testing for the Constancy of Parameters Over Time", *Journal of the American Statistical Association*, **84**, 223–230.

PHILLIPS, P. C. B. and PLOBERGER, W. (1996), "An Asymptotic Theory of Bayesian Inference for Time Series", *Econometrica*, **64**, 381–412.

PLOBERGER, W. (2004), "A Complete Class of Tests When the Likelihood is Locally Asymptotically Quadratic", *Journal of Econometrics*, **118**, 67–94.

POLLARD, D. (2001), "Contiguity" (Working Paper, Yale University).

PRIESTLEY, M. B. and CHAO, M. T. (1972), "Non-Parametric Function Fitting", *Journal of the Royal Statistical Society, Series B*, **34**, 385–392.

PRIMICERI, G. E. (2005), "Time Varying Structural Vector Autoregressions and Monetary Policy", *The Review of Economic Studies*, **72**, 821–852.

ROBINSON, P. M. (1989), "Nonparametric Estimation of Time-Varying Parameters", in P. Hackl (ed.) *Statistical Analysis and Forecasting of Economic Structural Change* (Berlin: Springer) 253–264.

ROBINSON, P. M. (1991), "Time-Varying Nonlinear Regression", in P. Hackl and A. H. Westlund (eds) *Economic Structural Change. Analysis and Forecasting* (Berlin: Springer) 179–190.

SCHERVISH, M. J. (1995), "Theory of Statistics" (New York: Springer).

SHEPHARD, N. and PITT, M. K. (1997), "Likelihood Analysis of Non-Gaussian Measurement Time Series", *Biometrika*, **84**, 653–667.

SHIVELY, T. S. (1988), "An Analysis of Tests for Regression Coefficient Stability", *Journal of Econometrics*, **39**, 367–386.

SIMS, C. A. and ZHA, T. (2006), "Where There Regime Switches in Us Monetary Policy?", *American Economic Review*, **96**, 54–81.

STOCK, J. H. (1994), "Unit Roots, Structural Breaks and Trends", in R. F. Engle and D. McFadden (eds) *Handbook of Econometrics*, Vol. 4 (New York: North Holland) 2740–2841.

STOCK, J. H. and WATSON, M. W. (1996), "Evidence on Structural Instability in Macroeconomic Time Series Relations", *Journal of Business and Economic Statistics*, **14**, 11–30.

STOCK, J. H. and WATSON, M. W. (1998), "Median Unbiased Estimation of Coefficient Variance in a Time-Varying Parameter Model", *Journal of the American Statistical Association*, **93**, 349–358.

STOCK, J. H. and WATSON, M. W. (2002), "Has the Business Cycle Changed and Why?" in M. Gertler and K. S. Rogoff (eds) *NBER Macroeconomics Annual 2002* (Cambridge, MA: MIT Press) 159–218.

WALD, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large", *Transactions of the American Mathematical Society*, **54**, 426–482.

WHITE, H. (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, **50**, 1–25.

WU, W. B. and ZHAO, Z. (2007), "Inference of Trends in Time Series", *Journal of the Royal Statistical Society, Series B*, **69**, 391–410.