Bypassing the Curse of Dimensionality: Feasible Multivariate Density Estimation

Minsu Chang*
Georgetown University

Paul Sangrey
Amazon

July 20, 2019

Abstract

Most economic data are multivariate making estimating multivariate densities a classic problem in the literature. However, given vector-valued data — $\{x_t\}_{t=1}^T$ — the *curse of dimensionality* makes nonparametrically estimating the data's density infeasible when the number of series, D, is large. Hence, we do not seek to provide estimators that perform well all of the time (it is impossible), but rather seek to provide estimators that perform well most of the time. We adapt the ideas in the Bayesian compression literature to density estimation by randomly binning the data. The binning randomly determines both the number of bins and which observation is placed in which bin. This novel procedure induces a simple mixture representation for the data's density. For any finite number of periods, T, the number of mixture components used is random. We construct a bound for this variable as a function of T that holds with high probability. We adopt the nonparametric Bayesian framework and construct a computationally efficient density estimator using Dirichlet processes. Since the number of mixture components is the key determinant of our model's complexity, our estimator's convergence rates — $\sqrt{\log(T)}/\sqrt{T}$ in the unconditional case and $\log(T)/\sqrt{T}$ in the conditional case — depend on D only through the constant term. We then analyze our estimators' performance in a monthly macroeconomic panel. Our procedure performs well in capturing the data's stylized features such as time-varying volatility and skewness.

Keywords: Curse of Dimensionality, Bayesian Nonparametrics, Big Data, Markov Process, Density Forecasting, Gaussian Mixtures, Transition Density

JEL Codes: C11, C14, C32, C55

*M. Chang: Department of Economics, Georgetown University, Washington, DC 20007. Email: minsu.chang@georgetown.edu (Chang), Paul Sangrey: Amazon, Seattle, WA 98121. Email: paul@sangrey.io (Sangrey). We are indebted to our advisors, Francis X. Diebold, Jesús Fernández-Villaverde, and Frank Schorfheide. We have also benefited greatly from conversations with Karun Adusumilli, Ben Connault, Frank DiTraglia, Laura Liu, Andriy Norets and seminar participants at the University of Pennsylvania and the 2018 NBER-NSF SBIES conference at Stanford University. All remaining errors are our own.

1 Introduction

Estimating multivariate densities is a classic problem across econometrics, statistics, and computer science. Researchers often find parametric assumptions restrictive and their models sensitive to deviations from these assumptions. On the other hand, given vector-valued data — $\{x_t\}_{t=1}^T$ — nonparametrically estimating the data's density is infeasible when the number of series, D, is large. This phenomenon is called the *curse of dimensionality*.

Nonparametric estimators simultaneously solve two problems. First, they approximate the density. Second, they estimate the parameters that govern this approximation. The original curse of dimensionality papers, such as Stone (1980), examine the approximation problem through the lens of the deterministic approximation literature. They show that requiring the estimators to be consistent causes the estimator and the deterministic approximation to use the same number of terms asymptotically. Solving this deterministic problem requires $T^{g(D)}$ terms for some g that depends upon the set of functions under consideration. To understand this, consider creating a multidimensional histogram. Dividing a D-dimensional hypercube into small hypercubes with width $1/T$ requires T^D terms. The various deterministic approximations essentially form these high-dimensional histograms asymptotically. The precise form of the "histogram" being formed depends on the application. In the years since Stone (1980), the minimax estimation literature has focused on mapping various applications to these high-dimensional histograms, e.g., Yang and Barron (1999) and Ichimura and Todd (2007).

Over the same period, various other authors have studied how random approximations behave in high dimensions,e.g., Johnson and Lindenstrauss (1984); Klartag and Mendelson (2005); Boucheron, Lugosi, and Massart (2013) and Talagrand (2014). This Bayesian compression literature develops parsimonious random approximations to high-dimensional datasets. Since high-dimensional random variables cluster on balls instead of hypercubes, the question is how should we approximate high-dimensional balls, not high-dimensional hypercubes. (We provide intuition below on both why random data tends to cluster on balls and why this dramatically simplifies the problem.) Thus far, the Bayesian compression literature has focused on the approximation problem and the closely related data compression problem. For example, Koop, Korobilis, and Pettenuzzo (2019) compress hundreds of variables and compute vector autoregressions on the compressed data. However, unlike the deterministic approximation case, no one has yet applied these ideas to density estimation. We apply these ideas and develop parsimonious high-dimensional approximations to feasibly estimate multivariate densities.

In particular, we develop a dynamic generalization of the infinite-mixture representation commonly used in the Bayesian nonparametric literature, (Ghosal and van der Vaart (2017)), as an alternative to current Bayesian conditional density estimators, e.g., Geweke and Keane (2007), Norets (2010); and Pati, Dunson, and Tokdar (2013). Infinite mixtures are commonly used to flexibly approximate cross-sectional densities, (Ghosal, Ghosh, and van der Vaart (2000) and van der Vaart and van Zanten (2008)). Because infinite-mixtures can approximate a broad class of densities, this procedure only requires a few assumptions on the data generating process (DGP). We can estimate both

unconditional and transition densities for both i.i.d. and Markov data.

We apply the results from the Bayesian compression literature to nonparametric density estimation in a series of steps. First, we construct a novel method for approximating high-dimensional balls that bins the data and endogenously determines both the number of bins and which vector — x_t — goes into which bin. Second, we show that this random binning induces an approximating mixture representation that is close to the true density.

For any finite T, we construct a bound for the number of mixture components as a function of T that holds with high probability. It is impossible to create a nonparametric estimator that is always parsimonious. This probability is with respect to the data-agnostic procedure that determines the number of mixture components. We convert these bounds on the mixture's complexity into convergence rates for the estimators. Our estimators' convergences rates — $\sqrt{\log(T)}/\sqrt{T}$ in the unconditional case and $\log(T)/\sqrt{T}$ in the conditional case — depend on D only through the constant term.

To summarize, we show that our estimator converges rapidly — it does not require many mixture components even when D is large — with arbitrarily high probability. We do this by tolerating a small chance of our estimator's converging slowly. Even though we cannot beat the minimax rate in general, we show that our estimators will perform well even when D is large and the true distribution is not smooth. In particular, we show that distance between the induced mixture representation and the data's true distribution, as measured by standard divergences, such as Hellinger and Kullback-Leibler, is small even when we take supremum over the set of true DGPs and D is large.

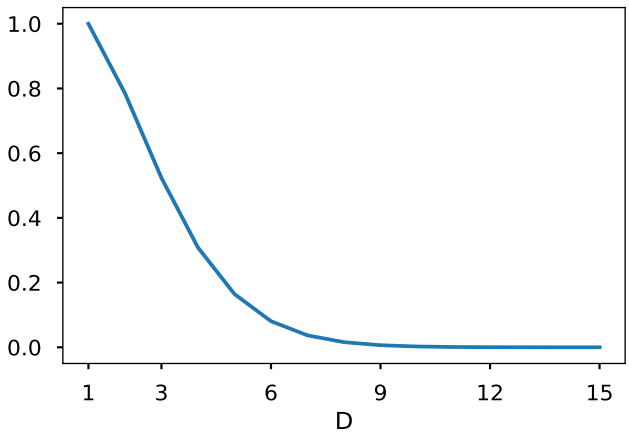
We organize the paper as follows. Section 2 provides the intuition underlying the results in the Bayesian compression literature, and consequently our results. Section 3 describes the data generating process. Section 4 constructs the sieve and provides conditions under which it approximates the true density well. Section 5 proves our estimators converge at the rates given above. Section 6 provides a computationally efficient Gibbs sampling algorithm to estimate our sieve. Section 7 introduces the data and the prior we use for the empirical analysis. Section 8 uses our method to empirically analyze a monthly macroeconomic panel showing our method works well in practice. Section 9 concludes. The appendices contain the proofs and additional empirical results.

2 Intuition

The convergence rates discussed above likely seem surprising, so we now explain why they are reasonable. We do this by discussing the intuition that drives the results in the Bayesian compression literature. As discussed above, the standard convergence rates are consequences of the number of bins of width $1/T$ required to fill a D-dimensional hypercube equaling T^D. The Bayesian compression algorithms use fewer terms than the deterministic approximations do by exploiting two facts. First, random data tend to cluster in balls. Second, the volume of a D-dimensional ball grows exponentially slower with the number of series than the hypercube does as shown in Figure 1. We exploit this

behavior by constructing a sieve for the D-dimensional ball instead of constructing a sieve for the D-dimensional hypercube as the literature usually does. Since the volume of the ball grows more slowly, our sieve requires far fewer terms, especially when D is large.

Figure 1: Volume of a Ball Relative to a Hypercube [1]



The goal in this paper is to exploit this simplicity to bound the number of terms required to estimate a density, instead of just compressing the data. In other words, similar to how Stone (1980) created a sieve for the hypercube, we construct a sieve for the ball. Previous methods have shown how to compress the data while only slightly perturbing the data's first two sample moments. We construct a sparse discretization operator (i.e., we bin the data) that does not significantly perturb the data's first two sample moments. To convert this distance between the sample moments into a distance between densities, we use the fact that as long as a process is locally asymptotically normal, its first two moments form a sufficient statistic for the density. Consequently, if the first two moments are close, the densities must be close as well.

We build a Dirichlet mixture process and adopt the standard Bayesian mixture framework. The number of mixture components determines the complexity of a Gaussian mixture and the estimator's convergence rate. For any fixed sample-size T, this is a random variable. Hence, we have a series of distributions indexed by T. As mentioned above, we can view the distance between the estimator and the truth itself as a random variable whose distribution is indexed by T. The critical difference between our results and the previous ones in the literature is that we only require the convergence rate to hold in a $1-2\delta$ probability region with respect to the prior. In other words, we want our estimator to converge rapidly "most of the time" where "most" means with probability at least $1-2\delta$ and this probability is only taken with respect to the prior. We still require the convergence rate to be uniform with respect to the likelihood.

Because the previous literature requires the convergence rate to be uniform with respect to randomness in the prior, they cannot exploit the smoothness that the prior induces in deriving their

[1] The ratio between the volume of a ball of hypercube with the same diameter equals $\frac{\pi^{\frac{D}{2}}}{2^D\Gamma\left(\frac{D}{2}+1\right)}$, where Γ denotes the Gamma function.

convergence rates. At a technical level, for any fixed T our sieve is not a measurable function of the data and so the Stone (1980) bounds do not apply.

3 Data Generating Process

Consider a D-dimensional time series: $X_T \coloneqq \{x_t\}_{t=1}^T$. Assume that X_T is first-order hidden Markov and is a Gaussian process. That is, there exists a latent state z_t, such that (x_t, z_t) are jointly Markov. The z_t may be a constant. We want to estimate x_t's conditional densities for $t = 1, \ldots, T$. Let $\mathcal{F}_t$ denote the time-t information set. We denote the true distribution P_T and the approximating distribution Q_T. They have associated densities p_T and q_T.

Definition 1 (Data Generating Process)**.**

$$p_T\left(x_t \,|\, \mathcal{F}_{t-1}\right) \coloneqq \sum_{k=1}^{\infty} \Pi_{t-1,k} \phi\left(x_t \,|\, x_{t-1}\beta_{k,t}, \Sigma_{k,t}\right). \tag{1}$$

Since X_T is Gaussian process, the conditional density — $p_T(x_t \,|\, F_{t-1})$ — has an infinite Gaussian mixture representation for each time period. Each mixture component has associated mixture probability, $\Pi_{t-1,k}$, and component-specific parameters, $\beta_{k,t}$, and $\Sigma_{k,t}$. We assume that all x_t have finite means and variances. The Gaussian process assumption is quite general allowing, for example, for both multiple modes and fat tails. We let the true DGP depend upon T because at this point we are only approximating the density for a fixed T. Of course, the Markov implies a consistency condition across p_T for all T.

Definition 2 (Approximating Model)**.**

$$q_T\left(x_t \,|\, \mathcal{F}_{t-1}\right) \coloneqq \sum_{k=1}^{K_T} \Pi_{t-1,k} \phi\left(x_t \,|\, x_{t-1}\beta_k, \Sigma_k\right). \tag{2}$$

The approximating model is a Gaussian mixture with K_T components, and so K_T governs the complexity of the model. As one would expect, K_T grows with T. Second, each cluster's components, (β_k, Σ_k), no longer have time t subscripts. The idea is that we can reuse the latent variables $(\beta_{k,t}, \Sigma_{k,t})$ across time without loss of generality. If two separate periods have sufficiently similar dynamics, we group them into one component with the same parameters. Since the clusters are defined differently in the true and approximating models, no simple relationship the parameters exists in general.

Throughout, we use μ_T to refer to the $T \times D$ mean matrix. We also consider the rescaled data:

$$\widetilde{X}_T \coloneqq \frac{X_T - \mu_T}{\sqrt{\|X_T - \mu_T\|_{L_2}}} \in S^{TD-1} = \left\{x \in \mathbb{R}^{TD} \,\middle|\, \|X\|_{L_2} = 1\right\}, \tag{3}$$

where $\|\cdot\|_{L_2}$ is the L_2-norm. Since we are on the unit hypersphere, we are in a compact space for any fixed T. Since $X_T - \mu_T$ is a zero-mean Gaussian process, its $TD \times TD$ covariance matrix completely

determines its distribution. We define the densities of $\widetilde{X}_T$ as we did for X_T above and denote them $\widetilde{p}_T$ and $\widetilde{q}_T$.

4 Sieve Construction

4.1 Setting up the Problem

We construct a sieve in this section that approximates a wide variety of data generating processes while still being as simple as possible. By simple, we mean that the metric entropy of these approximating models grows slowly with the number of datapoints. This property is useful because metric entropy controls the rate at which posteriors converge as shown by Ghosal, Ghosh, and van der Vaart (2000) and Shen and Wasserman (2001). It also controls the minimax rate at which estimators can converge, (Wong and Shen (1995) and Yang and Barron (1999)).

We approximate a marginal density in the space of densities over $\mathbb{R}^D$ — $\mathscr{P}(\mathbb{R}^D)$ — and a transition density which lies in associated the product space — $\mathscr{P}(\mathbb{R}^D) \times \mathscr{P}(\mathbb{R}^D)$. These approximations problems are not well-posed because multiple equivalent representations exist for each density given X_T that satisfy some bound on the distance to p_T in some metric on $\mathscr{P}\left(R^D\right)$. We can exploit this multiplicity by choosing a representation that is particularly amenable to estimation for each T. We want the most parsimonious density that still approximates well.

We construct our sieve as follows. Given some $\epsilon > 0$, we construct a mapping Θ_T that takes the TD-dimensional hypersphere and maps it onto a KD-dimensional hypersphere, where $K \ll T$. This mapping only perturbs the norms of the individual elements by at most ϵ. In other words, we construct an ϵ-isometry.

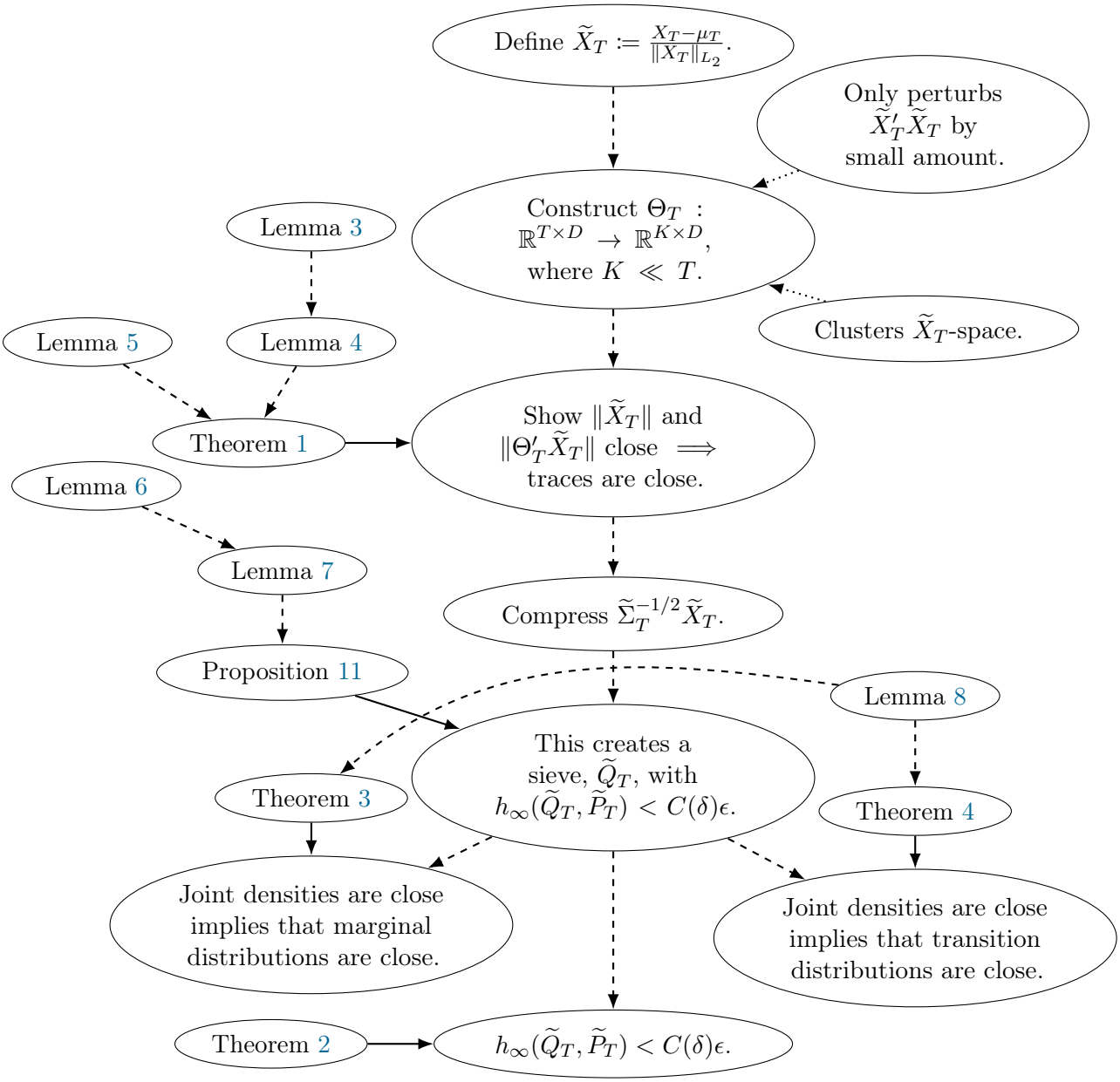
We then show the densities are also not perturbed significantly in Theorem 2. This result is true whenever the norm of the data matrix is a locally sufficient statistic for the density. In other words, we can use bounds on divergences between the norms, $\{\|\widetilde{x}_t\|_{L_2}\}$, to bound divergences between the densities in $\mathscr{P}(\widetilde{X}_T)$.

4.2 Bounding the Norm Perturbation

We construct our approximate sufficient statistic for $\widetilde{X}_T$ by "projecting" it onto a lower-dimensional space. The only reason this projection intuition is not exact is that the target space is not a subspace of the original space. We need the compressed data to have a mixture distribution. Hence, the compression operator Θ_T must be a discretization operator. A mixture distribution for some collection of data $\widetilde{X}_T$ is a random binning of the data where the data in each bin has the same parametric distribution. The question is how to construct the bins.

A standard discretization operator with K bins is a $T \times K$ matrix where each row θ_t contains exactly one 1 and the rest of the elements equal zero. A variable x_t is in bin k if and only if $\theta_{t,k} = 1$, i.e., Θ_T has a 1 in row t column k. We cannot use a standard discretization operator for two reasons. First, since all of the elements are weakly positive $\mathbb{E}[\theta_{t,k}] \neq 0$. Second, once we see a 1, the rest of

Figure 2: Sieve Construction[2]



the columns in the row must be identically zero. This property makes the columns too dependent for our results to hold.

Fixing the first issue is relatively straightforward. We let $\theta_{t,k}$ take on values from $\{-1, 0, 1\}$. Each x_t is in bin k if $\theta_{t,k} = 1$ and in bin $K + k$ if $\theta_{t,k} = -1$. There is no reason the elements of θ must be positive. The second issue is more problematic. We let each row have as many 1's and -1's as necessary. Once we do this, seeing a 1 in column k gives us no information about columns $k + 1$

[2]Because our construction of the sieve requires a number of steps, keeping track of the various dependencies between the various results is non-trivial. Consequently, we provide this graphic to display the dependencies between the various lemmas and theorems. Dashed lines are dependencies, solid lines are labels, and dotted lines are comments.

through K. It does complicate the analysis slightly, however. We are letting each period be in more than one component simultaneously. In other words, we do not just create a mixture distribution across periods but also create one in each period.

To make the discussion in the previous few paragraphs more formal, we now define the random operator Θ_T. We use a stick-breaking process to construct Θ_T, adapting the form often used to construct Dirichlet processes, as proposed by Sethuraman (1994).[3]

Definition 3 (Θ_T Operator)**.** Let b be a Bernoulli random variable with $\Pr(b = 1) \in (0, 1)$. Draw another random variable $\chi \in \{-1, 1\}$ with probability 1/2 each. Let $T \in \mathbb{N}$ be given. Draw T variables $\theta \coloneqq \chi \cdot b$ independently of all of the previous values, and form them into a column-vector — Θ_1. Form another column vector Θ_2 the same way and append it to the right of Θ_T. Continue this until all of the rows of Θ_T contain at least one nonzero element.

We form the Θ_T operator this way so that $\mathbb{E}[\theta] = 0$ and $\mathbb{V}\text{ar}(\theta) = \mathbb{E}[|\theta|] = \Pr(b = 1)$. Furthermore, its rows are independent and its columns form a martingale-difference sequence. The only dependence between the columns of Θ_T arises through the stopping rule, and stopped martingales are still martingales. In addition, Θ_T is independent of $\widetilde{X}_T$. Since Θ_T is discrete, Θ_T implicitly clusters $\widetilde{X}_T$. Consider some row θ_t of Θ_T. For each column of θ_t, define a bin as $|\theta_{t,k}| \times \text{sign}(\theta_{t,k})$. Clearly, if Θ_T has K_T columns, there are $2K_T$ possible total bins.

Our analysis requires a tight bound on the tail behavior of K_T. To create such a bound, we must understand its distribution. By Lemma 3, the probability density function of K_T is

$$\Pr(K_T \leq \widetilde{K}) \propto (1 - (1 - \Pr(b = 1))^{\widetilde{K}})^T. \tag{4}$$

Furthermore, we show in Lemma 4 that $K_T \propto \log(T)$ with high probability. The intuition behind this is that to get $K = \widetilde{K}$ the Bernoulli random variable must have $\widetilde{K}$ failures. The probability of this occurring declines exponentially fast in $\widetilde{K}$. This logarithmic growth is relied upon extensively in what follows.

We claimed above that Θ_T constructs an approximate sufficient statistic by binning $\widetilde{X}_T$. In other words, we are compressing the data. Equation (4) quantifies the amount by which we compress the data. Instead of considering each of the T values of x_t separately, we can bin them into K_T bins, and we can treat the values in each bin identically. Since $K_T \propto \log(T) \ll T$ this substantially reduces the complexity.

We also must show that Θ_T preserves the $\widetilde{x}_t$'s densities. It is not a sufficient statistic if we lose any necessary information. We turn to this now.

Theorem 1 (Bounding the Norm Perturbation)**.** *Let Θ_T be constructed as in Definition 3 with the number of columns denoted by K_T. Let $\epsilon > 0$ be given. Let $0 < \delta < 1$ be given such that $0 < \log(\frac{1}{\delta}) < c_1\epsilon^2 K_T$ for some constant c_1. Let $\widetilde{X}_T$ be in the unit hypersphere in $\mathbb{R}^{TD-1}$. Then*

[3] In Section 4.7, we show that a Dirichlet process can replace Θ_T without affecting our results. The intuition behind this is that we can construct both of them using similar stick-breaking processes. Consequently, they are mutually absolutely continuous, and so we a density that converts the measures exits.

with probability greater than $1-2\delta$ *with respect to* Θ_T*, there exists a constant* c_2 *such that for any* $\epsilon > \sqrt{\frac{\log T}{K_T}}$ *,*

$$\sup_t \left| \|\theta_t \widetilde{x}_t\|_{L_2} - \|\widetilde{x}_t\|_{L_2} \right| < c_2 \left(1 + \log\left(\frac{1}{\delta}\right)\right) \epsilon.$$

Theorem 1 implies that if we choose Θ_T with the number of columns $K \propto \log(T)$, applying Θ_t perturbs the norms of $\widetilde{x}_t$ by at most ϵ. This result holds with probability at least $1-2\delta$ with respect to the distribution over Θ_T. Since $\widetilde{X}_T \in S^{TD-1}$, we can map S^{TD-1} onto a smaller space $S^{K_T D-1}$, with $K \ll T$, without perturbing the individual elements' norms significantly.

The basic idea here is that we are pre-multiplying the data by a martingale, i.e., a process which expectation equal to one. This does not affect the mean or the variance. This increases randomness "smooths" the data. To gain intuition, one can think about the average value. Koop, Korobilis, and Pettenuzzo (2019) do precisely this, focusing on Bayesian model averaging. This allows us to get very tight bounds on the tails of the distribution with high probability. Since we have not changed the first two population moments and can tightly bound the tails of the distribution, we can place strong bounds on how we moved the sample moments. This is precisely what Theorem 1 does.

4.3 Distances on the Space of Densities

In the previous section, we showed that Θ_T does not affect $\widetilde{x}_t$'s norms significantly. These norms are not inherently interesting objects. Instead, they are interesting because they form a sufficient statistic for the Gaussian process. To show the densities are close, we must convert the distances between the norms into distances on $\mathscr{P}(\widetilde{X}_T)$.

The compressed data, $\Theta_T' \widetilde{X}_T$, has a distribution conditional on Θ_T. Since $\widetilde{X}_T$ is a normalized Gaussian process and Θ_T is a matrix, this process is Gaussian. Hence, there exists a distribution for $\widetilde{X}_T$ constructed by integrating out Θ_T. This integration creates an approximating distribution for $\widetilde{X}_T$: $\widetilde{Q}_T$.

Since Θ_T is almost surely discrete, this approximating distribution is a mixture, as in Definition 2. We represent it as an integral with respect to a latent mixing measure — G_t^Q — for each t. The parameters in each component are means and covariances, and so the G_t^Q measure is over the space of means and covariances. Because Θ_T can have more than one non-zero element, G_t^Q is a mixture distribution in each period, even conditional on Θ_T.

Let G^Q be the latent mixing measure over the space of G_t^Q. That is, each G_t^Q is a draw from G^Q. Since latent mixing measures are almost surely discrete, the G_t^Q share the same atoms. This dependence regularizes the mixing measures across time, i.e., it "smooths" the approximating model. However, since the atoms of G^Q are left arbitrary, it does not restrict the set of DGPs that we can approximate well.

Let δ_t^Q denote the mixture identity that determines which cluster contains Σ_t. Let $\phi(\cdot \,|\, \delta_t^Q)$ denote the mean-zero multivariate Gaussian density with covariance Σ_t. Then $\widetilde{Q}_T$ can be expressed as

$$\widetilde{q}_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\, dG^Q(dG_t^Q). \tag{5}$$

Likewise, if we replace q with p, we write the true model's density, $\widetilde{p}_T$, as

$$\widetilde{p}_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(\widetilde{x}_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P)\, dG^P\left(dG_t^P\right), \tag{6}$$

with its associated latent mixing measures and mixture identities. Note, The approximating cluster identities, $\{\delta^Q{}_t\}_{t=1}^T$, are different than the true cluster identities, $\{\delta^P{}_t\}_{t=1}^T$, because Θ_T induces Q's clustering. It is not induced by the underlying true clustering.

Since the densities are mixtures parameterized by their covariances, we must convert a clustering in $\widetilde{x}_t$-space into a clustering in Σ_t-space in order to convert the bounds above into bounds on the densities. The norms of $\widetilde{x}_t$ and $\widetilde{x}_{t^*}$ being close does not imply that the associated matrix norms for Σ_t and Σ_{t^*} are close. Consequently, we cluster $\Sigma_t^{-1/2}\widetilde{x}_t$ directly.

The error bound Theorem 1 provides is independent of δ_t^P and so it does not depend on Σ_t. In other words, for times t, t^* such that the associated $\widetilde{x}_t$ and $\widetilde{x}_{t^*}$ are contained in the same cluster, δ_k^Q, the following holds:[4]

$$\sup_{t,t^* \in \delta_k^Q} \left|\widetilde{x}_t \Sigma_t^{-1} \widetilde{x}_t - \widetilde{x}'_{t^*} \Sigma_{t^*}^{-1} \widetilde{x}_{t^*}\right| < \epsilon. \tag{7}$$

Here ϵ is independent of t, t^*, and the cluster identity. The right-hand side of (7) is a "distance" on the space of covariance matrices. That is, we introduce the following semimetric on the space of covariance matrices.[5]

Definition 4 (Weighted-L_2 Semimetric)**.**

$$\delta_{wl_2}(\Sigma_k, \Omega_k) \coloneqq \sup_{t,t^* \in \delta_k^Q} \left|\widetilde{x}'_t \Sigma_k^{-1} \widetilde{x}_t - \widetilde{x}'_{t^*} \Omega_k^{-1} \widetilde{x}_{t^*}\right|. \tag{8}$$

Note, δ_{wl_2} is compatible with, and weaker than, the max-norm.[6] The max-norm is equivalent to the L_2-norm up to a scale transformation, and the relevant scale is a constant since we only consider full-rank matrices. Hence, the space of covariance matrices forms a Polish space because the space of $D \times D$ matrices is isomorphic to $\mathbb{R}^{D \times D}$ and we are choosing an open subset of that space. In other words, δ_{wl_2} constructs a set of equivalence classes over the space of covariance matrices, where two sample covariances are equivalent if the implied second-moment behavior of the $\{\widetilde{x}_t \in \delta_k^Q\}$ is indistinguishable.

Definition 4 converts bounds on the norms into $\widetilde{x}_t$ into bounds on covariances. We must convert this bound to a bound on densities. The distance we use here is the Hellinger distance.

[4] We abuse notation slightly and use $t \in \delta_k^Q$ if the cluster identity associated with x_t equals δ_k^Q.

[5] It is a semimetric because we can have $\Sigma \neq \Omega$ but $\delta_{wl_2}(\Sigma, \Omega) = 0$. The two matrices may differ that cannot be identified by the set $x \in$ cluster k.

[6] If x, y in $x\Sigma^{-1}y$ are (possibly) different unit selection vectors we can pick out the maximum absolute deviation between elements in the two matrices. This difference is clearly at least as big as the δ_{wl_2} because that semimetric requires x, y to be the same.

Definition 5. Hellinger Distance

$$h(p,q) \coloneqq \frac{1}{\sqrt{2}}\sqrt{\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx}. \tag{9}$$

The Hellinger distance is useful because it is a valid norm on the space of densities. Since the covariance matrix is a sufficient statistic for a centered Gaussian process, we can convert bounds between the covariances into bounds in this distance. Instead of applying this directly to the joint distribution, we take the supremum over the conditional distributions.

Definition 6 (Supremum Hellinger Distance)**.**

$$h_\infty^2(p,q) \coloneqq \sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q, 1\le t\le T} h^2\left(p\left(\cdot \,\middle|\, \mathcal{F}_{t-1}^P\right), q\left(\cdot \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right). \tag{10}$$

The supremum Hellinger distance will prove useful because it is stronger than both the Hellinger distance and the Kullback-Leibler divergence applied to the joint density. As a consequence, once we bound h_∞, we can directly deduce other bounds as necessary.

4.4 Representing the Joint Density

We now show that the approximating distribution of $\widetilde{X}_T$ induced by Θ_T is close to the true distribution $\widetilde{P}_T$ using h_∞. We can do this whenever the rescaled trace is a locally sufficient statistic for the density. Hence, we can use bounds on divergences in $\widetilde{\mathcal{X}}$ to bound divergences in $\mathcal{P}(\widetilde{\mathcal{X}})$.

Theorem 2 (Representing the Joint Density)**.** *Let $\widetilde{X}_T \coloneqq \widetilde{x}_1, \ldots, \widetilde{x}_T$ be a D-dimensional Gaussian process with period-t finite stochastic means, μ_t, and covariances, Σ_t. Let Σ_t be positive-definite for all t. Let Θ_T be the generalized selection matrix constructed in Definition 3. Let $\widetilde{P}_T$ denote the distribution of $\widetilde{X}_T$. Then given $\epsilon > 0$ and $\delta \in (0,1)$, the approximating distribution, Q_T, which is the mixture distribution over $\widetilde{\mathcal{X}}$ that Θ_T induces, satisfies the following with probability at least $1 - 2\delta$ with respect to Θ_T for some constant C:*

$$h_\infty\left(\widetilde{P}_T(\widetilde{\mathcal{X}}), \widetilde{Q}_T(\widetilde{\mathcal{X}})\right) < C\left(1 + \log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

We represent the joint density as follows. Since $\widetilde{\mathcal{X}}$ lives in S^{TD-1}, we start by mapping S^{TD-1} onto a smaller space $S^{K_T D-1}$ where $K_T \ll T$. This argument is very similar to the various projection arguments that the literature makes when it projects S^{TD-1} into a "smaller" space. However, the operator Θ_T we use does not form a projection because it is not mapping the space onto itself. The unit sphere in $\mathbb{R}^{K_T D}$ is not a subset of the one in $\mathbb{R}^{TD}$.

Unlike the previous compression operators in the literature, Θ_T is discrete, and so it clusters $\widetilde{x}_t$. This property implies that the density of $\widetilde{x}_t$ is a process with respect to a discrete measure. That is, Q_T is a mixture distribution. Also, we show in Section 4.7, that we can assume that this latent measure is Dirichlet without loss of generality. In other words, our method represents the

$\widetilde{X}_T$ process as an integral with respect to a Dirichlet process. Consequently, since $\widetilde{X}_T$ is a Gaussian process, and hence locally Gaussian, we can represent $\widetilde{X}_T$ using a Gaussian mixture process whose latent mixing measure is a Dirichlet process.

The leading issue that remains is that we bounded the rescaled $\widetilde{X}_T$, not X_T. As one might expect, estimating the true joint density of X_T is impossible. Since $\|X_T\|^2 \propto T$, the bound we have is of the order $\sqrt{T}\epsilon$, which is useless. Instead, we consider simpler quantities such as X_T's marginal density (Section 4.5) and transition density (Section 4.6). We show that sample means of the marginal and transition densities converge to those implied by Q_T, and hence those implied by P_T. This convergence occurs because sample means converge to population means.

4.5 Representing the Marginal Density

We now derive a representation for the marginal density of X_T from the representation for the joint density. We first consider the case where the true density has a product form, i.e., the data are independent. The intuition behind the proof is that Theorem 2 implies that $T\epsilon^2$ bounds the maximum deviation of the approximating density. Standard arguments about the convergence of means for product measures give a $\frac{1}{T}$ term. Hence, the deviation between the the the means is bounded by ϵ^2. We use the Hellinger distance here instead of the sup-Hellinger distance because there is no conditioning information we need to take the supremum over.

Theorem 3 (Representing the Marginal Density)**.** *Let $x_1, \ldots, x_T$ be drawn independently from P_T where each x_t has a infinite-Gaussian mixture representation. Let Θ_T be constructed as in Theorem 2 for each t. Let ϵ be given. Construct Q_T by using the Θ_T operator to group the data and letting the data be Gaussian distributed within each component with component-wise means and covariances given by their conditional expectations. Then, with probability $1 - 2\delta$ with respect to Θ_T, there exists a constant C such that the following holds uniformly over T*

$$h\left(\int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t(\delta_t^Q)\right) < C\left(1 + \log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

We now extend Theorem 3 to the non-i.i.d. case. The hidden Markov assumption implies that the transitions are conditionally i.i.d. and this conditioning does not affect the convergence rate because we have a supremum-norm bound on the deviations in the joint density. Uniform ergodicity implies that the sample marginal density converges to the true marginal density. Consequently, using hidden Markov data instead of independent data does not affect the approximation results.

Corollary 3.1 (Representing the Marginal Density with Markov Data)**.** *Theorem 3 continues to hold when the x_t form a uniformly ergodic hidden Markov chain instead of being fully independent.*

4.6 Representing the Transition Density

We now show our model approximates transition densities well. Since the data are Markov, we can construct the sample transition density as an average of the transitions in the data. Component by

component, we solve for the correct conditional distributions in the approximating model. Similar to above, we relate the error in the transition densities and the error for the joint densities. We can consider the space of transitions as the product space: $\widetilde{X}_T \otimes \widetilde{X}_T$. We can construct the marginal density in the space. As in Section 4.5, the approximate product form gives us a $1/T$ term in the convergence rate. Theorem 2 gives us a $T\epsilon^2$ term. The T terms cancel, and so ϵ^2 bounds the distance between the densities.

Theorem 4 (Transition Density Representation)**.** *Let $x_1, \ldots, x_T \in R^{T \times D}$ be a uniformly ergodic Markov Gaussian process with density p_T. Let $\epsilon > 0$ be given. Let $K \geq c \log(T)^2/\epsilon$ for some constant c. Let δ_t be the cluster identity at time t. Then there exists a mixture density q_T with K clusters with the following form:*

$$q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}\right) \coloneqq \sum_{k=1}^{K} \phi\left(\beta_k x_{t-1}, \Sigma_k\right) \Pr\left(\delta_t = k \,|\, \delta_{t-1}\right).$$

Construct $q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^{Q}\right)$ from $q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}\right)$ by integrating out δ_{t-1} using $\Pr(\delta_{t-1} \,|\, X_T)$. Then with probability $1 - 2\delta$ with respect to the prior

$$h_\infty\left(p_T\left(x_t \,|\, \mathcal{F}_{t-1}^{P}\right), q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^{Q}\right)\right) < C\sqrt{1 + \log\left(\frac{1}{\delta}\right)}\epsilon.$$

4.7 Replacing Θ_T with a Dirichlet Process

The previous subsections use Θ_T to construct an approximating representation that is arbitrarily close to the truth. We want to construct an estimator that takes this representation to the data. (We do not claim that the representation is unique.) Here we argue that Θ_T can be chosen to be a Dirichlet process without loss of generality.

Consider the Θ_T process as in Definition 3 except we no longer stop when we no longer need columns. Then we can replace Θ_T with a Dirichlet process without altering the results. By doing this we can use standard Dirichlet-based samplers to estimate the sieve. In particular, it shows that the nonparametric Bayesian marginal density estimators in the literature satisfy the requirements of our theory, (Ghosal, Ghosh, and van der Vaart (2000) and Walker (2007)).

Lemma 1 (Replacing Θ_T with a Dirichlet Process)**.** *Let Q be a mixture distribution representable as an integral with respect to the Θ_T process defined in Definition 2. Then Q has a mixture representation as an integral with respect to the Dirichlet process.*

The intuition behind Lemma 1 is as follows. Theorem 2 shows that we can represent the density as an integral with respect to the random measure generated by Θ_T with probability $1 - 2\delta$. In other words, there exists a subset Θ_T space with Pr(that subset) $= (1 - 2\delta)$ such that the representation above holds. Since each realization Θ'_T in Θ'_T-space is a consistent sequence of categorical random variables, we can extend the probability space for these realizations by using a Dirichlet process. Intuitively, we are placing a Dirichlet prior on these categorical random variables.

To use the same notation we used to construct Q_T, we can view G_t^Q as a draw from G^Q and assume that both processes are Dirichlet, i.e., we are using a hierarchical Dirichlet process. Again by using the normalized completely random measure property of Dirichlet processes, this implies that the implied prior for the transition densities is Dirichlet.

5 Bayesian Nonparametrics and Convergence Rates

5.1 Problem Setup

We now use the sieve and associated bounds constructed in the previous section to derive the convergence rates of the associated estimators. We adopt a standard Bayesian nonparametric framework and show how fast the posteriors contract to the true model.

We start by recalling this setup. We assume the data $\{x_t\}_{t=1}^T$ are drawn from some distribution P_T which is parameterized $P_T(\cdot \,|\, \xi)$, for $\xi \in \Xi$. This parameter set is equipped with the Borel σ-algebra $\mathscr{B}$ with associated prior distribution $\mathcal{Q}_0(\xi)$. We further assume that there exists a regular conditional distribution of X_T given ξ — $P(X_T \,|\, \xi)$ — on the sample space $(\mathcal{X}, \mathscr{X})$. This implicitly defines a joint distribution over $(\mathcal{X} \times \Xi, \mathscr{X} \times \mathscr{B})$:

$$\Pr(X_T \in A, \xi \in B) = \int_B \Pr\left(A \,|\, \xi\right) d\mathcal{Q}_0(\xi). \tag{11}$$

Under some technical conditions, we can define a regular version of the conditional distribution of ξ given X_T, i.e., a Markov kernel from $(\mathcal{X}, \mathscr{X})$ into $(\Xi, \mathscr{B})$, which is called the posterior.

Definition 7. Posterior Distribution

$$\mathcal{Q}_T\left(B \,|\, X_T\right) \coloneqq \Pr\left(\{\xi \in B\} \,|\, X_T\right), B \in \mathscr{B}. \tag{12}$$

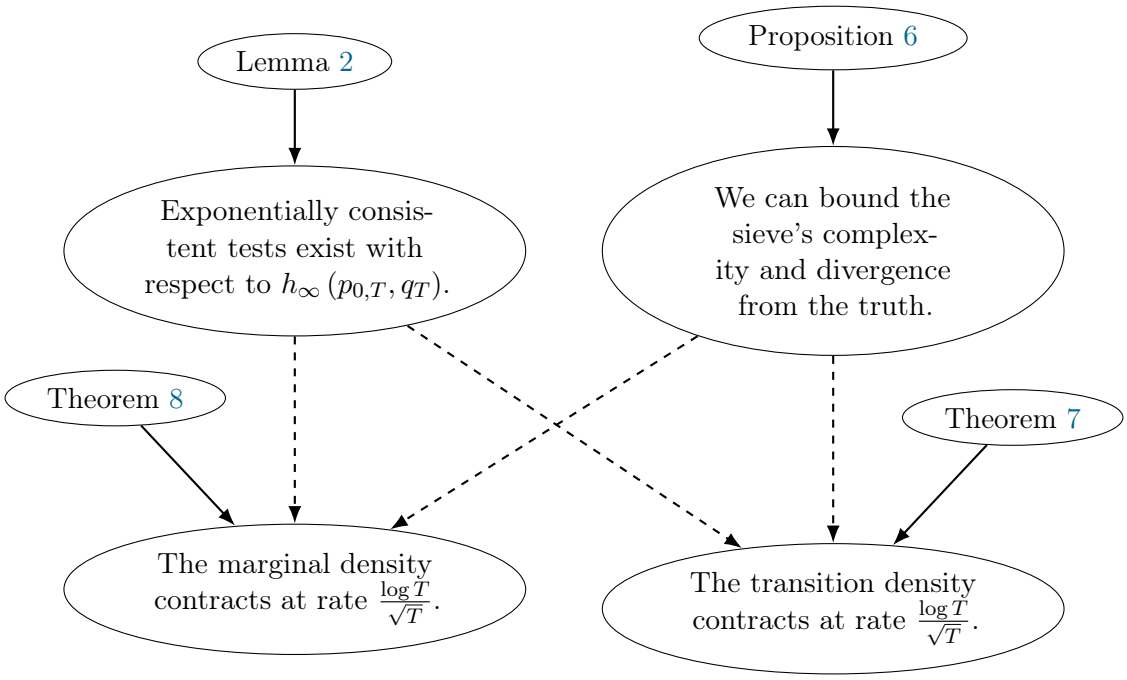
Posterior contraction rates characterize the speed at which the posterior distribution become close to the true value in a distributional sense. They are useful for two reasons. First, it puts upper bound on the convergence rate of point estimators such as the mean. Second, it tells you the speed at which inference using the estimated posterior distribution becomes valid. Our definition of this rate comes from (Ghosal and van der Vaart, 2017, Theorem 8.2).

Definition 8. Contraction Rate A sequence ϵ_T is a *posterior contraction rate* at parameter ξ^P with respect to the semimetric d if $\mathcal{Q}_T\left(\{\xi \,|\, d(\xi^P, \xi) \geq M_T \epsilon_T\} \,|\, X_T\right) \to 0$ in $P_T\left(X_T \,|\, \xi^P\right)$-probability for every $M_T \to \infty$.

To bound the asymptotic behavior of ϵ_T, we must simultaneously bound two separate quantities. First, we must show that our approximating model is close to the true density in the appropriate distance. We did this in the previous section. Second, we must bound the complexity (entropy) of our model, showing that it does not grow too rapidly.

We start by defining some notation that we use in deriving our theorems for the contraction rates. The concepts we use here are standard in the Bayesian nonparametrics literature. First, we

Figure 3: Contraction Rates[7]



define the metric (Kolmogorov) entropy for some small distance ϵ, some set Ξ, and some semimetrics, d_T and e_T. (One can, of course, use the same semimetric for both d_T and e_T.)

Definition 9. Metric Entropy $N(C\epsilon, d_T(\xi, \xi^P), e_T)$ is the function whose value for $\epsilon > 0$ is the minimum number of balls of radius $C\epsilon$ with respect to the d_T semimetric (i.e., d_T-balls of radius $C\epsilon$) needed to cover an e_T-ball of radius ϵ around the true parameter ξ^P.

The logarithm of this number — the *Le Cam Dimension* — is the relevant measure of the model's complexity, and hence the "size" of the sieve, and controls the minimax rate under some technical conditions. We define a ball with respect to the minimum of the Kullback-Leibler divergence and some related divergence measures. We also adopt the following two concepts used in Ghosal and van der Vaart (2007).

First, $V_{k,0}$ is "essentially" the k$^{\text{th}}$-centered moment of the Kullback-Leibler divergence between two densities f, g, and associated distributions F, G:

$$V_{k,0}(f, g) \coloneqq \int \left|\log(f/g) - \mathrm{D}_{\mathrm{KL}}\left(f \,||\, g\right)\right|^k dF. \tag{13}$$

Having defined $V_{k,0}(f, g)$, we define the relevant balls. $f_T(X \,|\, \xi)$ is the density of the length T data sequence X_T associated with parameter ξ. The ball is defined thus:

$$B_T(\xi^P, \epsilon, k) \coloneqq \left\{ \xi \in \Xi \,\middle|\, \begin{array}{l} \mathrm{D}_{\mathrm{KL}}\left(f\left(X_T \,|\, \xi^P\right) \,||\, f\left(X_T \,|\, \xi\right)\right) \leq T\epsilon^2, \\ V_{k,0}\left(f\left(X_T \,|\, \xi^P\right), f\left(X_T \,|\, \xi^P\right)\right) \leq T\epsilon^2 \end{array} \right\}. \tag{14}$$

[7]This graph displays the dependencies between the various lemmas and theorems. Dashed lines are dependencies, solid lines are labels.

We now quote (Ghosal and van der Vaart, 2007, Theoerem 1). This theorem provides general conditions for convergence of posterior distributions even if the data are not i.i.d.. It extends the results in Ghosal, Ghosh, and van der Vaart (2000), which is the most common way to derive convergence rates in the literature, to cover dependent data.

Theorem 5 (Ghosal and van der Vaart (2007) Theorem 1)**.** *Let d_T and e_T be semimetrics on Ξ. Let $\epsilon_T > 0, \epsilon_T \to 0$, $\left(\frac{1}{T\epsilon^2}\right)^{-1} \in O(1)$. $C_1 > 1$, $\Xi_T \in \Xi$ be such that for sufficient large $n \in \mathbb{N}$.*

1. *There exist exponentially consistent tests Υ_T as in Lemma 2 with respect to d_T.*

2. $$\sup_{\epsilon_T > \epsilon} \log N\left(\frac{C_2}{2}\epsilon, \left\{\xi \in \Xi_T \,\middle|\, d_T(\xi, \xi^P) \leq \epsilon\right\}, e_T\right) \leq T\epsilon_T^2 \tag{15}$$

3. $$\frac{\mathcal{Q}_T\left(\left\{\xi \in \Xi_T \,\middle|\, n\epsilon_T < d_T(\xi, \xi^P) \leq 2n\epsilon_T\right\} \,\middle|\, X\right)}{\mathcal{Q}_T\left(B_T(\xi^P, \epsilon_T, C_1) \,|\, X\right)} \leq \exp\left(\frac{C_2 T\epsilon_T^2 n^2}{2}\right) \tag{16}$$

Then for every $M_T \to \infty$, we have that

$$P_T\left(\mathcal{Q}_T\left(\left\{\xi \in \Xi_T \,\middle|\, d_T(\xi, \xi^P) \geq M_T\epsilon_T\right\} \,\middle|\, X\right) \,\middle|\, \xi^P\right) \to 0. \tag{17}$$

5.2 Contraction Rates

We now show that uniformly consistent tests exist with respect to the semimetric that we use: h_∞. This metric is stronger than the average squared Hellinger distance, which is usually used in the Bayesian nonparametric estimation of Markov transition densities, ((Ghosal and van der Vaart, 2017, 542)).

Note, h_∞^2 should be interpreted as a distance on the joint distributions because we can always factor a joint distribution as

$$f(X_T) = f\left(x_T \,|\, \mathcal{F}_{T-1}\right) \cdot f\left(x_{T-1} \,|\, \mathcal{F}_{T-2}\right) \cdots f\left(x_2 \,|\, \mathcal{F}_1\right) \cdot f\left(x_1 \,|\, \mathcal{F}_0\right), \tag{18}$$

where $\mathcal{F}_0$ denotes information that is always known, as is standard.

It is worth noting that h_∞^2 is a function of T even though we suppress it in the notation. We are only considering deviations between the densities over length-T sequences. The first goal is to show that consistent tests exit to separate two distributions in h_∞^2. To do this, we provide the following lemma.

Lemma 2 (Exponentially consistent tests exist with respect to h_∞)**.** *There exist tests Υ_T and universal constants $C_2 > 0$, $C_3 > 0$ satisfying for every $\epsilon > 0$ and each $\xi_1 \in \Xi$ and true parameter ξ^P with $h_\infty(\xi_1, \xi^P)$:*

1. $$P_T\left(\Upsilon_T \,\middle|\, \xi^P\right) \leq \exp(-C_2 T\epsilon^2) \tag{19}$$

2. $$\sup_{\xi \in \Xi,\, e_n(\xi_1, \xi) < \epsilon C_3} P_T\left(1 - \Upsilon_T \,\middle|\, \xi^P\right) \leq \exp(-C_2 T\epsilon^2) \tag{20}$$

Having shown there exist the appropriate tests, we now show that (15) and (16) hold. As noted in (Ghosal and van der Vaart, 2007, 197), the numerator is trivially bounded by 1, as long as $T\epsilon_T \to \infty$ which it does in this case. We do this by proving a proposition that covers both the marginal and transition dentin cases. We can deduce the mains theorems as results of it.

Proposition 6 (Bounding the Posterior Divergence). *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T := \sum_k \Pi_{k,t}\phi(x_t \mid \mu_t, \Sigma_t)$ with finite means and finite positive-definite covariances. Let $\Xi_T \subset \Xi$ and $T \to \infty$. Let Q_T be a mixture approximation with $\frac{K_T^i}{\eta_T}$ components. Assume the following condition holds with probability $1 - 2\delta$ for $\delta > 0$ and constants C and $i \in \mathbb{N}$:*

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_T\left(x_t \,|\, \mathcal{F}_{t-1}^P\right)\right) < C\eta_T. \tag{21}$$

Let $\epsilon_{i,T} := \frac{\log(T)^{\sqrt{i}}}{\sqrt{T}}$. Then the following two conditions hold with probability $1 - 2\delta$ with respect to the prior

$$\sup_{\epsilon_i \geq \epsilon_{T,i}} \log N\left((\epsilon_i, \left\{\xi \in \Xi_T \,|\, h_\infty(\xi, \xi^P) \leq \epsilon_i\right\}, h_\infty\right) \leq T\epsilon_{T,i}^2, \tag{22}$$

and

$$\mathcal{Q}_T\left(B_T\left(\xi^P, \epsilon_{T,i}, 2\right) \,\middle|\, X_T\right) \geq C\exp\left(-C_0 T\epsilon_{T,i}^2\right). \tag{23}$$

We can apply Proposition 6 to the transition density by taking $i = 2$. We use the representation for the transition density we proved in Theorem 4. As a consequence, by Theorem 5, the following result holds.

Theorem 7 (Contraction Rate of the Transition Density). *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T := \sum_k \Pi_{t,k}\phi(x_t \mid \mu_t, \Sigma_t)$ with finite means and finite positive-definite covariances. Let $T \to \infty$, then the following holds with $\epsilon_T := \sqrt{\frac{\log(T)^2}{T}}$ with probability $1 - 2\delta$ with respect to the prior. There exists a constant C independent of T such that the posterior over the transition densities constructed above and the true transition density satisfies*

$$P_T\left(\mathcal{Q}_T\left(\sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} h\left(p_T\left(x_t \,|\, \mathcal{F}_{t-1}^P\right), q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right) \geq C\epsilon_T \,\middle|\, X_T\right)\right) \to 0.$$

We also bound for the convergence rate of the marginal density. This should not be too surprising. Estimating the Markov transition density with respect to h_∞ is strictly harder than estimating the marginal distribution. You can integrate out the marginal distribution by using the stationary distribution. (In this context, the stationary and marginal distributions are the same.) Also, since i.i.d. data is trivially a uniformly ergodic Markov process, we cover the i.i.d. case as well.

Theorem 8 (Contraction Rate of the Marginal Density). *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T := \sum_k \pi_k \phi(\cdot \mid \mu_t, \Sigma_t)$ with finite mean and finite variance. Let $T \to \infty$, then the following holds with $\epsilon_T = \sqrt{\frac{\log(T)}{T}}$ and probability $1 - 2\delta$ with respect to the prior.*

There exists a constant C independent of T such that the posterior over the transition densities constructed above and the true transition density satisfies

$$P_T\left(\mathcal{Q}_T\left(h\left(p_T\left(x_t\right), q_T\left(x_t\right)\right) \geq C\epsilon_T \,|\, X\right)\right) \to 0.$$

6 Estimation Strategy

We estimate our model using Bayesian methods. So far, the discussion has been rather abstract. We have focused on proving theoretical results about our estimation strategy. We now construct a Gibbs sampler to estimate the model. Recall the definition of the approximating model for the transition density:

$$q_T\left(x_t \,|\, \mathcal{F}_{t-1}\right) = \sum_{k=1}^{K_T} \Pi\left(k = \delta_t \,|\, \delta_{t-1}\right) \phi\left(\beta_k x_{t-1}, \Sigma_k\right). \tag{24}$$

We must place a prior on each of the components in this model. We start by using a Dirichlet process to place a prior on $\Pi_{t,k} \coloneqq \Pi(\delta_t = k \,|\, \delta_{t-1})$ and, hence, implicitly on K_T. We then construct priors for β_k and Σ_k.

A substantial literature exists on efficiently estimating Dirichlet mixture models, e.g., Ishwaran and James (2001); Papaspiliopoulos and Roberts (2008); and Griffin and Walker (2011). We use the slice sampler of Walker (2007) to handle the potentially infinite number of clusters without truncation and compute a valid upper bound for K_T. Conditional on K_T we draw the δ_t's marginal distribution. This is straightforward because (24) is almost a standard Gaussian mixture model.[8] We then update the transition matrix Π so it has the correct marginal distributions, and draw the $\{\delta_t\}_{t=1}^T$. Given $\delta_t = k$ and the hyperparameters, we apply standard Bayesian regression methods to obtain β_k and Σ_k. We use a conditionally conjugate hierarchical prior and draw from the hyperparameters' posterior. We present the procedure in Algorithm 1.

6.1 Posterior of $\{\delta_t\}_{t=1}^T$

6.1.1 Bounding K_T

In each period, the approximating model and implied marginal density are Dirichlet mixtures. Consequently, we draw the cluster identities by adapting algorithms from the literature. We are in the standard situation, except our prior varies over time.

Sampling Dirichlet mixtures is difficult for two reasons. First, the prior allows for infinitely many clusters, and so we cannot sum the probabilities and cannot compute the resulting marginal cluster probabilities. This inability arises because we cannot numerically solve the probability of cluster k: $\Pr(k) = 1 - \sum_{k^* \neq k} \Pr(k^*)$. All Dirichlet mixture models share this property and so several authors have developed ingenious ways to deal with this issue. We adopt the algorithm developed by Walker (2007) because this algorithm is exact (we do not need to truncate the distribution) and

[8] Conditional on δ_{t-1} it is.

Algorithm 1 Gibbs Sampler

1. **Posterior of** $\{\delta_t\}_{t=1}^T$
   (a) Use Walker (2007) to determine the number of clusters K_T.
   (b) Draw the new marginal probabilities, π, and update the transition matrix, Π.
   (c) Given K_T and $\{x_t\}_{t=1}^T$, use multinomial sampling to draw δ_t with
   $$\Pr(\delta_t = k) \propto \phi\left(x_t \,|\, \beta_k, \Sigma_k\right) \Pi_{t,k}.$$
2. **Posterior of** π
   (a) Estimate the posterior of Π conditional on $\{\delta_t\}_{t=1}^T$:
   $$\Pi_{k,j} = \frac{\mathcal{Q}_0(\delta_{t-1} = k)\mathcal{Q}_0(\delta_t = j) + \sum_{t=2}^T \mathbf{1}(\delta_{t-1} = k)\mathbf{1}(\delta_t = j)}{\mathcal{Q}_0(\delta_{t-1} = k) + \sum_{t=2}^T \mathbf{1}(\delta_{t-1} = k)}.$$ [9]
3. **Posterior of Component-Specific Parameters**
   (a) Given each cluster k, use Bayesian regression to draw $\{\beta_k, \Sigma_k\}$.
4. **Posterior of Hyperparameters**
   (a) Draw the hyperparameters governing $\{\beta_k, \Sigma_k\}$ from their conjugate posteriors.
5. **Iterate**

computationally efficient. He does this by introducing a random variable — u_t — so that, conditional on u_t, the distributions are available in closed form.

Given the cluster parameters, we can write the distribution of x_t as

$$q_T(x_t) = \sum_{k=1}^{\infty} \Pi_{t,k} \phi\left(x_t \,|\, \beta_k, \Sigma_k\right). \tag{25}$$

As mentioned above, we introduce a latent variable $u_t \sim U(0, \Pi_{t,k})$ so we can rewrite (25) as

$$q_T(x_t) = \sum_{k=1}^{\infty} \mathbf{1}\left(u_t < \Pi_{t,k}\right) \phi\left(x_t \,|\, \beta_k, \Sigma_k\right) = \sum_{k=1}^{\infty} \Pi_{t,k} U\left(u_t \,|\, 0, \Pi_{t,k}\right) \phi\left(x_t \,|\, \beta_k, \Sigma_k\right). \tag{26}$$

Consequently, with probability $\Pi_{t,k}$, x_t and u_t are independent, and so the marginal density for u_t is

$$\Pr\left(u_t \,\middle|\, \{\Pi_{t,k}\}_{k=1}^K\right) = \sum_{k=1}^{\infty} \Pi_{t,k} U\left(u_t \,|\, 0, \Pi_{t,k}\right) = \sum_{k=1}^{\infty} \mathbf{1}\left(u_t < \Pi_{t,k}\right). \tag{27}$$

Then we can condition on $\{u_t\}_{t=1}^T$ as a vector, but not on $\Pi_{t,k}$.

$$\Pr\left(\{v_k\}_{k=1}^K \,\middle|\, \{\delta_t\}_{t=1}^T\right) = \mathcal{Q}_0\left(\{v_k\}_{k=1}^K\right) \prod_{t=1}^T \mathbf{1}\left(v_{k=\delta_t} \prod_{\kappa<\delta_t}(1-v_\kappa) > u_{k=\delta_t}\right), \tag{28}$$

where the v_k are the sticks in the stick-breaking representation of the prior.

The dependence between the u_t does not affect (28) because the v_k do not depend upon t. Hence, the v_k are conditionally independent given $\{u_t\}_{t=1}^T$. Exploiting this independence and the stick-breaking representation of the prior, we can draw v_k from (28); it only shows up once in the product. By adopting the prior for the sticks implied by standard Dirichlet process — Beta$(1,\alpha)$, we use (28) to draw v_k. As shown by Papaspiliopoulos and Roberts (2008), this implies v_k are distributed:

$$v_k \sim \text{Beta}\left(1 + \sum_{t=1}^T \mathbf{1}(\delta_t = k), T - \sum_{\kappa=1}^k \sum_{t=1}^T \mathbf{1}(\delta_t = \kappa) + \alpha\right) \tag{29}$$

for $k = 0, 1, \ldots$. We only need to do this for the v_k where that $k \leq \max(\delta_t)$. These sticks are the only sticks at affect the likelihood. We can calculate the marginal cluster probabilities π_k:

$$\pi_k = v_k \prod_{\kappa=1}^k (1 - v_\kappa). \tag{30}$$

6.1.2 Correcting Π to have the Correct Marginal Distribution

If the data were i.i.d., we could convert the v_k into π_k, and then compute the set of possible δ_t. This step is precisely the references above use. However, the data are not i.i.d. because $\Pi_{t,k}$ depends on δ_{t-1}. The question at hand is how to transform the algorithm to update the marginal distribution in the presence of i.i.d. data into one that does not change the dependence structure in non-i.i.d. data.

We must construct a probability matrix such that the relationship between two clusters, k and k^*, remain the same as they did in the previous draw of the sampler, but the marginal distribution is updated appropriately. We know that Markov transition matrices and their associated marginal distributions have the following relationship for each cluster k:[10]

$$\pi_k = \sum_{j=1}^\infty \Pi_{k,j}\pi_j. \tag{31}$$

Let $\widetilde{\pi}$ be a new marginal distribution that is equivalent (in the measure-theoretic sense) to π. Define a transition matrix $\widetilde{\Pi}$ whose elements satisfy $\widetilde{\Pi}_{j,k} = \Pi_{j,k}\frac{\widetilde{\pi}_k}{\pi_k}\frac{\pi_j}{\widetilde{\pi}_j}$. We now show that $\widetilde{\pi}$ is the marginal

[10] This condition holding for all k is the standard condition that a stationary distribution is a left-eigenvector of the transition matrix.

distribution associated with $\widetilde{\Pi}$ by showing it satisfies (31).[11]

$$\widetilde{\pi}_k = \pi_k \frac{\widetilde{\pi}_k}{\pi_k} = \sum_{j=1}^{\infty} \Pi_{j,k} \pi_j \frac{\widetilde{\pi}_k}{\pi_k} = \sum_{j=1}^{\infty} \Pi_{j,k} \frac{\widetilde{\pi}_k}{\pi_k} \frac{\pi_j}{\widetilde{\pi}_j} \widetilde{\pi}_j = \widetilde{\pi}_k \sum_{j=1}^{\infty} \widetilde{\Pi}_{k,j} \widetilde{\pi}_j \tag{32}$$

We constructed a matrix $\widetilde{\Pi}$ that induces the correct marginal distributions. In doing this, we only changed the marginal distribution. The relative probabilities between different states has not been affected, i.e., for all states j, k, the relative probabilities $\Pi(k \,|\, \delta_t)/\Pi\pi(k \,|\, \delta_t) = \pi_k/\widetilde{\pi}_j$.

To run a Gibbs sampler, we view the operation in (32) as a draw from a conditional posterior. We condition on all but the first left eigenvector (the one associated with the eigenvector 1) of the transition matrix, Π and replace it with the one associated with $\widetilde{\Pi}$. We then calculate the resulting transition matrix. Transition matrices associated with irreducible Markov chains have exactly one stationary distribution, and that stationary distribution is the first left eigenvector. So this algorithm computes the unique new transition matrix.

6.1.3 Conditionally Drawing the $\{\delta_t\}_{t=1}^T$

If the new stationary distribution, $\widetilde{\pi}$, has more clusters than the previous one, π, did, we use the prior for Π to draw them. We do not have to transform them to have the appropriate dynamics because they contain no datapoints under Π, implying that π and $\tilde{\pi}$ coincide as they have the same prior.

From $\widetilde{\Pi}$ can compute $\Pi_{t,k}$ for each t by drawing the first cluster identity, δ_0 from the stationary distribution, and then using using the Markov property of δ_{t-1} for $t > 1$, and iterating forward. We can now compute $\{k \,|\, \Pi_{t,k} > u_t\}$ for each t. Then the posterior of δ_t is

$$\Pr\left(\delta_t = k \,|\, \Pi_{t,k}, u_t, x_t, \beta_k, \Sigma_k\right) \propto \mathbf{1}\left(k \in \{k \,|\, \Pi_{t,k} > u_t\}\right) \phi\left(x_t \,|\, \beta_k x_{t-1}, \Sigma_k\right). \tag{33}$$

This is a finite set with known probabilities, and the δ_t are categorical variables. These can sampled directly.

6.2 Posterior on the Transition Matrix

We place the Dirichlet process prior over these cluster identities in each period to allow for an arbitrary number of clusters. By stacking the Dirichlet processes over time, we obtain a Dirichlet process over the (δ_{t-1}, δ_t) product space. Intuitively, we are constructing the transition matrix, Π, as a Dirichlet-distributed infinite-dimensional square matrix as noted by Lin, Grimson, and Fisher (????).

Given the cluster identities, δ_t, which we drew in Section 6.1, we draw the transition matrices. We do this by noting that the prior probability of a transition is the product of the unconditional probabilities normalized appropriately. We can update this by counting the proportion of realized

[11] The multiplication and division in (32) is the scalar version.

transitions:

$$\Pi_{k,j} = \frac{\mathcal{Q}_0(\delta_{t-1} = k)\mathcal{Q}_0(\delta_t = j) + \sum_{t=2}^{T} \mathbf{1}(\delta_{t-1} = k)\mathbf{1}(\delta_t = j)}{\mathcal{Q}_0(\delta_{t-1} = k) + \sum_{t=2}^{T} \mathbf{1}(\delta_{t-1} = k)}.$$

Each element, $\Pi_{k,j}$, determines the probability of transitions in (δ_{t-1}, δ_t) and is updated by counting the number of transitions from k to j.

6.3 Identification Strategy and Cluster Labeling Problem

The other problem endemic to mixture models is that the cluster identities are not uniquely identified. In particular, we have a label switching problem. A model with clusters labeled 0 and 1 is the same model as one with those clusters labeled 1 and 0. This lack of uniqueness is particularly problematic in i.i.d. environments because there is no natural way to order the clusters.

In time series environments, like the one we consider here, we can label the clusters by when they first appear. The first period is always in cluster zero. The second cluster to arrive is always labeled cluster two. This labeling procedure has two nice features relative to the existing methods of labeling the clusters by their probability ordering. First, it imposes a strict order of the clusters. We have no ties, such as occur in probability-based labeling when two probabilities are equal. Second, the ordering is invariant to estimation uncertainty. We do not have to estimate which datapoint comes first in time, and so it is easy to maintain the same ordering across draws.

In order to enforce this identification restriction, we re-order the cluster identities immediately before returning the next posterior draw so that they always arrive in time order. This reordering does not solve the identification problem, but it does reduce the amount of multi-modality in our posterior.

6.4 Posterior for the Coefficient Parameters

The component-specific likelihood is given in Definition 10 where $X_k \coloneqq \{x_t \,|\, t \in \delta_k\}$, $Y_t \coloneqq \{x_t \,|\, t-1 \in \delta_k\}$, and T_k is the number of datapoints in cluster k. We are factoring the likelihood into the component specific components.

Definition 10. Component-Specific Likelihood

$$\{x_t\}_{t=1}^{T} \,|\, \{\delta_t\}_{t=1}^{T}, \{\beta_k, \Sigma_k\}_{k=1}^{K} \sim \prod_{k=1}^{K} \frac{|\Sigma_k|^{-T_k/2}}{(2\pi)^{T_k/2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left\{(Y_k - X_k\beta_k)\,\Sigma_k^{-1}\,(Y_k - X_k\beta_k)'\right\}\right),$$

We can factor the likelihood into component-specific parts, and estimate the parameters component by component. Because the components have varying amounts of data, we cannot assume that the number of datapoints in each of the components approaches infinity. Also, when we forecast, we sometimes must add more components. To do this effectively, we want to use all of the information the observed data gives us. We cannot condition on the data in the new component because there is none. Consequently, we specify a hierarchical model to pool information across components.

The first level is the standard Gaussian Inverse-Wishart prior that is conjugate to the prior specified in Definition 10.[12] The only difference is that we parameterize the innovation covariance distribution in terms of its mean: Ω.[13] If we need to add a new component during the course of the algorithm we draw from the distribution of β_k, Σ_k conditional on the $\bar{\beta}, U, \Omega, \mu_1$. We cannot condition on the data in the new component because none exists.

Definition 11. Component-Specific Parameters' Prior

$$\{\beta_k\}_{k=1}^K \,|\, \Sigma_k, \bar{\beta}, U \sim \mathcal{MN}\left(\bar{\beta}, \Sigma_k, U\right) \tag{34}$$

$$\{\Sigma_k\}_{k=1}^K \,|\, \Omega \sim \mathcal{W}^{-1}\left((\mu_1 - 2)\Omega, \mu_1 + D - 1\right) \tag{35}$$

This prior is the conjugate prior for the likelihood in Definition 10, and so we can use the standard formulas to estimate it. This gives the following marginal posterior for the Σ_k:

$$\begin{aligned}\Sigma_k \,|\, X_k, Y_k \sim & \mathcal{W}^{-1}\Bigg(\bar{\beta}' U^{-1}\bar{\beta} + (\mu_1 - 2)\Omega + Y_k' Y_k \\ & - \left(U^{-1}\bar{\beta} + X_k' Y_k\right)' \left(U^{-1} + X_k' X_k\right)^{-1} \left(U^{-1}\bar{\beta} + X_k' Y_k\right), \mu_1 + D - 1 + T_k\Bigg).\end{aligned} \tag{36}$$

We can also compute the following conditional posterior for β_k given Σ_k:

$$\bar{\beta}, \Sigma_k \,|\, X_k, Y_k \sim \mathcal{MN}\left(\left(U^{-1} + X_k' X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k' Y_k\right), \Sigma_k, \left(U^{-1} + X_k' X_k\right)^{-1}\right) \tag{37}$$

We now specify the prior and posterior for the hyperparameters. As is common in the literature, we draw $\bar{\beta}$ and U from their posteriors to control level of smoothing in a data-dependent way by placing prior distributions on the hyperparameters and estimating them. As we did above, we place a conjugate matrix-normal prior on the coefficient matrix and an Inverse-Wishart prior on the covariance matrix.

Definition 12. Coefficient Hyperparameters' Prior

$$\bar{\beta}, U \sim \mathcal{MN}(\beta^\dagger, \mathbb{I}_D, U)\mathcal{W}^{-1}(\Psi_U, \nu_U)$$

The product of the priors for β_k's given in (34) now behaves as the likelihood. Since we have Gaussian priors and likelihoods this a fairly standard posterior calculation. The only complication is

[12] Throughout, we use the parametric formulas given in the Wikipedia pages for the distribution. For example, the Matrix-Normal distribution is parameterized as it is at https://en.wikipedia.org/wiki/Matrix_normal_distribution.

[13] The scale parameter and the degrees of freedom parameter are chosen in the appropriate way to make Ω the mean matrix: $\mathbb{E}[\Sigma_k] = \text{Scale}/(\text{Degrees of freedom} - D - 1) = (\mu_1 - 2)\Omega/(\mu_1 + D - 1 - D - 1) = \Omega$.

that the $\{\beta_k\}_{k=1}^K$ are heteroskedastic.[14] Consequently, we provide the derivation in Appendix E.2:

$$U \,|\, \{\Sigma_k, \beta_k\}_{k=1}^K \sim \mathcal{W}^{-1}\left(\beta^\dagger\beta^{\dagger\prime} + \Psi_U + \sum_{k=1}^K \beta_k\Sigma_k^{-1}\beta_k' - \left(\beta^\dagger + \sum_{k=1}^K \beta_k\Sigma_k^{-1}\right)\left(\sum_{k=1}^K \Sigma_k^{-1} + \mathbb{I}_D\right)^{-1} \left(\beta^\dagger + \sum_{k=1}^K \beta_k\Sigma_k^{-1}\right)', \nu_U + (K+1)D\right), \tag{38}$$

and

$$\bar{\beta} \,|\, U, \{\Sigma_k, \beta_k\}_{k=1}^K \sim \mathcal{MN}\left(\left(\beta^\dagger + \sum_{k=1}^K \beta_k\Sigma_k^{-1}\right)\left(\sum_{k=1}^K \Sigma_k^{-1} + \mathbb{I}_D\right)^{-1}, \left(\sum_{k=1}^K \Sigma_k^{-1} + \mathbb{I}_D\right), U\right). \tag{39}$$

To draw Ω from its posterior, we adapt the hierarchical prior Huang and Wand (2013) construct. We deviate from them to allow Ω to have off-diagonal elements. Our covariance matrices are i.i.d. in expectation, but the prior for a new covariance matrix is not necessarily i.i.d. Also, the model of Huang and Wand (2013) does not necessarily have a density with respect to Lebesgue measure for the covariance matrix itself. We only allow for the hyperparameters to take on values where Σ_k's distribution has a both mean and density.

In particular, we parameterize the hierarchy for the Σ_k as follows. We have two degree of freedom parameters, μ_1 and μ_2, a mean matrix, $\Omega = \mathbb{E}[\Sigma_k]$, and D scale parameters for Ω: $a_1, \ldots, a_D$.

Definition 13 (Prior for the Covariances)**.**

$$\Omega \sim \mathcal{W}\left(\frac{\text{diag}(a_1, \ldots, a_D)}{\mu_2 + D - 1}, \mu_2 + D - 1\right)$$

If we send $\mu_2 \to \infty$, the implied prior for the prior for Ω becomes fully dogmatic. If $\nu_2 = 1/2$ and $D = 1$, the root diagonal elements — $\sqrt{(\Sigma_k)_{dd}}$ — have half-t distributions. In general, the $(\Sigma_k)_{dd}$ have appropriately scaled F-distributions.[15] If the off-diagonal elements of Ω almost surely equal to 0, the diagonal elements satisfy $(\Sigma_k)_{dd} \sim \Gamma^{-1}\left(\mu_1/2, (\frac{\mu_1}{2} - 1)\Omega_{dd}\right)$. This is why we let the number of degrees of freedom in (35) depend upon D. In general, the mean of these elements is the same, but the distribution is different since the off-diagonal elements of Ω affect the distribution of $(\Sigma_k)_{dd}$.

Obviously, conditional on Ω, everything is independent. The posterior distribution of Σ_k given

[14] They must be in order for the prior in (34) to be conjugate with its likelihood because the likelihood is heteroskedastic itself.

[15] $\sigma^2 \sim F(1, \mu_1 + D - 1) \implies \sigma \sim \text{half-}t(\mu_1 + D - 1)$. In the one dimensional case, $\mu_1 + D - 1_2 = 1/2$ implies that $\sigma^2 \sim F(1, \mu_1 + D - 1)$. This result is not feasible in the multivariate case while maintaining a density with respect to Lebesgue measure. If we let $\mu_1 \to 2$, we recover this expression. However, Ω is not well-defined in this case.

$\Omega, \{x_t \,|\, \delta_t = k\}$ is

$$\Omega \,|\, \{\Sigma_k\}_{k=1}^K \sim \mathcal{W}\Bigg(K(\mu_1 + D - 1) + (\mu_2 + D - 1), \\ \left(\text{diag}(a_1, \dots a_D)^{-1} + (\mu_1 - 2) \sum_{k=1}^K \Sigma_k^{-1} \right)^{-1} \Bigg). \tag{40}$$

As noted by Huang and Wand (2013), if Ω is almost surely diagonal, then the correlation parameters in Σ_k have a prior density of the form $p(\rho_{ij}) \propto (1 - \rho^{ij})^{\mu_1/2-1}$, $-1 < \rho_{ij} < 1$. Note, this implies that as $\mu_1 \to 2$, then the distribution of these off-diagonal elements approaches $U(-1, 1)$. Conversely, as $\mu_1 \to \infty$, the distribution of these off-diagonal elements converges to point masses at the off-diagonal elements of Ω. The off-diagonal elements of Ω are normal variance-mean mixtures where the mixing density is a χ^2-distribution, as is standard for Wishart priors.

7 Data and Prior

We downloaded monthly data on real consumption (DPCERAM1M225NBEA), the personal consumption expenditure price index (PCEPI), industrial production (INDPRO), housing supply (MSACSR), the M2 measure of money supply (M2), unemployment rate (UNRATE), and 10-year Government bond yields (IRLTLT01USM156N) from the Federal Reserve Bank of Saint Louis economic database, (FRED). We chose these data series because they are several of the fundamental economic series underlying the macroeconomy, and they span much of the interesting variation.

All of the data are seasonally-adjusted by FRED. We convert to approximate percent changes by log-differencing all of the data except for the consumption measure, which is already measured in percent changes, the unemloyment rate and the long-term interest rate. We then demean the data and rescale them so they have standard deviations equal to 1. This is useful because it puts all of the data on the same scale.

The data covers the January 1963 to December 2018. The time dimension is 671, and the cross-sectional dimension is 7. Figure 4 shows the standardized monthly macroeconomic data used in this subsection. The gray bars are the NBER recessions.

We use the same prior for both datasets and for the simulation, as in Table 1 to make our results more easily interpretable. The prior we use for the component coefficients has a Kronecker structure, and so we specify prior beliefs over the relationship between regressands and regressors separately. In particular, the parameters are a priori independent across different regressands.

The prior we use for the component parameters and base Dirichlet measure is rather flat. We are not imposing a great deal of a priori structure. In addition, the theory tells us it will not matter asymptotically.

Figure 4: Monthly Macroeconomic Series

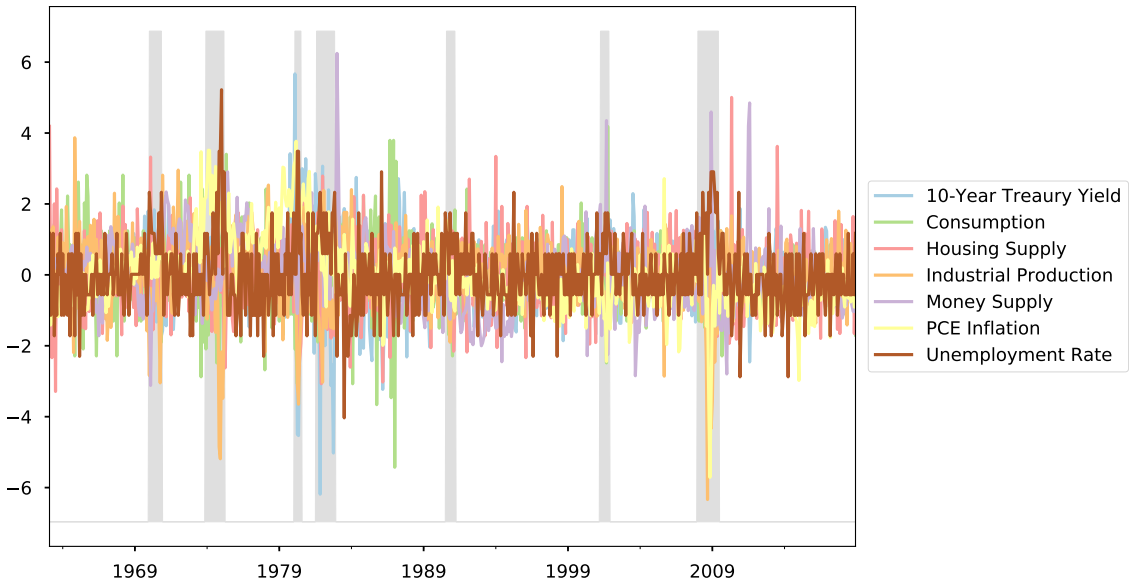


Table 1: Prior

Degrees of freedom for the hierarchical prior	5
Expected Number of Components	5
Component Coefficients	
Intercept	0
Expected Diagonal Autocorrelation	0.9
Expected Off-Diagonal Autocorrelation	0
Component Covariances	
Mean	$.25^2\mathbb{I}_D$
μ_1	3
μ_2	3

8 Empirical Results

8.1 Monthly Macroeconomic Series

Using the macroeconomic data, we obtain the posterior draws from our sampler which are summarized in Fig. 5. In Fig. 5a, we see that the conditional mean tracks the dynamics of data quite well. We can divide the conditional variance in each period into two components using the law of total volatility:

$$\mathbb{V}\mathrm{ar}\left(x_t \,|\, \mathcal{F}_{t-1}\right) = \mathbb{V}\mathrm{ar}\left(\mathbb{E}\left[x_t \,|\, \delta_t\right] | \, \mathcal{F}_{t-1}\right) + \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left(x_t \,|\, \delta_t\right) | \, \mathcal{F}_{t-1}\right]. \tag{41}$$

Since the model is linear conditional on the cluster identity δ_t, the first term comes from variation in $\beta_k x_t$, while the second arises from variation in the innovations. Figure 5d shows the volatility associated with autoregressive coefficients, whereas Fig. 5b shows the volatility associated with

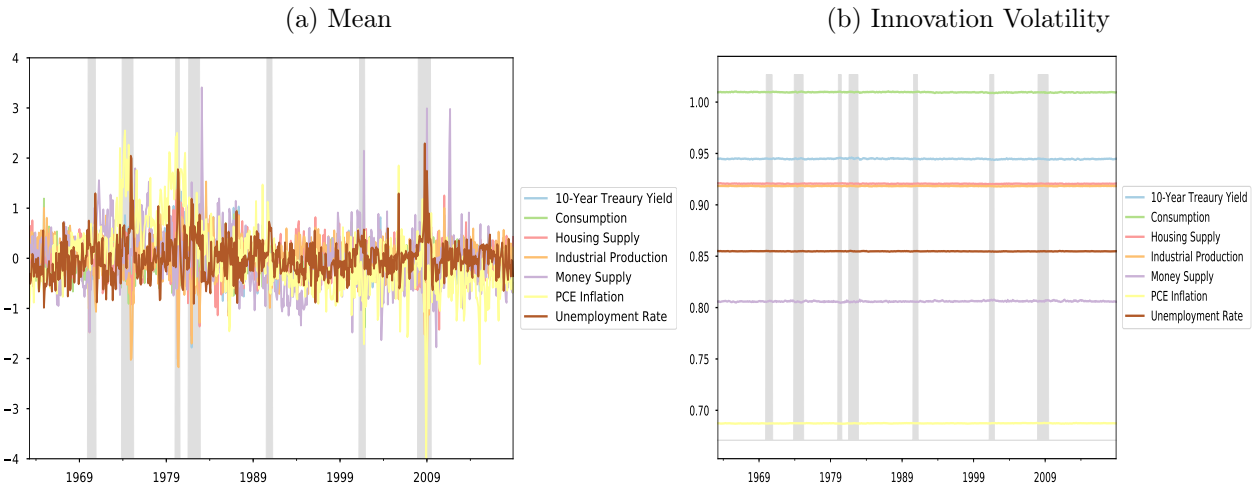
innovations. The total volatility, which we graph for consumption in Fig. 8a, is the sum of the two. Interestingly, most of the variation arises from the variation in the conditional means, not variation in the conditional variances.

Comparing these two volatilities, we observe bigger changes in dynamics for the coefficient volatility. This implies that the stochastic volatility in macroeconomic data studied in papers such as Fernández-Villaverde and Rubio-Ramírez (2010) and Fernández-Villaverde, Guerrón-Quintana, Kuester, and Rubio-Ramírez (2015) can be more parsimoniously modeled using variation in the conditional mean than by using stochastic volatility.

Figure 5c shows the mixture probability of the first cluster in each period. From the empirical results, we see that 5 clusters become active in our sample but the mixture probability of the first cluster is very high. Hence, our model is very parsimonious, which we did not impose. We can also see that the mixture probabilities fluctuates at the monthly frequency.

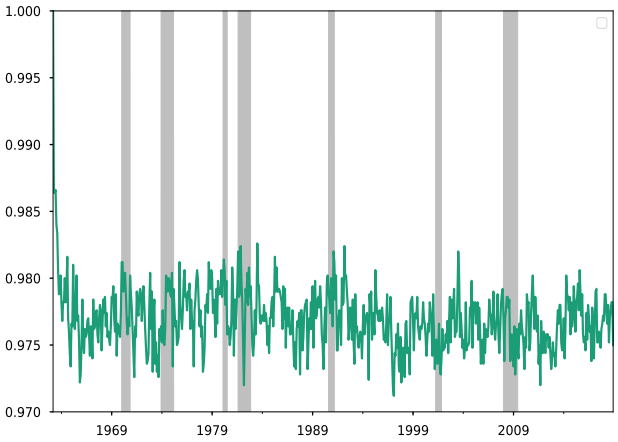
Since our estimator uses a mixture representation, we can link it to regime-switching models by viewing the mixtures as regimes. Our model provides more flexibility than the standard regime-switching models, because it endogenously determines the number of regimes and multiple regimes with different probabilities can be active in each period. We are not interested in identifying the underlying regimes, (which is impossible) but only in approximating the true density. Unlike many regime-switching models, we do not have a "recession" regime and "normal-times" regimes.

Figure 5: Empirical Results with Monthly Macroeconomic Series



To show that our algorithm works reasonably well in practice, we display the conditional density forecast for consumption in Fig. 6. Predictive Densities and PIT's for the other series are provided in Appendix D. The qualitative performance of the estimator and its its relationship to the VAR estimator holds for all of the series considered. If the model works perfectly, the probability integral transform (PIT) should be independent and distributed $U[0,1]$. As we can see, it is roughly independent and distributed approximately uniformly. This implies it is picking up the underlying data's time-varying volatility and skewness. It is worth noting that since the main objective of this paper is density estimation, not forecasting, we report in-sample fits.

(c) Mixture Probability of the First Cluster



(d) Coefficient Volatility

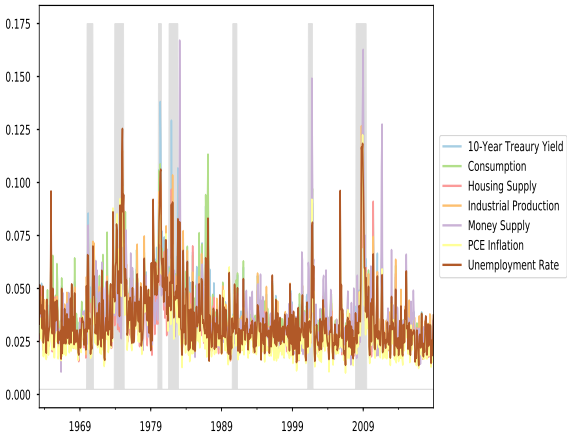
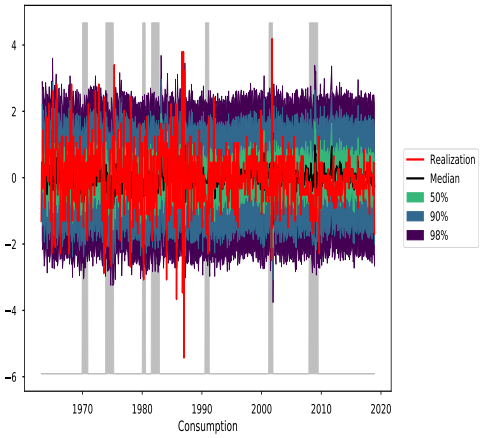
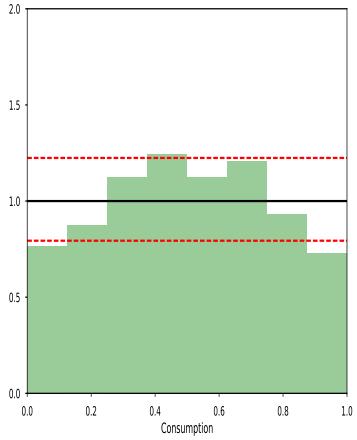


Figure 6: 1-Period Ahead Conditional Forecasts: Consumption Expenditure

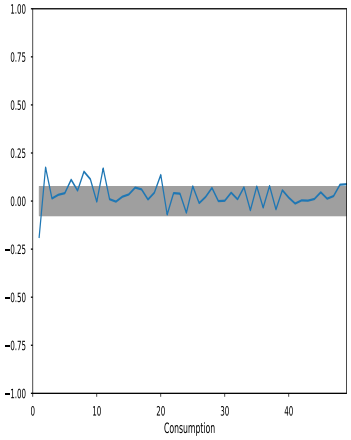
(a) Posterior Density



(b) PIT Histogram



(c) PIT ACF



The dynamics of the data in Fig. 6a are not obviously non-Gaussian or non-linear. Are we effectively just estimating a VAR? The answer is no. First of all, we can compare the 1-period ahead conditional forecasts from our model and those from Bayesian VAR(1). We can clearly see in Fig. 7b that the PIT for the VAR is far from being uniformly distributed. For example, the underlying true DGP is skewed, but the VAR is not. In addition, by examining the conditional variance, (Fig. 8a), we see that it spikes a great deal in recessions when we compute the rolling averages over 1 year. In other words, we can see stochastic volatility in consumption data that varies with the business cycle which would not be captured by VAR(1). Interestingly, although skewness (Fig. 8b) and kurtosis (Fig. 8c) exhibit significant time-variation, this time variation does not obviously co-move with the business cycle. For instance, skewness of consumption fluctuated more in magnitude in the 70s and 80s than it did in the period after 2000.

This time-variation in volatility but not in higher-moments is interesting on a number of dimensions. For example, similar to Schorfheide, Song, and Yaron (2018), we find stochastic volatility for consumption growth at business cycle frequencies using purely macroeconomic data. Conversely,

Figure 7: 1-Period Ahead Conditional Forecasts: Consumption Expenditure (VAR(1))

(a) Posterior Density (b) PIT Histogram (c) PIT ACF

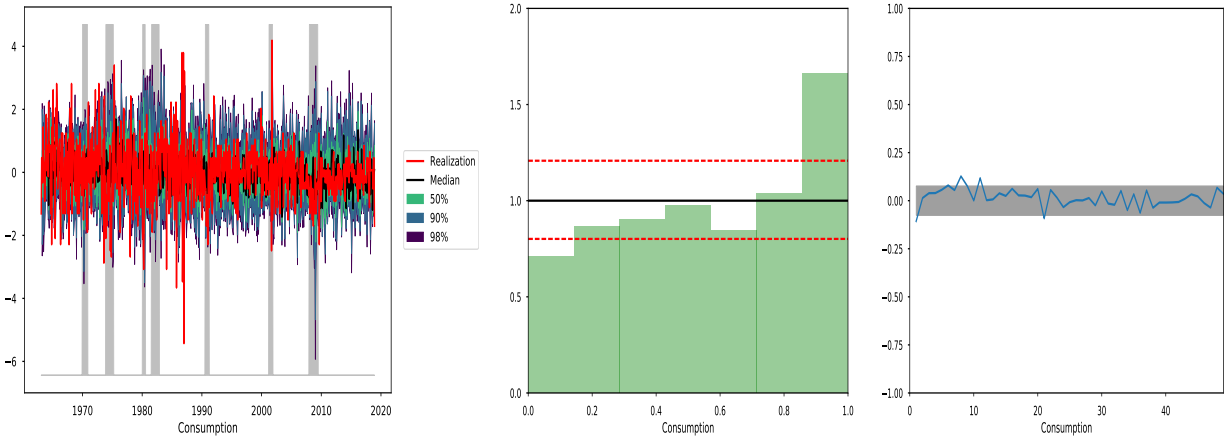
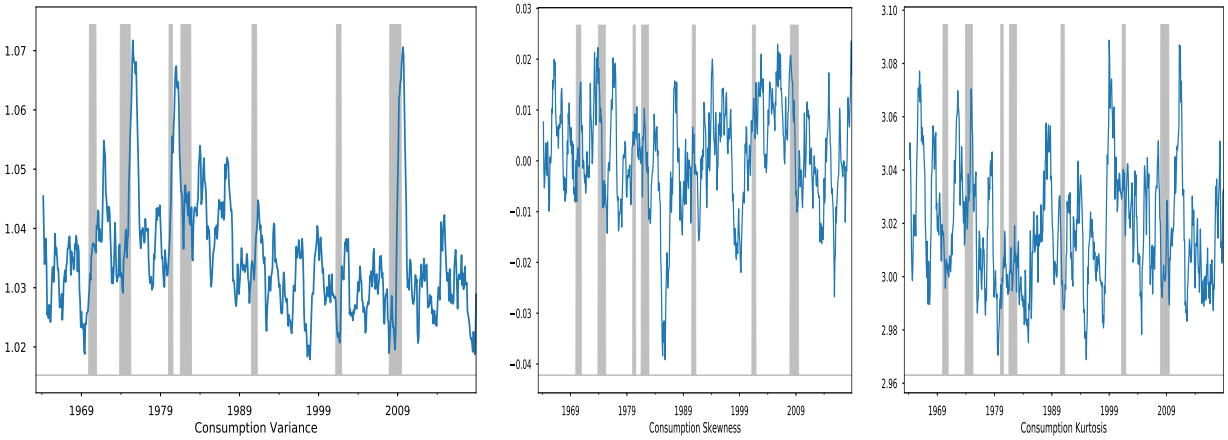


Figure 8: Consumption Variability

(a) Variance (b) Skewness (c) Kurtosis



disaster models such as Barro and Jin (2011) and Tsai and Wachter (2016) predict that kurtosis should either always be high (not approximately 3) or increase substantially during disasters.

9 Conclusion

In this paper, we show how to practically estimate marginal and transition densities of multivariate processes. This is a classic question in econometrics because most economic datasets are multivariate and parametric approximations often perform poorly. Furthermore, even outside of economics, other data-based disciplines face the same issues. We develop a Dirichlet Gaussian mixture model to estimate a wide variety of processes quite rapidly. Our method scales to a more series than the literature has thus far been able to handle and performs reasonably well in practice.

We provide new theory that shows, under some general assumptions, the posterior distribution of our estimators converges more rapidly than the previous literature has been able to achieve.

In particular, we exploit the tail behavior of probability distributions in high dimensions to show that our estimator for the marginal densities converges at a $\sqrt{\log(T)/T}$ rate and our estimator for the transition densities converge at a $\log(T)/\sqrt{T}$ rate with high probability. They are noteworthy because they are the parametric rate up to a logarithmic term. These rates are remarkable because they do not depend on the number of series except through the constant term.

We show that this estimation strategy performs well in simulations and when applied to various macroeconomic and financial data. In the empirical applications, we show that macroeconomic and financial data's dynamics are often far from Gaussian and the dynamic structure moves across the business cycle. We further find that our proposed representation requires more than one mixture component, but only a few, to handle the data's dynamics well.

References

Barro, R. J., and T. Jin (2011): "On the Size Distribution of Macroeconomic Disasters," *Econometrica*, 79(5), 1567–1589.

Birgé, L. (2013): "Robust Tests for Model Selection," in *From Probability to Statistics and Back: High-Dimensional Models and Processes — A Festscrift in Honor of Jon A. Wellner*, ed. by M. Banerjee, F. Bunea, J. Huang, M. Koltchinksii, and M. H. Maathius, vol. 9 of *IMS Collections*, pp. 47–68. Institute of Mathematical Statistics.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

de la Peña, V. H. (1999): "A General Class of Exponential Inequalities for Martingales and Ratios," *The Annals of Probability*, 27(1), 537–564.

Fernández-Villaverde, J., P. Guerrón-Quintana, K. Kuester, and J. Rubio-Ramírez (2015): "Fiscal Volatility Shocks and Economic Activity," *American Economic Review*, 105(11), 3352–84.

Fernández-Villaverde, J., and J. Rubio-Ramírez (2010): "Macroeconomics and Volatility: Data, Models, and Estimation," .

Geweke, J., and M. Keane (2007): "Smoothly Mixing Regressions," *Journal of Econometrics*, 138(1), 252–290, 50th Anniversary Econometric Institute.

Ghosal, S., J. K. Ghosh, and A. W. van der Vaart (2000): "Convergence Rates of Posterior Distributions," *The Annals of Statistics*, 28(2), 500–531.

Ghosal, S., and A. W. van der Vaart (2007): "Convergence Rates of Posterior Distributions for Non-i.i.d. Observations," *The Annals of Statitics*, 35, 192–223.

——— (2017): *Fundamentals of Nonparametric Bayesian Inference*, vol. 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Griffin, J. E., and S. G. Walker (2011): "Posterior Simulation of Normalized Random Measure Mixtures," *Journal of Computational and Graphical Statistics*, 20(1), 241–259.

Huang, A., and M. P. Wand (2013): "Simple Marginally Noninformative Prior Distributions for Covariance Matrices," *Bayesian Analysis*, 8(2), 439–452.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," chap. 74, pp. 5369–5468.

Ishwaran, H., and L. F. James (2001): "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96(453), 161–173.

Johnson, W. B., and J. Lindenstrauss (1984): "Extensions of Lipschitz Maps into a Hilbert Space," *Contemporary Mathematics*, 26, 189–206.

Klartag, B., and S. Mendelson (2005): "Empirical Processes and Random Projections," *Journal of Functional Analysis*, 225(1), 229–245.

Koop, G., D. Korobilis, and D. Pettenuzzo (2019): "Bayesian Compressed Vector Autoregressions," *Journal of Econometrics*, 210(1), 135–154, Annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective".

Lin, D., E. Grimson, and J. Fisher (????): "Construction of Dependent Dirichlet Processes Based on Poisson Processes," pp. 1396–1404.

Nguyen, X. (2016): "Borrowing Strengh in Hierarchical Bayes: Posterior Concentration of the Dirichlet Base Measure," *Bernoulli*, 22(3), 1535–1571.

Norets, A. (2010): "Approximation of Conditional Densities by Smooth Mixtures of Regressions," *The Annals of Statistics*, 38(3), 1733–1766.

Papaspiliopoulos, O., and G. O. Roberts (2008): "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models," *Biometrika*, 95(1), 169–186.

Pati, D., D. B. Dunson, and S. T. Tokdar (2013): "Posterior Consistency in Conditional Distribution Estimation," *Journal of Multivariate Analysis*, 116, 456–472.

Schorfheide, F., D. Song, and A. Yaron (2018): "Identifying Long-Run Risks: A Bayesian Mixed-Frequency Approach," *Econometrica*, 86(2), 617–654.

Sethuraman, J. (1994): "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4(2), 639–650.

Shen, X., and L. Wasserman (2001): "Rates of Convergence of Posterior Distributions," *The Annals of Statistics*, 29(3), 687–714.

Stone, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8(6), 1348–1360.

Talagrand, M. (1996): "Majorizing Measures: The Generic Chaining," *The Annals of Probability*, 24(3), 1049–1103.

——— (2014): *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, vol. 60. Springer Science & Business Media.

Tsai, J., and J. A. Wachter (2016): "Rare Booms and Disasters in a Multisector Endowment Economy," *The Review of Financial Studies*, 29(5), 1113–1169.

van der Vaart, A. W., and H. J. van Zanten (2008): "Rates of Contraction of Posterior Distributions based on Gaussian Process Priors," *The Annals of Statistics*, 36(3), 1435–1463.

Walker, S. G. (2007): "Sampling the Dirichlet Mixture Model with Slices," *Communications in Statistics – Simulation and Computation*, 36(1), 45–54.

Wong, W. H., and X. Shen (1995): "Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLEs," *The Annals of Statistics*, 23(2), 339–362.

Yang, Y., and A. Barron (1999): "Information-Theoretic Determination of Minimax Rates of Convergence," *The Annals of Statistics*, 27(5), 1564–1599.

Appendix A Measure Concentration

A.1 Generic Chaining

We start with recalling a few definitions and fixing some notation. Recall the definition of a γ-functional, where the infimum is taken with respect to all subsets $\mathcal{X}_s \subset \mathcal{X} \subset \mathbb{R}^{T\times D}$ such that the cardinality $|\mathcal{X}_s| \leq 2^{2^s}$ and $|\mathcal{X}_0| = 1$, and d is a metric: $\gamma_\alpha(\mathcal{X}, d) = \inf \sup_{x\in\mathcal{X}} \sum_{s=0}^{\infty} 2^{s/\alpha} d(s, \mathcal{X}_s)$. This $\gamma_2(\mathcal{X}, d)$ functional is useful because it controls the expected size of a Gaussian process by the majorizing measures theorem, (Talagrand (1996)).

Recall the definition of the Orlicz norm of order n: $\psi_n \coloneqq \inf_{C>0} \mathbb{E}\left[\exp\left(\frac{|X|^n}{C^n} - 1\right) \leq 1\right]$. This is useful because a standard argument shows if X has a bounded ψ_n norm then the tail of X decays faster than $2\exp\left(-\frac{x^n}{\|x\|_{\psi_n}^n}\right)$. Hence, if x has a finite ψ_2-norm, it is subgaussian.

A.2 Definition and Properties of the Θ_T-operator

Lemma 3. *Let K be the number of columns of Θ_T as defined in Definition 3. Then its probability density function has the following form, where $\mu \coloneqq \Pr(b = 1)$.*

$$\Pr(K \leq \widetilde{K}) = \left(1 - (1-\mu)^{\widetilde{K}}\right)^T \tag{42}$$

Proof. Let θ_t denote a row of Θ_T. Then

$$\Pr(K \leq \widetilde{K}) = \Pr(1 \in \theta_t \text{ for all } t) = \Pr(1 \in \theta_t)^T = (1 - \Pr(1 \notin \theta_t))^T = \left(1 - (1-\mu)^{\widetilde{K}}\right)^T. \tag{43}$$

□

Lemma 4. *There exists a constant $\gamma \in (0, 1)$ and constants c_1, c_2, such that with probability at least γ, the following holds.*

$$c_1 \log(T) \leq K \leq c_2 \log(T) \tag{44}$$

Proof. Let $B \coloneqq \exp(\widetilde{K})$ We set the cumulative distribution function equal to $1-\gamma$, i.e. the survival function equal to γ:

$$(1-\gamma) = (1-(1-\mu)^{\widetilde{K}})^T \implies \log(1-\gamma)/T = \log(1-(1-\mu)^{\tilde{K}}). \tag{45}$$

Note, for positive a and b, $a^{\log(b)} = b^{\log(a)}$, which can be verified by taking logs of both sides.

$$\log(1-\gamma)/T = \log\left(1-\left(\frac{1}{1-\mu}\right)^{-\log B}\right) = \log\left(1-\left(\frac{1}{B}\right)^{-\log(1-\mu)}\right) = \log\left(1-B^{\log(1-\mu)}\right). \tag{46}$$

Taking the Taylor series approximation of the logarithm function around 1 gives

$$-\log(1-\gamma)/T \approx B^{\log(1-\mu)} \implies T \propto B^{-\log(1-\mu)} \implies B \propto T^{-1/\log(1-\mu)}. \tag{47}$$

This implies

$$K \propto -\frac{1}{\log(1-\mu)}\log(T) \propto \log(T). \tag{48}$$

We can bound this in the opposite direction by replacing $1-\gamma$ with γ since $\gamma \in (0,1)$.

□

A.3 Relationship between the Orlicz and L_2 norms.

We use the following lemma in our proof of Theorem 1. We need it to bound the tail deviations using a bound on the 2nd moment deviations.

Lemma 5. *Let Θ_T be a matrix constructed as in Definition 3. Let $\{x_t\}_{t=1}^T$ be a sequence of known random vectors of length D. Then we have the following.*

1. *The squared L_2-norm of x is equivalent to $\mathbb{E}\left[\langle \Theta_k, x\rangle^2\right]$.*
2. *The squared L_2-norm of x, $\|x\|_{L_2}^2$ dominates the 2nd-order Orlicz norm.*

Proof.

Part 8.1. First, we start by showing Item 1. Let Θ_k denote a column of the matrix. The root of the proof follows from realizing that Θ_T is a generalized selection matrix, and covariances are dominated by variances:

$$\mathbb{E}_\Theta\left[X'\Theta_k\Theta_k'X\right] = \mathbb{E}_\Theta\left[\sum_{t=1}^T x_t\theta_{t,k}\theta_{t,k}x_t'\right] = \mathbb{E}_{\Theta_k}\left[\sum_{t=1}^T|\theta_{t,k}|x_tx_t'\right] = \frac{1}{K}\sum_{t=1}^T x_tx_t', \tag{49}$$

where the last line follows by the independence of the rows of Θ_k.

Consider $\mathbb{E}_\Theta\left[X'\Theta\Theta'X\right]$. Since the columns of Θ_T are a martingale difference sequence, variances of sums are sums of variances:

$$\mathbb{E}_\Theta\left[X'\Theta\Theta'X\right] = \sum_{k=1}^{K}\mathbb{E}_{\Theta_k}\left[X'\Theta_k\Theta_k'X\right] = \sum_{t=1}^{T}x_t x_t'. \tag{50}$$

Part 8.2. Now that we have shown Item 1, we must show that L_2-norm dominates the ψ_2-norm. This is useful because it implies that if we can control the variance of the distribution, we automatically control the tails as well:

$$\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{|\langle\Theta_k,x\rangle|^2}{C^2}\right)\right]-1\leq 1\right\} \tag{51}$$

$$=\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{\sum_{t=1}^{T}|\theta_{t,k}|x_t'x_t+2\sum_{t,\tau\neq t}\theta_{t,k}\theta_{\tau,k}x_t'x_\tau}{C^2}\right)\right]\leq 2\right\}. \tag{52}$$

Since the cross-terms are proportional to squares, and the Θ_k are generalized selection vectors this bounded by

$$\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{2\sum_{t=1}^{T}|\theta_{t,k}|x_t'x_t}{C^2}\right)\right]\leq 2\right\}. \tag{53}$$

By the definition of the exponential function, $|\theta_{t,k}|\in\{0,1\}$, and the multinomial theorem, this equals

$$\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\sum_{h=0}^{\infty}\frac{2^h\left(\sum_{t=1}^{T}|\theta_{t,k}|x_t'x_t\right)^h}{C^{2h}h!}\right]\leq 2\right\} \tag{54}$$

$$=\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\sum_{h=0}^{\infty}\frac{2^h\sum_{\sum k_t=h}\binom{h}{k_1,k_2,\ldots k_T}\prod_{t=1}^{T}|\theta_{t,k}|(x_t'x_t)^{k_t}}{C^{2h}h!}\right]\leq 2\right\}. \tag{55}$$

Since everything is absolutely convergent, we can interchange expectations and infinite sums, and so this equals

$$\inf\left\{C>0\,\middle|\,\sum_{h=0}^{\infty}\frac{2^h\sum_{\sum k_t=h}\binom{h}{k_1,k_2,\ldots,k_T}\prod_{t=1}^{T}\frac{1}{K}(x_t'x_t)^{k_t}}{C^{2h}h!}\leq 2\right\}. \tag{56}$$

Then we can use the multinomial theorem and the formula for the exponential function in the reverse direction, implying this equals

$$\inf\left\{C>0\,\middle|\,\frac{1}{K}\exp\left(\frac{2\|x\|_{L_2}^2}{C^2}\right)\leq 2\right\}=\inf\left\{C>0\,\middle|\,\frac{2\|x\|_{L_2}^2}{C^2}=\log(2K)\right\}\leq\frac{\sqrt{2}\|x\|_{L_2}}{\sqrt{\log(2)}}, \tag{57}$$

where the last inequality follows because $K\geq 1$. Hence, we have that the L_2-norm dominates the ψ_2-norm.

□

A.4 Norm Equivalence

In the section below we reproduce (Klartag and Mendelson, 2005, Prososition 2.2). The one change that we make is that we spell out one of the constants as a function of its arguments. We do this because we will need to take limits with respect to δ on what follows.

Proposition 9 (Klartag and Mendelson (2005) Proposition 2.2)**.** *Let $(\mathcal{X}, d)$ be a metric space and let $\{Z_x\}_{x\in\mathcal{X}}$ be a stochastic process. Let $K > 0, \Upsilon : [0,\infty) \to \mathbb{R}$ and set $W_x := \Upsilon(|Z_x|)$ and $\epsilon := \frac{\gamma_2(\mathcal{X},d)}{\sqrt{K}}$. Assume that for some $\eta > 0$ and $\exp(-c_1(\eta)K) < \delta < \frac{1}{4}$, the following hold.*

1. *For any $x, y \in \mathcal{X}$ and $u < \delta_0 := \frac{4}{\eta}\log\frac{1}{\delta}$,*

$$\Pr\left(|Z_x - Z_y| > ud(x,y)\right) < \exp\left(-\frac{\eta}{\delta_0}Ku^2\right)$$

2. *For any $x, y \in \mathcal{X}$ and $u > 1$*

$$\Pr\left(|W_x - W_y| > ud(x,y)\right) < \exp\left(-\eta K u^2\right)$$

3. *For any $x \in \mathcal{X}$, with probability larger than $1-\delta$, $|Z_x| < \epsilon$.*

4. *Υ is increasing, differentiable at zero and $\Upsilon'(0) > 0$.*

Then, with probability larger than $1-2\delta$, with $C(\Upsilon,\delta,\eta) := \left(c(\Upsilon)c(\eta)(\frac{2}{\eta}(\log\frac{1}{\delta}+1))\right) > 0$, and both $c(\Upsilon)$ and $c(\eta)$ depend solely on their arguments.

$$\sup_{x\in\mathcal{X}}|Z_x| < C(\Upsilon,\delta,\eta)\epsilon.$$

Here we quote a version of Bernstein's inequality for martingales due to ((de la Peña, 1999, Theorem 1.2A)), which we use later.

Theorem 10 (Bernstein's Inequality for Martingales)**.** *Let $\{x_i, \mathcal{F}_i\}$ be a martingale difference sequence with $\mathbb{E}[x_i \,|\, \mathcal{F}_{i-1}] = 0, \mathbb{E}\left[x_i^2 \,\middle|\, \mathcal{F}_{i-1}\right] = \sigma_i^2, v_k = \sum_{i=1}^k \sigma_i^2$. Furthermore, assume that $\mathbb{E}[|x_i|^n \,|\, \mathcal{F}_{i-1}] \le \frac{n!}{2}\sigma_i^2 M^{n-2}$ almost everywhere. Then, for all $x, y > 0$,*

$$\Pr\left(\left\{\left|\sum_{i=1}^k x_i\right| \ge u, v_k \le y \text{ for some } k\right\}\right) \ge 2\exp\left(-\frac{u^2}{2(y+uM)}\right). \tag{58}$$

If we choose c small enough, this implies

$$\Pr\left(\left\{\left|\frac{1}{k}\sum_{i=1}^k x_i\right| \ge u, v_k \le y \text{ for some } k\right\}\right) \ge 2\exp\left(-c\min\left\{\frac{u^2k^2}{v}, \frac{uk}{M}\right\}\right). \tag{59}$$

Theorem 1 (Bounding the Norm Perturbation)**.** *Let Θ_T be constructed as in Definition 3 with the number of columns denoted by K_T. Let $\epsilon > 0$ be given. Let $0 < \delta < 1$ be given such that*

$0 < \log(\frac{1}{\delta}) < c_1\epsilon^2 K_T$ *for some constant* c_1. *Let* $\widetilde{X}_T$ *be in the unit hypersphere in* $\mathbb{R}^{TD-1}$. *Then with probability greater than* $1-2\delta$ *with respect to* Θ_T, *there exists a constant* c_2 *such that for any* $\epsilon > \sqrt{\frac{\log T}{K_T}}$,

$$\sup_t \left| \|\theta_t \widetilde{x}_t\|_{L_2} - \|\widetilde{x}_t\|_{L_2} \right| < c_2 \left(1 + \log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

Proof. We mimic the proof of (Klartag and Mendelson, 2005, Theorem 3.1), verifying the conditions of Proposition 9. Similar to them we use $\Upsilon(t) \coloneqq \sqrt{1-t}$. Our conclusion is stated in terms of the logarithm of the sample size — T. This conclusion is weaker than theirs as $\gamma_2\left(\widetilde{\mathcal{X}}, \|\cdot\|_{L_2}\right) < C\sqrt{\log(T)}$. We can see this by combining the majorizing measure theorem, ((Talagrand, 2014, Theorem 2.4.1)), and the minoration theorem, ((Talagrand, 2014, Lemma 2.4.2)). Effectively, we have an upper bound for the supremum of a Gaussian process and tighter upper bound for the same process.

We start by fixing some notation. Let $x, y \in \mathcal{X}$. We use the functional notation $x(\theta_k)$ to refer $\sum_{d=1}^{D} \theta'_k x_d$.

$$Z_x^K \coloneqq \frac{1}{K}\sum_{k=1}^{K} x^2(\theta_k) - \|x\|_{L_2}^2 \tag{60}$$

Consider $Z_x^K - Z_y^K$.

$$Z_x^K - Z_y^K = \frac{1}{K}\sum_{k=1}^{K} x^2(\theta_k) - y^2(\theta_k) = \frac{1}{K}\sum_{k=1}^{K}(x-y)(\theta_k)(x+y)(\theta_k) \tag{61}$$

Part 10.1. Let $Y_k \coloneqq x^2(\theta_k) - y^2(\theta_k)$, then

$$\begin{aligned} &\Pr(|Y_k| > 4u\|x-y\|_{\psi_2}\|x+y\|_{\psi_2}) \\ &\quad \leq \Pr(|x(\theta_k) - y(\theta_k)| > 2\sqrt{u}\|x-y\|_{\psi_2}) + \Pr(|x(\theta_k) + y(\theta_k)| > 2\sqrt{u}\|x+y\|_{\psi_2}) \\ &\quad \leq 2\exp(-u), \end{aligned} \tag{62}$$

where the last inequality comes from the sub-exponential tails of $\theta_{t,k}$ and the first by the union bound. This implies that $\|Y_k\|_{\psi_1} \leq c_1\|x-y\|_{\psi_2}\|x+y\|_{\psi_2} \leq c_2\|x-y\|_{\psi_2}$. We do not need the β used by Klartag and Mendelson because the entries in our θ operator are uniformly bounded by 1 in absolute value.

The Y_k are a martingale difference sequence, and so we can apply Theorem 10. They are a martingale difference sequences because the expectation in the next period is either the current value because the increments are mean zero if the sum does not stop or identically zero if they do. If we set $v = 4K\|Y_k\|_{\psi_1}^2$ we can use Bernstein's inequality for martingales mentioned above. $\sum_{k=1}^{K}\sigma_k^2 \leq v$ with probability 1 because this variance is either the same as it is in the independent case or zero. Consequently, by Theorem 10, we have the following if set $v \coloneqq 4K\|\theta\|_{\psi_1}^2$ and $M = \|\theta\|_{\psi_1}$:

$$\Pr\left(\left\{\left|\frac{1}{K}\sum_{k=1}^{K}\theta_k\right| > u\right\}\right) \leq 2\exp\left(-cK\min\left\{\frac{u^2}{\|\theta\|_{\psi_1}^2}, \frac{u}{\|\theta\|_{\psi_1}}\right\}\right) \tag{63}$$

Then by applying (63) to $\Pr(\left|z_x^k - z_y^k\right| > u)$, we have the following.

$$\Pr\left(\left|Z_x^k - Z_y^k\right| > u\right) \leq 2\exp\left(-c\min\left\{\frac{u^2}{\|x-y\|_{L_2}^2}, \frac{u}{\|x-y\|_{L_2}}\right\}\right) \tag{64}$$

The estimate for $\Pr\left(\left|Z_x^k\right| > u\right)$ follows from the same method, but we define $Y_k \coloneqq x^2(\theta_k) - 1$, and use the fact that $\|x(\theta)\|_{\psi_2} \leq 1$, which we verified in Lemma 5. The L_2-norm is bounded above by 1 because we are using rescaled data.

We fix $\eta \leq c$. Assume that $u < \delta_0 = 4\frac{1}{\eta}\log\frac{1}{\delta}$. Then we have

$$\Pr\left(\left|Z_x^k - Z_y^k\right| > 2\|x-y\|_{L_2}\right) \leq 2\exp\left(\eta K \min\left\{u, u^2\right\}\right) < \exp\left(-\eta K\frac{u^2}{\delta_0}\right). \tag{65}$$

Part 10.2. By the triangle inequality,

$$|W_x - W_y| = \left|\left(\frac{1}{K}\sum_{k=1}^K x^2(\theta_i)\right)^{1/2} - \left(\frac{1}{K}\sum_{k=1}^K y^2(\theta_i)\right)^{1/2}\right| \leq \left(\frac{1}{K}\sum_{k=1}^K (x-y)^2(\theta_i)\right)^{1/2}. \tag{66}$$

Applying (63) for $u > 1$:

$$\begin{aligned}\Pr\left(|W_x - W_y| > u\|x-y\|_{\psi_2}\right) &\leq \Pr\left(\frac{1}{K}\sum_{k=1}^K (x-y)^2(\theta_k) > u^2\|x-y\|_{\psi_2}^2\right) \\ &\leq \Pr\left(\frac{1}{K}\sum_{k=1}^K (x-y)^2(\theta_k) > u^2\left\|(x-y)^2\right\|_{\psi_1}\right) \\ &< \exp\left(-cku^2\right).\end{aligned} \tag{67}$$

Since $\eta < c$, this is bounded by $\exp(-\eta K u^2)$.

Part 10.3. For any $x \in \mathcal{X}$ by (63),

$$\Pr(|Z_x| > \epsilon) < \exp(-\eta K\epsilon^2) < \delta. \tag{68}$$

Part 10.4. We can bound the derivative of Υ:

$$\Upsilon'(0) = 1/2 > 0. \tag{69}$$

□

Appendix B Representation Theory

B.1 The Joint Density

Lemma 6 (Bouding Ratio of Sums by Max Ratio)**.** *Let x_t, y_t be a sequence of numbers whose sum is absolutely convergent. Then the ratio of the sums is bounded by the supremum of the ratios, i.e.,*

$$\frac{\sum x_t}{\sum y_t} \leq \sup_t \frac{x_t}{y_t}.$$

Proof. Clearly, if $\#t = 1$, the result holds. Assume $\#t = 2$. Assume the claim is false. Then

$$\begin{aligned} &\frac{x_1 + x_2}{y_1 + y_2} > \max\left\{\frac{x_1}{y_1}, \frac{x_2}{y_2}\right\} \implies x_1 + x_2 > \max\left\{x_1 + \frac{x_1 y_2}{y_1}, x_2 + \frac{x_2 y_1}{y_2}\right\} \\ &\implies x_1 > \frac{x_2 y_1}{y_2} \text{ and } x_2 > \frac{x_1 y_2}{y_1} \implies x_1 > \frac{y_1}{y_2}\frac{x_1 y_2}{y_1} \implies x_1 > x_1. \end{aligned} \tag{70}$$

This is a contradiction. To see the general case we proceed by induction,

$$\frac{\sum_t x_t}{\sum_t y_t} \leq \max\left\{\frac{\sum_{t \neq T} x_t}{\sum_{t \neq T}}, \frac{x_T}{y_T}\right\} \leq \cdots \leq \max\left\{\frac{x_t}{y_t}\right\}, \tag{71}$$

where the first inequality holds by the first step. Clearly, as long as everything convergent, this still holds if we take limits. □

Lemma 7. *Consider the ratio of the densities between p_T and q_T. Let δ_k^q be a clustering of x_t with respect to q_T. Let these clusters δ_k^q satisfy the following, where $\mu_k^q = \mathbb{E}_{P_T}\left[x_t \,\middle|\, t \in \delta_k^q\right]$ and $\Sigma_k^q = \mathbb{C}\text{ov}_{P_T}\left[x_t \,|\, x_t \in \delta_t^q\right]$:*

$$\sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - \left(x_t - \mu_k^q\right)' \left(\Sigma_k^q\right)^{-1} \left(x_t - \mu_k^q\right) \right| < C(\delta)\epsilon. \tag{72}$$

Then the log-divergence satisfies

$$\sup_{x_t, x_t^*} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - (x_{t^*} - \mu_{t^*})' \Sigma_{t^*}^{-1} (x_{t^*} - \mu_{t^*}) \right| < \epsilon \implies \sup_{x_t, x_t^*} \left| \log\left(\frac{p_T(x_t)}{p_T(x_{t^*})}\right) \right| \propto \epsilon. \tag{73}$$

Proof. Consider the log-ratio of Gaussian kernels, by assumption

$$\sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - \left(x_t - \mu_k^q\right)' \left(\Sigma_k^q\right)^{-1} \left(x_t - \mu_k^q\right) \right| \leq \epsilon. \tag{74}$$

Consider the ratio of the proportionality constants χ^p and χ^q associated with the kernels k^p, k^q above:

$$\chi^p = \int_{\mathcal{X}} k^p(x)\, dx, \quad \chi^q = \int_{\mathcal{X}} k^q(x)\, dx. \tag{75}$$

By the definition of proportionality constant, we can write

$$\log\left(\frac{\chi^q}{\chi^p}\right) = \log\left(\frac{\sum k^q(x)\,dx}{\sum k^p(y)\,dy}\right) = \log\left(\frac{\sum k^q(x)/p_T(x)\,dP_T(x)}{\sum k^p(y)/p_T(y)\,dP_T(y)}\right), \tag{76}$$

where we can change measures to P_T. By Lemma 6, this is bounded by the supremum of the ratios, since we are integrating over the same space in both sums:

$$\leq \sup_x \log\left(\frac{k^q(x)/p_T(x)}{k^p(x)/p_T(x)}\right) \leq \sup_x \log\left(\frac{k^q(x)}{k^p(x)}\right), \tag{77}$$

because the Jacobian terms cancel. We can bound the inverse-ratio of the proportionality constants — $\frac{\mu_q}{\mu_p}$ — in the same way. We just interchange the labels on the kernels. Consequently, the proportionality constants satisfy

$$\left|\log\frac{\mu_1}{\mu_2}\right| = \frac{1}{2}c(\delta)\epsilon \tag{78}$$

because the $k^{\cdot}(x)$ are Gaussian kernels, and we bounded the log-ratio in (74). The total deviation is the sum of the deviation in the constants and in the kernels. The result holds by combining (78) and (74).

□

Proposition 11 (Bounding the Supremum of the Rescaled Data)**.** *Let $\widetilde{X} \coloneqq \widetilde{x}_1, \ldots, \widetilde{x}_T$ be a D-dimensional Gaussian process with finite stochastic means μ_t and covariances Σ_t, where Σ_t is positive-definite for all t. Let Θ_T be the generalized selection matrix defined in Definition 3. Let $\widetilde{P}_T$ denote the distribution of $\widetilde{X}$. Then given $\epsilon > 0$ and for some $\delta \in (0,1)$, the approximating distribution $\widetilde{Q}_T$, which is the mixture distribution over $\{\widetilde{\Sigma}_t^{-1/2}\widetilde{x}_t\}_{t=1}^T$ defined by the clustering induced by Θ_T satisfies the following with probability at least $1-2\delta$ with respect to Θ_T.*

$$\sup_t h^2\left(\int_{G_t}\phi\left(\widetilde{x}_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P), \int_{G_t}\phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\right) < c\left(1+\log\frac{1}{\delta}\right)^2\epsilon^2 \tag{79}$$

Proof. In this proof, we drop the tilde's over the x_t because all of the terms have them. Let G^P and G^Q be the associated mixing measures of the covariances. Let $\mathcal{K}$ be a coupling from between the space of G^P and G^Q. Consider the supremum of the squared Hellinger distance — h^2 — between P_T and Q_T:

$$\sup_t h^2\left(\int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P), \int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\right). \tag{80}$$

Combining the integrals with respect to the marginals (G_t^P, G_t^Q) into a integral with respect to the joint, and exploiting the convexity of the supremum and of the squared Hellinger distance gives:

$$\leq \int_{G_t^P \times G_t^Q} \sup_t h^2\left(\phi\left(x_t \,\middle|\, \delta_t^P\right), \phi\left(x_t \,\middle|\, \delta_t^Q\right)\right) d\mathcal{K}(G_t^P, G_t^Q). \tag{81}$$

We expand the definition of h^2 using its formula as an f-divergence:

$$\leq \int_{G_t^P \times G_t^Q} \sup_t \int_{\mathbb{R}^D} \left| \left(\frac{\phi\left(x_t \,\middle|\, \delta_t^P\right)}{\phi\left(x_t \,\middle|\, \delta_t^Q\right)} \right)^{1/2} - 1 \right|^2 d\Phi\left(x_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q). \tag{82}$$

Since we are only considering the density for one period within the integral:

$$= \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} \sup_t \left| \left(\frac{\phi\left(x_t \,\middle|\, \delta_t^P\right)}{\phi\left(x_t \,\middle|\, \delta_t^Q\right)} \right)^{1/2} - 1 \right|^2 d\Phi\left(x_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q). \tag{83}$$

By Lemma 7 and a first-order Taylor series of the exponential function around the logarithm of the original argument, after pulling the square-root inside

$$\leq C_1 \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} \sup_t \left| (x_t - \mu_t^P)' \Sigma_t^P (x_t - \mu_t^P) - (x_t - \mu_t^Q) \Sigma_t^Q (x_t - \mu_t^Q) \right| d\Phi\left(x_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q). \tag{84}$$

Since Q_T was defined through applying Θ_T to $(\Sigma_t^P)^{-1/2}(x_t - \mu_t^P)$, by Theorem 1 this norm perturbation is bounded by ϵ^2; we just have to square the constant:

$$\leq C \left(1 + \log \frac{1}{\delta}\right)^2 \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} |\epsilon|^2 \, d\Phi\left(x_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q) = C \left(1 + \log \frac{1}{\delta}\right)^2 \epsilon^2, \tag{85}$$

where the last equality holds because all of the integrals integrate to 1. □

Theorem 2 (Representing the Joint Density)**.** *Let $\widetilde{X}_T \coloneqq \widetilde{x}_1, \ldots, \widetilde{x}_T$ be a D-dimensional Gaussian process with period-t finite stochastic means, μ_t, and covariances, Σ_t. Let Σ_t be positive-definite for all t. Let Θ_T be the generalized selection matrix constructed in Definition 3. Let $\widetilde{P}_T$ denote the distribution of $\widetilde{X}_T$. Then given $\epsilon > 0$ and $\delta \in (0, 1)$, the approximating distribution, Q_T, which is the mixture distribution over $\widetilde{\mathcal{X}}$ that Θ_T induces, satisfies the following with probability at least $1 - 2\delta$ with respect to Θ_T for some constant C:*

$$h_\infty\left(\widetilde{P}_T(\widetilde{\mathcal{X}}), \widetilde{Q}_T(\widetilde{\mathcal{X}})\right) < C \left(1 + \log\left(\frac{1}{\delta}\right)\right) \epsilon.$$

Proof. Let G^P, G^Q be the associated mixing measures of the associated covariances. Let $\mathcal{K}$ be a coupling from between the space of G^P and G^Q, and the space of such couplings be $\mathcal{T}(G^P, G^Q)$. Consider the squared supremum Hellinger distance — h_∞^2 — between P_T and Q_T. The proof here is based on a combination of proofs of (Nguyen, 2016, Lemma 3.1) and (Nguyen, 2016, Lemma 3.2). Let δ_t be the latent mixture identity that tells you which cluster μ_t, Σ_t is in.

We can represent both densities succinctly as follows. Importantly, we do not require that the

G_t^P are independent:

$$p_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P\left(\delta_t^P\right) dG^P\left(dG_t^P\right). \tag{86}$$

We represent q_T in the same fashion replacing the P's in the expression above with Q's:

$$q_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q\left(\delta_t^Q\right) dG^Q\left(dG_t^Q\right). \tag{87}$$

Then the squared sup-Hellinger distance between the two measures has the following form:

$$\begin{aligned} &h_\infty^2\left(p_T(\widetilde{\mathcal{X}}), q_T(\widetilde{\mathcal{X}})\right) \\ &= h_\infty^2\left(\int\int \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)\, dG^P(dG^P), \int\int \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\, dG^Q(dG_t^Q)\right). \end{aligned} \tag{88}$$

Letting $\mathcal{K}(G^P, G^Q)$ be any coupling between the two densities, we can combine G^P and G^Q into one process. We want to integrate with respect to their joint density:

$$\begin{aligned} = h_\infty^2\Bigg(&\int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)\, d\mathcal{K}(dG_t^P, dG_t^Q), \\ &\int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\, d\mathcal{K}(dG_t^P, dG_t^Q)\Bigg). \end{aligned} \tag{89}$$

Since supremum of squared Hellinger distance is convex as is the supremum, by Jensen's inequality that is bounded

$$\leq \int_{G\times G} \sup_t h^2\left(\int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\right) d\mathcal{K}(dG_t^P, dG_t^Q). \tag{90}$$

If we can bound the supremum of the deviations over the periods, we have bounded the joint. This is true even in the dependent case.

We can place the bound obtained in Proposition 11 inside (90). Since we are integrating $C\epsilon^2$ over a joint density, the density is bounded above by 1, and we are done.

In other words, we have with probability $1 - 2\delta$:

$$h_\infty^2(P_T(\widetilde{\mathcal{X}}), Q_T(\widetilde{\mathcal{X}})) < C\left(\log\frac{1}{\delta}\right)^2 \epsilon^2. \tag{91}$$

□

Lemma 8. *Let f, g be two densities of locally asymptotically normal (LAN) processes with respect to the sample size T.*[16] *Squared Hellinger distance and Kullback-Leibler divergence are equivalent.*

[16] This trivially covers all Gaussian processes with finite-means and variances.

Proof. Consider the following decomposition of the Hellinger distance:

$$\int (\sqrt{f/g} - 1)\, dG = \int \left(\exp\left(\frac{1}{2} (\log f - \log g) \right) - 1 \right) dG. \tag{92}$$

Taking a Taylor expansion of the exponential function:

$$= \int \left(1 + \frac{1}{2} \log\left(\frac{f}{g} \right) + O\left(\log\left(\frac{f}{g} \right)^2 \right) - 1 \right) dG \tag{93}$$

$$= \int \frac{1}{2} \log\left(\frac{f}{g} \right) dG + O\left(\int \log\left(\frac{f}{g} \right)^2 dG \right). \tag{94}$$

Consider one-half the Kullback-Leibler divergence:

$$\frac{1}{2} \int \log\left(\frac{f}{g} \right) \frac{f}{g}\, dG = \frac{1}{2} \int \log\left(\frac{f}{g} \right) \exp\left(\log\left(\frac{f}{g} \right) \right) dG. \tag{95}$$

Taking a 1st-order Taylor expansion of the exponential function:

$$= \frac{1}{2} \int \log\left(\frac{f}{g} \right) \left(1 + \log\left(\frac{f}{g} \right) \right) dG = \frac{1}{2} \int \log\left(\frac{f}{g} \right) dG + O\left(\int \log\left(\frac{f}{g} \right) \log\left(\frac{f}{g} \right) dG \right). \tag{96}$$

The first terms in (93) and (96) are the same. Consequently, is of the same-asymptotic order as $\log(f/g)^2$. By the locally asymptotically normal assumption $\log f(x) \propto (x - \mu_f)' \Sigma_f^{-1} (x - \mu_f) + o(T)$ Choose $\epsilon \propto \frac{1}{T}$. Let z denote the deviation above. By the convexity of the square function and Jensen's inequality, it is sufficient to bound the value inside the integral:

$$\int \log(f/g)^2\, dG \leq \int |z|^2\, dG + O(\epsilon) \leq \int |z|\, dG + O(\epsilon) = \int \log(f/g)\, dG + O(\epsilon), \tag{97}$$

where the first inequality holds by the LAN property, the second inequality holds since $|z| < 1$, and the third-inequality holds by the LAN property. By (93) and (96), the last term in (97) is bounded by both the Hellinger and Kullback-Leibler divergences.

□

B.2 Representing the Marginal Density

Theorem 3 (Representing the Marginal Density)**.** *Let $x_1, \ldots, x_T$ be drawn independently from P_T where each x_t has a infinite-Gaussian mixture representation. Let Θ_T be constructed as in Theorem 2 for each t. Let ϵ be given. Construct Q_T by using the Θ_T operator to group the data and letting the data be Gaussian distributed within each component with component-wise means and covariances given by their conditional expectations. Then, with probability $1 - 2\delta$ with respect to Θ_T, there exists a constant C such that the following holds uniformly over T*

$$h\left(\int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t(\delta_t^Q)\right) < C\left(1 + \log\left(\frac{1}{\delta}\right)\right)\epsilon.$$

Proof. We start by comparing the Hellinger distance between the joint densities, which are both product measures. We want to compare the difference between the marginal densities in terms of the difference between the joint densities. In particular, we show that the difference between the marginal densities is $1/T$ times the difference between the joint densities if the joint densities have a product form. By Theorem 2, we know that is bounded by $T\epsilon^2$, and so we have the desired result. The unusual thing is that we are trying to bound the difference between the joint density and its components in the opposite direction as is usually done. We want to bind the component distance in terms of the joint density distance instead of the other way around.

We can write the squared Hellinger distance between the joint distributions as follows. Let G_m be the marginal distribution over δ_t. Note, the following holds:

$$\prod_{t=1}^{T} \int_{G_t} \phi\left(x_t \,|\, \delta_t\right) dG_t(\delta_t) = \prod_{t=1}^{T} \int_{G_m} \phi\left(x_t \,|\, \delta_t\right) dG_m(\delta_t). \tag{98}$$

All (98) is saying is that the joint T independent draws from the marginal are the same as T independent draws from a sequence $G_1, \ldots G_T$, which is drawn from G. By assumption G has a product form. The Kullback-Leibler divergence between the two joint distributions is

$$\mathrm{D}_{\mathrm{KL}}\left(q_T \,||\, p_T\right) = \int_{\mathbb{R}^{T\times D}} \log\left(\frac{q_T}{p_T}\right) dP_T = \int_{\mathbb{R}^{T\times D}} \log\left(\frac{\prod_{t=1}^{T} \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\prod_{t=1}^{T} \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) dP_T. \tag{99}$$

Ratios of products are products of ratios, and logs of products are sums of logs, and we can substitute in the definition of the marginal

distribution, (98), giving

$$= \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) dP_T = \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_m^Q(\delta_t^Q)}{\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_m^P(\delta_t^P)}\right) dP_T. \tag{100}$$

We can rewrite P_T in terms of its mixture representation:

$$\int_{G_t} \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) \prod_{t=1}^{T} \phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t). \tag{101}$$

The only interactions between the two terms are the x_t:

$$= \sum_{t=1}^{T} \left(\left(\int_{G_t} \int_{\mathbb{R}^D} \log\left(\frac{\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) \phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t)\right) \left(\int_{\mathbb{R}^{(T-1)\times D}} \prod_{\tau\neq t} \phi\left(x_\tau \,|\, \delta_\tau\right) dx\, dG_m^P(\delta_\tau)\right)\right). \tag{102}$$

The second integrals all equal 1, and so their product does as well, giving

$$= \sum_{t=1}^{T} \left(\int_{G_t} \int_{\mathbb{R}^D} \log\left(\frac{\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}{\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}\right) \phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t)\right). \tag{103}$$

The term inside the sum is the Kullback-Leibler divergence between the two marginal distributions, which does not depend upon t:

$$= \sum_{t=1}^{T} \mathrm{D}_{\mathrm{KL}}\left(\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_m^Q(\delta_t^Q) \,\middle\|\, \int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_m^P(\delta_t^P)\right) \tag{104}$$

$$= T\mathrm{D}_{\mathrm{KL}}\left(\int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_m^Q(\delta_t^Q) \,\middle\|\, \int_{G_m} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_m^P(\delta_t^P)\right). \tag{105}$$

In other words, the distance between the joint densities is at least T times the distance between the distance marginal densities. Also, by Lemma 8 this is proportional to squared Hellinger distance. In other words, the difference between the joint densities is at least T times the distance between the distance between the marginal densities. We know by Theorem 2 that this is bounded above by $CT\epsilon^2$. The T

arises because we are no longer using the rescaled data, and $\|X\|^2 \propto T$. This gives

$$h^2\left(\int_{G_m}\phi\left(x_t\,\middle|\,\delta_t^Q\right)dG_m^Q(\delta_t^Q),\int_{G_m}\phi\left(x_t\,|\,\delta_t^P\right)dG_m^P(\delta_t^P)\right)\leq\frac{1}{T}h^2(q_T,p_T)\leq C\frac{T}{T}\epsilon^2=C\epsilon^2. \tag{106}$$

□

Corollary 3.1 (Representing the Marginal Density with Markov Data)**.** *Theorem 3 continues to hold when the x_t form a uniformly ergodic hidden Markov chain instead of being fully independent.*

Proof. Let z_1 be a latent variable such that (x_t, z_t) forms Markov sequence. Consider a reshuffling $(\tilde{x}_1, \tilde{z}_1), \ldots (\tilde{x}_T, \tilde{z}_T)$. Now both of these sequences clearly have the same marginal distribution. (They likely do not have the same joint distribution.) Hence, by Theorem 3 the result follows since the reshuffled data has a product density.

□

B.3 Representing the Transition Density

Theorem 4 (Transition Density Representation)**.** *Let $x_1, \ldots, x_T \in R^{T\times D}$ be a uniformly ergodic Markov Gaussian process with density p_T. Let $\epsilon > 0$ be given. Let $K \geq c\log(T)^2/\epsilon$ for some constant c. Let δ_t be the cluster identity at time t. Then there exists a mixture density q_T with K clusters with the following form:*

$$q_T\left(x_t\,|\,x_{t-1},\delta_{t-1}\right):=\sum_{k=1}^{K}\phi\left(\beta_k x_{t-1},\Sigma_k\right)\Pr\left(\delta_t=k\,|\,\delta_{t-1}\right).$$

Construct $q_T\left(x_t\,\middle|\,\mathcal{F}_{t-1}^Q\right)$ from $q_T\left(x_t\,|\,x_{t-1},\delta_{t-1}\right)$ by integrating out δ_{t-1} using $\Pr(\delta_{t-1}\,|\,X_T)$. Then with probability $1-2\delta$ with respect to the prior

$$h_\infty\left(p_T\left(x_t\,|\,\mathcal{F}_{t-1}^P\right),q_T\left(x_t\,\middle|\,\mathcal{F}_{t-1}^Q\right)\right)<C\sqrt{1+\log\left(\frac{1}{\delta}\right)}\epsilon.$$

Proof. We need the conditional density of $\widetilde{x}_t\,|\,\widetilde{x}_{t-1},\delta_{t-1}$. By Theorem 2, there exists a generalized selection matrix Θ_T satisfying the statement of the theorem. Conditional on Θ_T, the distribution is Gaussian. So consider the following where θ_t is the t^{th} row of Θ_T. (Throughout, we will implicitly prepend a 1 to $\tilde{x}_{t-1}$ to allow for a non-zero mean as is standard in regression notation.)

By the linearity of Gaussian conditioning in $\theta_t\widetilde{x}_t, \theta_{t-1}\widetilde{x}_{t-1}$ space, for some $\beta_{k,k'}$, $\Sigma_{k,k'}$.

$$\theta_t\widetilde{x}_t \,|\, \widetilde{x}_{t-1}, \theta_t, \theta_{t-1} \overset{\mathcal{L}}{=} \theta_t\widetilde{x}_t \,|\, \theta_{t-1}\widetilde{x}_{t-1}, \theta_t, \theta_{t-1} \overset{\mathcal{L}}{=} \phi(\beta_{k,k'}\theta_{t-1}\widetilde{x}_{t-1}, \Sigma_{k,k'}) \overset{\mathcal{L}}{=} \phi(\beta_{k,k'}\widetilde{x}_{t-1}, \Sigma_{k,k'}). \tag{107}$$

The first equality holds because the elements in each cluster have the same Gaussian distribution under Q_T. The last equality holds because the elements of θ_{t-1} are in $\{-1, 0, 1\}$, we can absorb the θ_{t-1} into the $\beta_{k,k'}$ without increasing the number of clusters more than two-fold. This is because the vectors θ_{t-1} that contain at most one non-zero element form a convex hull, and we take the weighted averages over them in (108).

We want the distribution of $\widetilde{x}_t$ given $\theta_{t-1}, \widetilde{x}_{t-1}$. We do not want to condition on θ_t. So we can just integrate over θ_t using its distribution. Its predictive distribution does not depend upon $\widetilde{x}_{t-1}$ because we construct Θ_T independently of $\widetilde{x}$:

$$\widetilde{x}_t \,|\, \theta_{t-1} = k, \widetilde{x}_{t-1} \sim \sum_{k'} \phi(\beta_{k,k'}\widetilde{x}_{t-1}, \Sigma_{k,k'}) \Pr\left(\theta_t = k'\right) \tag{108}$$

The last probability — $\Pr(\theta_t = k')$ — does not have any conditioning information because the rows of the Θ_T process are independent except for the stopping rule, which is not relevant here. Define a set of clusters in $(\widetilde{x}_t, \widetilde{x}_{t-1})$ space by grouping the ones whose associated $\{\beta, \Sigma\}$ are equal. In other words, take the Cartesian product of the clusters used in (108) and denote the cluster identities by δ_t's. Integrating out the cluster identities gives

$$\widetilde{x}_t \,|\, \widetilde{x}_{t-1}, \delta_{t-1} \sim \sum_{j} \phi(\beta_j\widetilde{x}_{t-1}, \Sigma_j) \Pr\left(\delta_t = j \,|\, \delta_{t-1}\right). \tag{109}$$

Clearly, there are $\log(T)^2 = K_T^2$ different clusters.[17]

We now make a similar argument to the one we made in the marginal density case. Again, we must show that the Kullback-Leibler divergence between the joint density is T times an average Kullback-Leibler divergence. The tricky issue is that we no longer have a product distribution. Instead, we must show that appropriately constructed conditional densities satisfy the necessary inequalities. We start by considering the Kullback-Leibler divergence between the joint distributions. We assumed that p_T is a hidden Markov model. That implies there exists a hidden state z_t such that (x_t, z_t) are jointly Markov. We use capital letters to refer to the entire processes, i.e. Δ_T^P is

[17] The number of clusters used here is of the same asymptotic order as in the prior. This bound may no longer be tight.

the vector of cluster identities with respect to P_T. Consider the supremum of the deviations in each period:

$$\sup_t \mathrm{D}_{\mathrm{KL}} \left(\int_{G_t} \phi \left(x_t \,|\, \delta_t^P \right) dG_t^P(\delta_t^P) \,\middle\|\, \int_{G_t} \phi \left(x_t \,\middle|\, \delta_t^Q \right) dG_t^Q(\delta_t^Q) \right). \tag{110}$$

We can rewrite this as follows by the definition of filtration, since we can condition on only past events without loss of generality, where we G_M to refer to a Markov density:

$$= \sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \sup_t \mathrm{D}_{\mathrm{KL}} \left(\int_{G_M} \phi \left(x_t \,|\, \delta_t^P \right) dG_M^P \left(\delta_t^P \,|\, \mathcal{F}_{t-1}^P \right) \,\middle\|\, \int_{G_M} \phi \left(x_t \,\middle|\, \delta_t^Q \right) dG_M^Q \left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q \right) \right). \tag{111}$$

The goal is to show that the integral of (111) with respect to P_T can be rewritten as the sum of the individual conditionals. We start by considering the Kullback-Leibler divergence between the joint distributions. We assumed p_T is a hidden Markov model. This implies there exists a hidden state z_t such that (x_t, z_t) are jointly Markov. We use capital letters to refer to the entire processes, i.e. Δ_T^P is the vector of cluster identities with respect to P_T. The Kullback-Leibler divergence is

$$\mathrm{D}_{\mathrm{KL}} \left(P_T \,||\, Q_T \right) = \mathrm{D}_{\mathrm{KL}} \left(\prod_{t=1}^T p_T \left(x_t \,|\, \mathcal{F}_{t-1}^P \right) \,\middle\|\, \prod_{t=1}^T q_T \left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q \right) \right). \tag{112}$$

Consider the Kullback-Leibler divergence. We can rewrite the density period-by-period in terms of the transitions, the hidden Markov assumption implies that the G_t from the are constant functions of $\mathcal{F}_{t-1}$. Since the filtrations are measurable with respect to $x_1, \ldots, x_{t-1}$:

$$\int_{\mathbb{R}^{T \times D}} \sup_t \log \frac{q_T(X)}{p_T(X)} \, dP_T = \int_{\mathbb{R}^{T \times D}} \left(\sup_{x_t, \mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \log \frac{\int_{G_M} \phi \left(x_t \,\middle|\, x_{t-1}, \delta_t^Q \right) dG_M^Q \left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q \right)}{\int_{G_M} \phi \left(x_t \,|\, x_{t-1}, \delta_t^P \right) dG_P \left(\delta_t \,|\, \mathcal{F}_{t-1}^P \right)} \right) dP_T. \tag{113}$$

Clearly, the supremum with respect to x_t is greater than the average with respect to the x_t, and so this is

$$\geq \int_{\mathbb{R}^{T \times D}} \left(\sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \log \frac{\int_{G_M} \phi \left(x_t \,\middle|\, x_{t-1}, \delta_t^Q \right) dG_M^Q \left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q \right)}{\int_{G_M} \phi \left(x_t \,|\, x_{t-1}, \delta_t^P \right) dG_P \left(\delta_t \,|\, \mathcal{F}_{t-1}^P \right)} \right) dP_T. \tag{114}$$

Let $\mathcal{K} \left(dG^P(\Delta^P), dG^Q(\Delta^Q) \right)$ be a coupling between the joint distributions of Δ^P and Δ^Q. Note, this a coupling over the entire sequence

of δ_t^P and δ_t^Q. Then we can rewrite above as

$$\int_{\mathbb{R}^{T\times D}} \int_{G^P\times G^Q} \left(\sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \log \frac{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right) dG_M^Q\left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q\right)}{\int_{G_M} \phi\left(x_t \mid x_{t-1}, \delta_t^P\right) dG_P\left(\delta_t \mid \mathcal{F}_{t-1}^P\right)} \right) d\mathcal{K}\left(dG^P(\Delta^P), dG^Q(\Delta^Q)\right) dP_T. \tag{115}$$

Conditional on both Δ^P and Δ^Q, $\mathcal{F}_{t-1}^P$ and $\mathcal{F}_{t-1}^Q$ contain no information regarding δ_t^P and δ_t^Q. By the law of iterated expectations, we can rewrite this as integral with respect to the joint distribution as we did above. The reason that we can pull the logarithm through integral is because conditional on δ_t^Q, and δ_t^Q, the integral contains only a single element. We then factor $\mathcal{K}$:

$$= \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P} \log \left(\prod_{t=1}^{T} \int_{G_t^P\times G_t^Q} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right)}{\phi\left(x_t \mid x_{t-1}, \delta_t^P\right)} d\mathcal{K}\left(dG_t^P(\delta_t^P), dG_t^Q(\delta_t^Q) \,\middle|\, \mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P\right) \right) dP_T. \tag{116}$$

The Markov assumption on x_t, z_t implies that the δ_t^P and δ_t^Q will be hidden Markov as well. In addition since the δ_t are almost surely discrete, we can assume without loss of generality that the hidden state that makes x_t be a hidden Markov is almost surely discrete.

$$\int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P} \sum_{t=1}^{T} \log \left(\int_{G_M^P\times G_M^Q} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right)}{\phi\left(x_t \mid x_{t-1}, \delta_t^P\right)} d\mathcal{K}\left(dG_M^P(\delta_t^P), dG_M^Q(\delta_t^Q) \,\middle|\, \delta_{t-1}^Q, \delta_{t-1}^P, x_{t-1}, z_{t-1}\right) \right) dP_T. \tag{117}$$

We apply Jensen's inequality to the logarithm, and pull the supremum through the sum because it does not depend upon t:

$$\geq \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \sup_{\mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P} \left(\int_{G_M^P\times G_M^Q} \log \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right)}{\phi\left(x_t \mid x_{t-1}, \delta_t^P\right)} d\mathcal{K}\left(dG_M^P(\delta_t^P), dG_M^Q(\delta_t^Q) \,\middle|\, \delta_{t-1}^Q, \delta_{t-1}^P, x_{t-1}, z_{t-1}\right) \right) dP_T. \tag{118}$$

The terms inside the sum are all the same after interchanging the sum and the integral:

$$= T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}_{t-1}^Q, \mathcal{F}_{t-1}^P} \left(\int_{G_M^P\times G_M^Q} \log \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right)}{\phi\left(x_t \mid x_{t-1}, \delta_t^P\right)} d\mathcal{K}\left(dG_M^P(\delta_t^P), dG_M^Q(\delta_t^Q) \,\middle|\, \delta_{t-1}^Q, \delta_{t-1}^P, x_{t-1}, z_{t-1}\right) \right) dP_T. \tag{119}$$

since couplings preserve marginals, and the δ_t are almost surely discrete, by the law of iterated expectations

$$= T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} dP_T. \tag{120}$$

We can factor P_T, and pull the supremum outside of the expectation, because we have finitely many terms. Given $\mathcal{F}^P_{t-1}$ and $\mathcal{F}^Q_{t-1}$, the only place where the two terms share a value is x_t. The other terms integrated to one.

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \int_{\mathbb{R}^D} \cdots \int_{R^D} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} \prod_{t=1}^{T} p_T\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right) dx_t \tag{121}$$

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \int_{\mathbb{R}^D} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} p_T\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right) dx_t \tag{122}$$

This is the formula for the Kullback-Leibler divergence between the conditional expectations.

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \mathrm{D_{KL}}\left(q_T\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right) \middle|\middle| p_T\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right)\right) \tag{123}$$

By Lemma 8 the supremum of the Kullback-Leibler divergences is proportional to squared Hellinger distance. By Proposition 11 the initial equation is bounded above by $CT\epsilon^2$: (The T comes from using non-rescaled data.)

$$\sup_t h^2\left(q_T\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right), p_T\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right)\right) \leq \sup_t \frac{1}{T} h^2(q_T(X), p_T(X)) \leq C\frac{T}{T}\epsilon^2 = C\epsilon^2. \tag{124}$$

□

Lemma 1 (Replacing Θ_T with a Dirichlet Process)**.** *Let Q be a mixture distribution representable as an integral with respect to the Θ_T process defined in Definition 2. Then Q has a mixture representation as an integral with respect to the Dirichlet process.*

Proof. We can represent a Dirichlet process as $\Pr(x) = \sum_{i=1}^{\infty} \beta_i \delta_{x_i}(x)$, where δ_{x_i} is a indicator function with $\delta_{x_i}(x_i) = 1$, and the β_i satisfy a stick-breaking process. In other words, $\beta_i = \beta'_i \prod_{j=1}^{i=1}(1-\beta'_j)$ with $\beta'_j \sim \mathrm{Beta}(1,\alpha)$ for some positive scalar α. Consider the probability

mass function of a row of Θ_T, θ_t. Then $\Pr(|i| = 1) = b\prod_{j=1}^{j-1}(1-b)$. Since draws from the beta distribution lie in $(0, 1)$ with probability 1, these two stick-breaking processes are clearly mutually absolutely continuous. If we take $x \in \{-1, 1\}$ with the probability $1/2$ each as the Dirichlet base measure, the process are mutually absolutely continuous after possibly extending the space so that the Beta random variables are well-defined.

Because these two processes are mutually absolutely continuous, a Radon-Nikodym derivative exists because both measures are σ-finite. Since the rows are independent, and Dirichlet processes are normalized random measures, (Lin, Grimson, and Fisher (????)), we can extend this to the entire Θ_T process. Consequently, any process that is representable as an integral with respect to Θ_T can be represented as an integral with respect to to a Dirichlet process. □

Appendix C Contraction Rates

C.1 Constructing Exponentially Consistent Tests with Respect to h_∞

Lemma 2 (Exponentially consistent tests exist with respect to h_∞)**.** *There exist tests Υ_T and universal constants $C_2 > 0$, $C_3 > 0$ satisfying for every $\epsilon > 0$ and each $\xi_1 \in \Xi$ and true parameter ξ^P with $h_\infty(\xi_1, \xi^P)$:*

1. $$P_T\left(\Upsilon_T \,\middle|\, \xi^P\right) \leq \exp(-C_2 T\epsilon^2) \tag{19}$$

2. $$\sup_{\xi\in\Xi,\, e_n(\xi_1,\xi)<\epsilon C_3} P_T\left(1-\Upsilon_T \,\middle|\, \xi^P\right) \leq \exp(-C_2 T\epsilon^2) \tag{20}$$

Proof. As done in the proof of the representing the Markov data, we can represent the joint density as a product density conditionally on a sequence of latent mixing measures G_t:

$$f\left(X_T \,|\, G_1, \ldots G_T\right) = \prod_{t=1}^{T} \int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f). \tag{125}$$

Since we are letting G_t differ every period, we can do this for both Q_T and P_T. We can define a distance between these conditional densities as the sum of the squared Hellinger distances between each period. This is not the same as the Hellinger distance between the joint measures:

$$\begin{aligned} &h^2_{\text{avg}}\left(f\left(X \,\middle|\, \{G_t^f\}\right), g\left(X \,|\, \{G_t^g\}\right)\right) \\ &\quad := \frac{1}{T}\sum_{t=1}^{T} h^2\left(\int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi\left(x_t \,|\, \delta_t^g\right) dG_t^g(\delta_t^g)\right). \end{aligned} \tag{126}$$

Then by (Birgé, 2013, Corollary 2), there exists a test ϕ_T that satisfies the following:[18]

$$\begin{aligned} &\Pr_T\left(\phi_T(X) \,\middle|\, \left\{G_t^f, G_t^g\right\}\right) \\ &\quad \leq \exp\left(-\frac{1}{3} T h^2_{\text{avg}}\left(\int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi\left(x_t \,|\, \delta_t^g\right) dG_t^g(\delta_t^g)\right)\right) \end{aligned} \tag{127}$$

and

$$\begin{aligned} &\Pr_T\left(1-\phi_T(X) \,\middle|\, \left\{G_t^f, G_t^g\right\}\right) \\ &\quad \leq \exp\left(-\frac{1}{3} T h^2_{\text{avg}}\left(\int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi\left(x_t \,|\, \delta_t^g\right) dG_t^g(\delta_t^g)\right)\right). \end{aligned} \tag{128}$$

The issue with these equations is that they are not in terms of h_∞ and only hold conditionally. The reason that we can get around this is because they hold for all G_t^f and for all G_t^g. Consequently, we can take the infimum of both sides, and bound the right-hand side of both equations by

[18] To map his notation into ours, take his $z = 0$, and take his measure R equal to P. Equation (127) is obvious then, and (128) follows by taking the exponential of both sides in the inequality inside the probability and rearranging.

$$\frac{T}{3}\sup_{\{(G_t^f,G_t^g)\}} h_{\text{avg}}^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right) \tag{129}$$

for any length T sequence. This equals the least favorable G_t^f and G_t^g repeated T times. This joint distribution exists in our set because we are not placing any restrictions on the dynamics besides ergodicity, and stationary distribution are clearly ergodic. Hence, this equals

$$=\frac{T}{3}\frac{1}{T}\sum_{t=1}^{T}h^2\left(\int_{G_{sup}^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_{sup}^f(\delta_t^f),\int_{G_{sup}^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_{sup}^g(\delta_t^g)\right). \tag{130}$$

The terms inside the sum are all the same:

$$=\frac{T}{3}h^2\left(\int_{G_{sup}^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_{sup}^f(\delta_t^f),\int_{G_{sup}^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_{sup}^g(\delta_t^g)\right) \tag{131}$$

$$=\frac{T}{3}\sup_{(G_t^f,G_t^g)}h^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right) \tag{132}$$

$$=\frac{T}{3}h_\infty^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right). \tag{133}$$

Taking the supremum over G_t^f and G_t^g is equivalent to taking supremum over $\mathcal{F}_{t-1}^f$ and $\mathcal{F}_{t-1}^g$ because the G_t^f and G_t^g are measurable functions of the later, and we are taking the supremum outside of the integral. They both span the same information sets:

$$=\frac{T}{3}h_\infty^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f),\int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right). \tag{134}$$

Since we can bound the error probabilities in both directions, using exponentially consistent tests, we have shown both items in Lemma 2 hold. □

C.2 Bounding the Posterior Divergence

Proposition 6 (Bounding the Posterior Divergence)**.** *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T \coloneqq \sum_k \Pi_{k,t}\phi(x_t\,|\,\mu_t,\Sigma_t)$ with finite means and finite positive-definite covariances. Let $\Xi_T \subset \Xi$ and $T\to\infty$. Let Q_T be a mixture approximation with $\frac{K_T^i}{\eta_T}$ components. Assume the following condition holds with probability $1-2\delta$ for $\delta>0$ and constants C and $i\in\mathbb{N}$:*

$$\sup_t h\left(q_T\left(x_t\,\middle|\,\mathcal{F}_{t-1}^Q\right),p_T\left(x_t\,|\,\mathcal{F}_{t-1}^P\right)\right)<C\eta_T. \tag{21}$$

Let $\epsilon_{i,T}\coloneqq\frac{\log(T)^{\sqrt{i}}}{\sqrt{T}}$. Then the following two conditions hold with probability $1-2\delta$ with respect to

the prior

$$\sup_{\epsilon_i \geq \epsilon_{T,i}} \log N\left((\epsilon_i, \left\{\xi \in \Xi_T \,\middle|\, h_\infty(\xi, \xi^P) \leq \epsilon_i\right\}, h_\infty\right) \leq T\epsilon_{T,i}^2, \tag{22}$$

and

$$\mathcal{Q}_T\left(B_T\left(\xi^P, \epsilon_{T,i}, 2\right) \,\middle|\, X_T\right) \geq C \exp\left(-C_0 T \epsilon_{T,i}^2\right). \tag{23}$$

Proof. We are looking at locally asymptotically normal models, as discussed in Lemma 8, and we bind the Hellinger distance and Kullback-Leibler divergence in terms of $(x_t - \mu_t)'\Sigma_t^{-1}(x_t - \mu_t)$. In addition, the supremum of the deviations is clearly greater than the average of the deviations, and so the h_∞-norm forms smaller balls than both $\mathrm{D}_{\mathrm{KL}}(f \,||\, g)$ and $V_{k,0}$. Consequently, we can replace $B_T(\xi_0, \epsilon_T, 2)$ with $\{\xi \in \Xi \,|\, h_\infty^2(\xi, \xi_0) < T\epsilon_T^2\}$. We use 2 as the last argument of B because we are using $V_{2,0}$, i.e., effectively the 2nd-moment of the Kullback-Leibler divergence.

To prove the result we need to find a sequence $\epsilon_{T,i} \to 0$ that satisfies the following two conditions:

$$\sup_{\epsilon_i > \epsilon_{T,i}} \log N\left(\epsilon_i, \left\{\xi \in \Xi_T \,|\, h_\infty(\xi, \xi_0) \leq \epsilon_i\right\}, h_\infty\right) \leq T\epsilon_{T,i}^2 \tag{135}$$

and

$$\mathcal{Q}_T\left(\left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi, \xi_0) < \epsilon_{T,i}\right\}\right) \geq C \exp\left(-T\epsilon_{T,i}^2\right). \tag{136}$$

These two conditions work in opposite directions. The first criterion is easier to satisfy the larger $\epsilon_{T,i}$ is, but to achieve a fast rate of convergence we want a small $\epsilon_{T,i}$ in the second condition.

By assumption, there exists a covering with $\frac{K_T^i}{\eta_T}$ components such that the following holds:

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right)\right) < C\sqrt{1 + \log\frac{1}{\delta}}\eta_T. \tag{137}$$

Equation (136) is satisfied if

$$\eta_T^2 \geq \frac{C_0}{1 + \log(1/\delta)} \exp\left(-T\epsilon_{T,i}^2\right) \propto \exp\left(-T\frac{\log(T)^i}{T}\right) = \frac{1}{T^i}. \tag{138}$$

To satisfy (135), h_∞^2 must be bounded below and decline exponentially fast. The expressions above hold for any $\eta_T^* \geq \eta_T$. Let $\eta_T^* = \frac{\log(T)^n}{T^n}$. We know there exists a covering with $K_T = \frac{\log(T)^i}{\eta_T^*}$ components. This implies that

$$K_T = \frac{\log(T)^i}{\eta_T^*} = \frac{\log(T)^i}{\log(T)^i / T^i} = T^i. \tag{139}$$

This K_T is proportional to the number of terms we are using, and the bracketing number is proportional to the covering number:

$$N\left(\epsilon_n, \left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi, \xi_0)\right\} \leq \epsilon_i, h_\infty^2\right) \leq T^i = \exp\left(\log\left(T^i\right)\right) = \exp\left(T\epsilon_{T,i}^2\right). \tag{140}$$

Taking logarithms of both sides of (140) finishes the proof.

□

C.3 Contraction Rate of the Marginal Density

Theorem 8 (Contraction Rate of the Marginal Density)**.** *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T \coloneqq \sum_k \pi_k \phi\left(\cdot \,|\, \mu_t, \Sigma_t\right)$ with finite mean and finite variance. Let $T \to \infty$, then the following holds with $\epsilon_T = \sqrt{\frac{\log(T)}{T}}$ and probability $1 - 2\delta$ with respect to the prior. There exists a constant C independent of T such that the posterior over the transition densities constructed above and the true transition density satisfies*

$$P_T\left(\mathcal{Q}_T\left(h\left(p_T\left(x_t\right), q_T\left(x_t\right)\right) \geq C\epsilon_T \,|\, X\right)\right) \to 0.$$

Proof. To prove this result, note that the existence of exponentially consistent tests with respect to the average Hellinger metric for independent data is well-known ((Ghosal and van der Vaart, 2017, 540)). We can represent the density as product density by a resampling argument as we did in the construction of the sieve.

Having done that we can verify the conditions in Proposition 6. If we take $i = 1$ in (21), Theorem 3 implies the necessary bound on the sieve complexity exists. In addition, since h_∞ is bounded above by the Hellinger distance, h, the conclusions of Proposition 6 trivially go through in this weaker topology.

This verifies the three conditions in Theorem 5 on a set with with probability $1 - 2\delta$ with respect to the prior. This then gives us the posterior contraction rate $\epsilon_T = \sqrt{\frac{\log T}{T}}$.

□

C.4 Contraction Rate of the Transition Density

Theorem 7 (Contraction Rate of the Transition Density)**.** *Let P_T be a uniformly ergodic Hidden Markov Gaussian process, i.e., $p_T \coloneqq \sum_k \Pi_{t,k} \phi(x_t \,|\, \mu_t, \Sigma_t)$ with finite means and finite positive-definite covariances. Let $T \to \infty$, then the following holds with $\epsilon_T \coloneqq \sqrt{\frac{\log(T)^2}{T}}$ with probability $1 - 2\delta$ with respect to the prior. There exists a constant C independent of T such that the posterior over the transition densities constructed above and the true transition density satisfies*

$$P_T\left(\mathcal{Q}_T\left(\sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} h\left(p_T\left(x_t \,|\, \mathcal{F}_{t-1}^P\right), q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right) \geq C\epsilon_T \,\middle|\, X_T\right)\right) \to 0.$$

Proof. The proof of this is essentially identical to the marginal density case, mutatis mutandis. Lemma 2 implies the that h_∞ has the required exponentially consistent tests. We verify the conditions in Proposition 6. If we take $i = 2$ in (21), Theorem 4 implies the necessary bound on the sieve complexity exists.

This verifies the three conditions in Theorem 5 on a set with with probability $1-2\delta$ with respect to the prior. This then gives us the posterior contraction rate $\epsilon_T = \sqrt{\frac{\log(T)^2}{T}}$.

□

Appendix D Macroeconomic Empirical Results

Figure 9: Unemployment Rate

(a) Posterior Density (b) PIT Histogram (c) PIT ACF

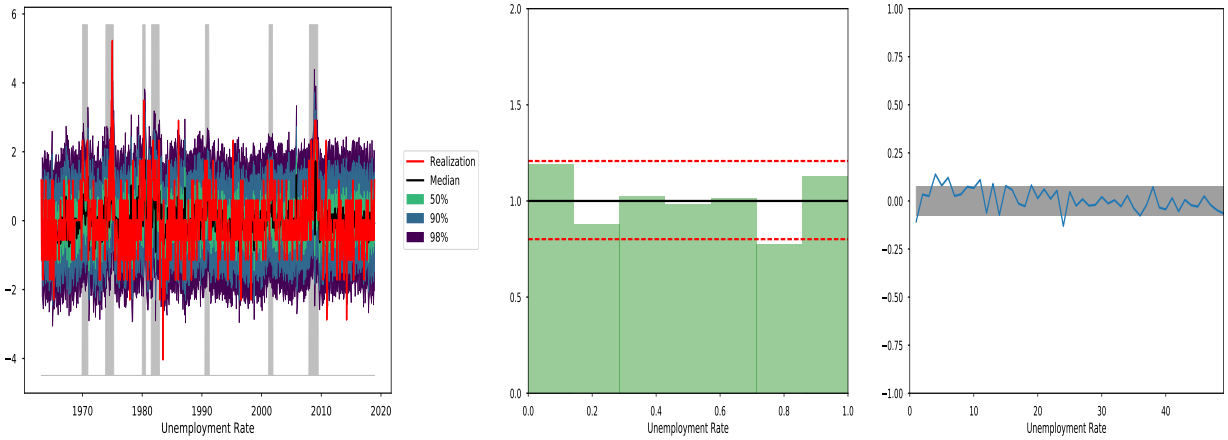


Figure 10: Housing Supply

(a) Posterior Density (b) PIT Histogram (c) PIT ACF

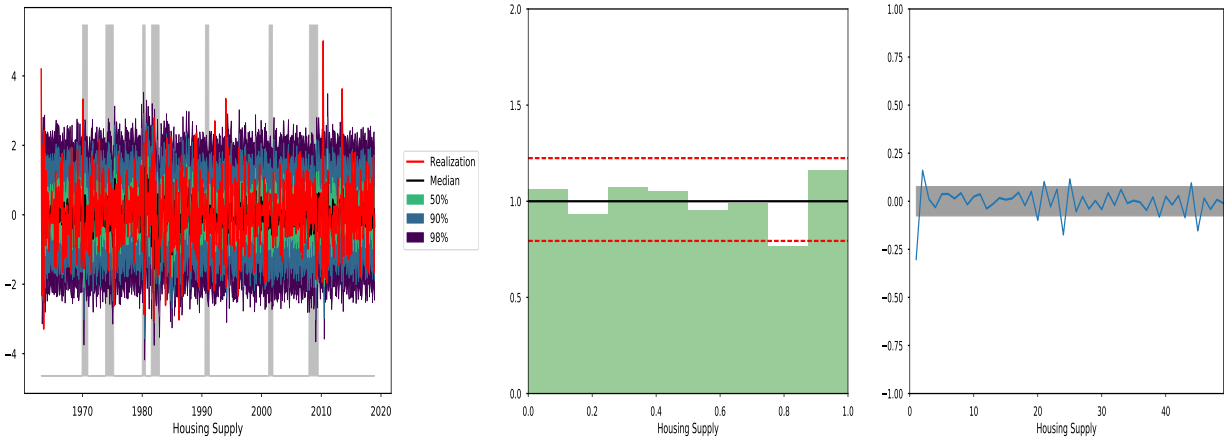


Figure 11: Industrial Production

(a) Posterior Density (b) PIT Histogram (c) PIT ACF

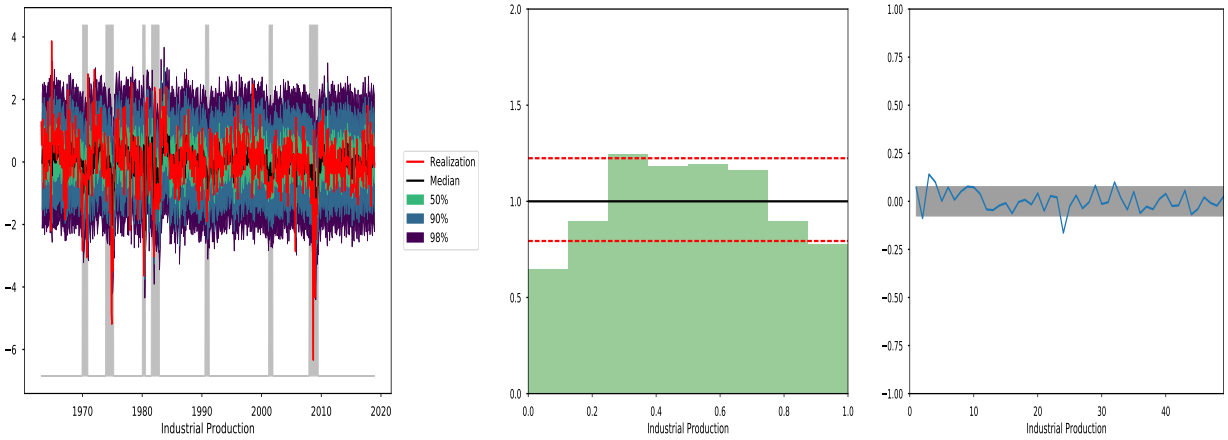


Figure 12: Long-Term Interest Rate

(a) Posterior Density (b) PIT Histogram (c) PIT ACF

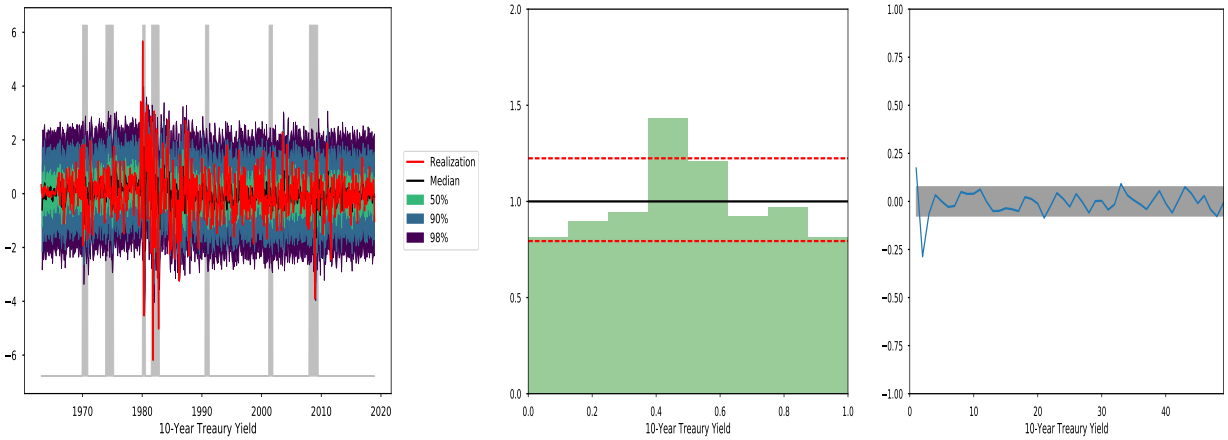


Figure 13: Money Supply (M2)

(a) Posterior Density (b) PIT Histogram (c) PIT ACF

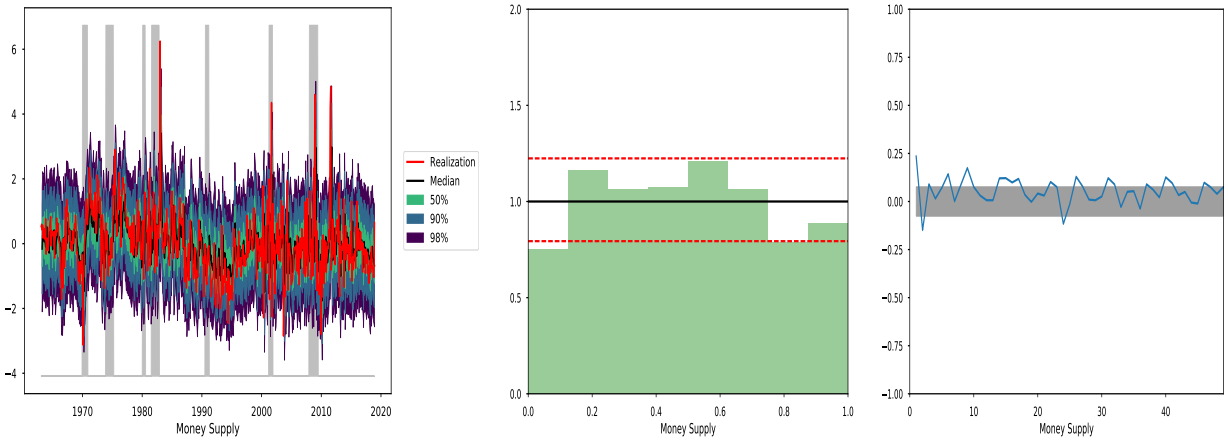
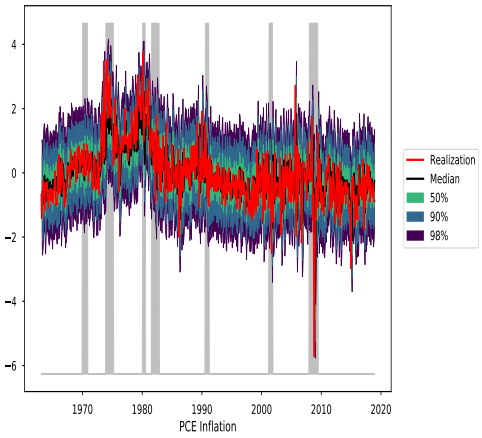
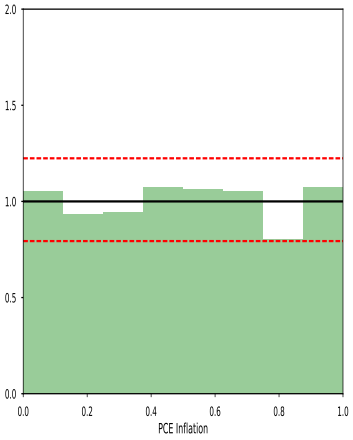


Figure 14: Personal Consumption Expenditures (PCE) Inflation

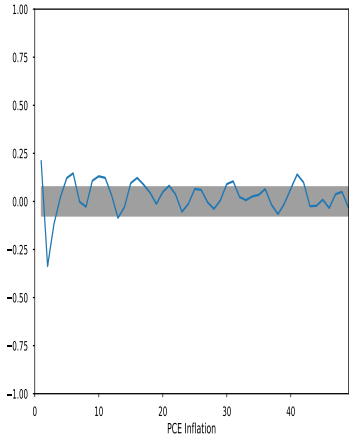
(a) Posterior Density



(b) PIT Histogram



(c) PIT ACF



Appendix E Posterior Derivations

E.1 Component Coefficient Posterior

Let X_k be the $T_k \times N$ vector and Y_k be the $T_k \times D$ vector of data in component K. This implies that Σ_k is a $D \times D$ matrix and β_k is an $N \times D$ matrix.[19] Meanwhile, V is a $D \times D$ matrix and U is a $N \times N$ matrix.

The joint density is

$$\Pr\left(Y_k, \beta_k, \Sigma_k \mid X_k\right) = \exp\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\left(\beta_k - \bar{\beta}\right)' U^{-1}\left(\beta_k - \bar{\beta}\right)\right\}\right) \exp\left(-\frac{1}{2}\operatorname{tr}\left\{\left(Y_k - X_k\beta_k\right)\Sigma_k^{-1}\left(Y_k - X_k\beta_k\right)'\right\}\right)$$
$$\frac{|\Sigma_k|^{-T_k/2}}{(2\pi)^{T_k/2}} \frac{1}{\sqrt{(2\pi)^{ND}|V|^N|U|^D}} \frac{|(\mu_1 - 2)\Omega|^{\nu/2}}{\sqrt{2^{\nu D}}\Gamma_D(\frac{\nu}{2})} |\Sigma_k|^{-\frac{\nu+D+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left\{(\mu_1 - 2)\Omega\Sigma_k^{-1}\right\}\right) \tag{141}$$

By the additivity and circular commutativity of the trace, and associativity of matrix multiplication:

$$\propto |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\left(\beta_k - \bar{\beta}\right)' U^{-1}\left(\beta_k - \bar{\beta}\right)\right\}\right) \exp\left(-\frac{1}{2}\operatorname{tr}\left\{\left(\left(Y_k - X_k\beta_k\right)'\left(Y_k - X_k\beta_k\right) + (\mu_1 - 2)\Omega\right)\Sigma_k^{-1}\right\}\right). \tag{142}$$

Combining the two kernels of β_k and expanding gives

$$\propto |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\left(\left(\beta_k - \bar{\beta}\right)' U^{-1}\left(\beta_k - \bar{\beta}\right)\right) + \left(\left(Y_k - X_k\beta_k\right)'\left(Y_k - X_k\beta_k\right) + (\mu_1 - 2)\Omega\right)\Sigma_k^{-1}\right\}\right) \tag{143}$$
$$= |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\left(\beta_k' U^{-1}\beta_k - 2\beta_k' U^{-1}\bar{\beta} + \bar{\beta}' U^{-1}\bar{\beta}\right) + \Sigma_k^{-1}\left(Y_k'Y_k - 2\beta_k' X_k' Y_k + \beta_k' X_k' X_k\beta_k + (\mu_1 - 2)\Omega\right)\right\}\right). \tag{144}$$

Isolating the terms that have a β in them:

$$= \exp\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\left(-2\beta_k' U^{-1}\bar{\beta} + \beta_k' U^{-1}\beta_k\right) + \Sigma_k^{-1}\left(-2\beta_k' X_k' Y_k + \beta_k' X_k' X_k\beta_k\right) + V^{-1}\bar{\beta}' U^{-1}\bar{\beta} + \Sigma_k^{-1}\left(Y_k'Y_k + (\mu_1 - 2)\Omega\right)\right\}\right).$$
$$|\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \tag{145}$$

[19] The likelihood in (141) is correct because the trace is the sum of the diagonal elements.

Rewriting the traces in terms of the vectorization operator:

$$= \exp\left(-\frac{1}{2}\left(\operatorname{tr}\{V^{-1}(-2\beta_k' U^{-1}\bar{\beta})\} + \operatorname{vec}\{\beta_k\}' \operatorname{vec}\{U^{-1}\beta_k V^{-1}\} \operatorname{tr}\{\Sigma_k^{-1}(-2\beta_k' X_k' Y_k)\} + \operatorname{vec}\{\beta_k\}' \operatorname{vec}\{X_k' X_k \beta_k \Sigma_k^{-1}\}\right)\right)$$
$$\exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\bar{\beta}' U^{-1}\bar{\beta} + \Sigma_k^{-1}\left(Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right)|\Sigma_k|^{-\frac{\nu+D+T+1}{2}}.$$

Exploiting the relationship between vectorization and the Kronecker product and then combining squared terms:

$$\propto \exp\left(\operatorname{tr}\{\beta_k'\left(U^{-1}\bar{\beta}V^{-1} + X_k' Y_k \Sigma_k^{-1}\right)\} - \frac{1}{2}\operatorname{tr}\{\left(\left(V^{-1}\otimes U^{-1}\right) + \left(\Sigma_k^{-1}\otimes X_k' X_k\right)\right)\operatorname{vec}\{\beta_k\}\operatorname{vec}\{\beta_k\}'\}\right)$$
$$\exp\left(-\frac{1}{2}\operatorname{tr}\{V^{-1}\bar{\beta}' U^{-1}\bar{\beta} + \Sigma_k^{-1}\left(Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right)|\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{146}$$

If we assume that $V = \Sigma_k$, we can simplify this as

$$= \exp\left(\operatorname{tr}\{\beta_k'\left(U^{-1}\bar{\beta} + X_k' Y_k\right)\Sigma_k^{-1}\} - \frac{1}{2}\operatorname{tr}\{\left(\Sigma_k^{-1}\otimes\left(U^{-1} + X_k' X_k\right)\right)\operatorname{vec}\{\beta_k\}\operatorname{vec}\{\beta_k\}'\}\right) \tag{147}$$
$$\exp\left(-\frac{1}{2}\operatorname{tr}\{\Sigma_k^{-1}\left(\bar{\beta}' U^{-1}\bar{\beta} + Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right)|\Sigma_k|^{-\frac{\nu+D+T+1}{2}}$$
$$= \exp\left(\operatorname{vec}\{\beta_k\}'\operatorname{vec}\{\left(U^{-1}\bar{\beta} + X_k' Y_k\right)\Sigma_k^{-1}\} - \frac{1}{2}\operatorname{vec}\{\beta_k\}'\left(\Sigma_k^{-1}\otimes\left(U^{-1} + X_k' X_k\right)\right)\operatorname{vec}\{\beta_k\}\right)$$
$$\exp\left(-\frac{1}{2}\operatorname{tr}\{\Sigma_k^{-1}\left(\bar{\beta}' U^{-1}\bar{\beta} + Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right)|\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{148}$$

We now use the multivariate completion of squares: $u'Au - 2\alpha' u = (u - A^{-1}\alpha)' A (u - A^{-1}\alpha) - \alpha' A^{-1}\alpha$. Let $Z_k \coloneqq (U^{-1}\bar{\beta} + X_k' Y_k)$ and $W_k \coloneqq (U^{-1} + X_k' X_k)$. We can now rewrite (148) as

$$= \exp\left(-\frac{1}{2}\left(\operatorname{vec}\{\beta_k\} - (\Sigma_k^{-1}\otimes W_k)^{-1} Z_k \Sigma_k^{-1}\right)'\left(\Sigma_k^{-1}\otimes W_k\right)\left(\operatorname{vec}\{\beta_k\} - (\Sigma_k^{-1}\otimes W_k)^{-1} Z_k \Sigma_k^{-1}\right)\right)$$
$$\exp\left(\frac{1}{2}\Sigma_k^{-1} Z_k' (\Sigma_k^{-1}\otimes W_k)^{-1} Z_k \Sigma_k^{-1}\right)\exp\left(-\frac{1}{2}\operatorname{tr}\{\Sigma_k^{-1}\left(\bar{\beta}' U^{-1}\bar{\beta} + Y_k' Y_k + (\mu_1 - 2)\Omega\right)\}\right)|\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{149}$$

I now eliminate all of the Kronecker products:

$$= \exp\left(-\frac{1}{2}\operatorname{vec}\left\{\beta_k - W_k^{-1}Z_k\right\}'\operatorname{vec}\left\{W_k\left(\beta_k - W_k^{-1}Z_k\right)\Sigma_k^{-1}\right\}\right) \tag{150}$$

$$\exp\left(\frac{1}{2}\operatorname{vec}\left\{(U^{-1}\bar{\beta} + X_k'Y_k)\Sigma_k^{-1}\right\}'\operatorname{vec}\left\{W_k^{-1}Z_k\right\} - \frac{1}{2}\operatorname{tr}\left\{\Sigma_k^{-1}\left(\bar{\beta}'U^{-1}\bar{\beta} + Y_k'Y_k + (\mu_1 - 2)\Omega\right)\right\}\right)|\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{151}$$

We rewrite this in terms of the traces, reorder some of the terms, and substitute the definitions of Z_k and W_k back in:

$$= \exp\left(-\frac{1}{2}\operatorname{tr}\left\{\Sigma_k^{-1}\left(\beta_k - \left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)\right)'\left(U^{-1} + X_k'X_k\right)\left(\beta_k - \left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)\right)\right\}\right) \tag{152}$$

$$\exp\left(-\frac{1}{2}\operatorname{tr}\left\{\Sigma_k^{-1}\left(\left(\bar{\beta}'U^{-1}\bar{\beta} + Y_k'Y_k + (\mu_1 - 2)\Omega\right) - \left(U^{-1}\bar{\beta} + X_k'Y_k\right)'\left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)\right)\right\}\right)|\Sigma_k|^{-\frac{\nu+D+T+1}{2}}.$$

The first expression is kernel of a matrix-normal distribution. The mean is $\left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)$, and the two covariance parameters are Σ_k, and $\left(U^{-1} + X_k'X_k\right)^{-1}$. The second expression is the kernel of a Inverse-Wishart distribution. Its scale parameter is $\left(\bar{\beta}'U^{-1}\bar{\beta} + Y_k'Y_k + (\mu_1 - 2)\Omega\right) - \left(U^{-1}\bar{\beta} + X_k'Y_k\right)'\left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right)$. It has $\mu_1 + D - 1 + T_k$ degrees of freedom. To see the intuition behind this, note that if U^{-1} and Ω both equal zero, this equals $Y_k'Y_k - Y_k'X_k'(X_k'X_k')^{-1}X_kY_k$, i.e., the sum of squared residuals. Since the β_k parameter does not show up in the second expression, we can draw from the posterior by drawing the Σ_k from its marginal posterior, and then drawing from the posterior of β_k conditional on Σ_k .

E.2 Hypermean Posterior with Heteroskedastic Data

We now compute the posterior of the hierarchical mean for the coefficients conditional on the covariance matrices, $\{\Sigma_k\}_{k=1}^{K_T}$:

$$\Pr\left(\{\beta\}_{k=1}^{K}, \bar{\beta}, \{\Sigma\}_{k=1}^{K}\right) = \exp\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\left(\bar{\beta} - \beta^{\dagger}\right)'U^{-1}\left(\bar{\beta} - \beta^{\dagger}\right)\right\}\right)\exp\left(\sum_{k=1}^{K} -\frac{1}{2}\operatorname{tr}\left\{\Sigma_k^{-1}\left(\beta_k - \bar{\beta}\right)'U^{-1}\left(\beta_k - \bar{\beta}\right)\right\}\right)$$

$$\sqrt{(2\pi)^{ND}|U|^{D}}|U|^{-\frac{\nu_U+N+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\left\{\Psi_U U^{-1}\right\}\right)\prod_{k=1}^{K}\frac{1}{\sqrt{(2\pi)^{ND}|\Sigma_k|^{N}|U|^{D}}} \tag{153}$$

Dropping all of the terms that contain neither $\bar{\beta}$ nor U:

$$\propto |U|^{-\frac{\nu_U+N+(K+1)D+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\left(\bar{\beta}-\beta^{\dagger}\right)' U^{-1}\left(\bar{\beta}-\beta^{\dagger}\right)+\sum_{k=1}^{K}\Sigma_k^{-1}(\bar{\beta}-\beta_k)'U^{-1}(\bar{\beta}-\beta_k)\right\}\right)\exp\left(-\frac{1}{2}\operatorname{tr}\{\Psi_U U^{-1}\}\right). \tag{154}$$

Expanding out the terms and dropping terms that do not involve $\bar{\beta}$ or U:

$$\begin{aligned}\propto \exp&\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\bar{\beta}'U^{-1}\bar{\beta}-2V^{-1}\beta^{\dagger\prime}U^{-1}\bar{\beta}+V^{-1}\beta^{\dagger\prime}U^{-1}\beta^{\dagger}+\sum_{k=1}^{K}\Sigma_k^{-1}(\bar{\beta}'U^{-1}\bar{\beta}-2\beta_k'U^{-1}\bar{\beta}+\beta_k'U^{-1}\beta_k\right\}\right) \\ &|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{\Psi_U U^{-1}\}\right).\end{aligned} \tag{155}$$

Exploiting properties of the trace and vectorization, where $B \coloneqq \operatorname{vec}\{\bar{\beta}\}$:

$$\begin{aligned}\propto \exp&\left(-\frac{1}{2}\operatorname{vec}\left\{\beta^{\dagger}\right\}'\left(V^{-1}\otimes W^{-1}\right)B+\operatorname{vec}\left\{W^{-1}\beta^{\dagger\prime}V^{-1}\right\}'B-\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}\{(\Sigma_k^{-1}\otimes U^{-1})BB'\}+\operatorname{vec}\left\{\sum_{k=1}^{K}U^{-1}\beta_k\Sigma_k^{-1}\right\}'B\right) \\ &|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\left\{V^{-1}\beta^{\dagger\prime}U^{-1}\beta^{\dagger}+\sum_{k=1}^{K}\Sigma_k^{-1}\beta_k'U^{-1}\beta_k+\Psi_U U^{-1}\right\}\right).\end{aligned} \tag{156}$$

We can simplify using the circular commutativity of the trace:

$$\begin{aligned}\propto \exp&\left(-\frac{1}{2}\operatorname{vec}\{\bar{\beta}\}'\left(\left(\sum_{k=1}^{K}\Sigma_k^{-1}\right)\otimes U^{-1}+V^{-1}\otimes U^{-1}\right)\operatorname{vec}\{\bar{\beta}\}+\operatorname{vec}\left\{U^{-1}\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}U^{-1}\beta_k\Sigma_k^{-1}\right\}'\operatorname{vec}\{\bar{\beta}\}\right) \\ &|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\left\{\beta^{\dagger}V^{-1}\beta^{\dagger\prime}U^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k'U^{-1}+\Psi_U U^{-1}\right\}\right).\end{aligned} \tag{157}$$

Collecting terms:

$$\propto \exp\left(-\frac{1}{2}\operatorname{vec}\{\bar{\beta}\}'\left(\left(\sum_{k=1}^{K}\Sigma_k^{-1}+V^{-1}\right)\otimes U^{-1}\right)\operatorname{vec}\{\bar{\beta}\}+\operatorname{vec}\left\{U^{-1}\left(\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)\right\}'\operatorname{vec}\{\bar{\beta}\}\right) \tag{158}$$

$$|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\left\{\left(\beta^{\dagger}V^{-1}\beta^{\dagger\prime}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k'+\Psi_U\right)U^{-1}\right\}\right)$$

$$\propto \exp\left(-\frac{1}{2}\operatorname{tr}\left\{\left(\sum_{k=1}^{K}\Sigma_k^{-1}+V^{-1}\right)\bar{\beta}'U^{-1}\bar{\beta}+\left(\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)'U^{-1}\bar{\beta}\right\}\right) \tag{159}$$

$$|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\left\{\left(\beta^{\dagger}V^{-1}\beta^{\dagger\prime}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k'+\Psi_U\right)U^{-1}\right\}\right).$$

We now vectorize the first line of (159) after using the circular commutativity of the trace to the square term. We drop the second line for now to simplify the exposition. We will bring it back in later. This gives

$$\exp\left(-\frac{1}{2}\operatorname{vec}\{\bar{\beta}\}'\left(\left(\sum_{k=1}^{K}\Sigma_k^{-1}+V^{-1}\right)\otimes U^{-1}\right)\operatorname{vec}\{\bar{\beta}\}-2\operatorname{vec}\left\{U^{-1}\left(\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)\right\}'\operatorname{vec}\{\bar{\beta}\}\right) \tag{160}$$

We then apply the multivariate equation of squares, and let $Z \coloneqq (\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1})$ and $W \coloneqq (\sum_{k=1}^{K}\Sigma_k^{-1}+V^{-1})$:

$$=\exp\left(-\frac{1}{2}\left(\operatorname{vec}\{\bar{\beta}\}-\left(W\otimes U^{-1}\right)^{-1}\operatorname{vec}\{U^{-1}Z\}\right)\left(W\otimes U^{-1}\right)\left(\operatorname{vec}\{\bar{\beta}\}-\left(W\otimes U^{-1}\right)^{-1}\operatorname{vec}\{U^{-1}Z\}\right)\right)$$
$$\exp\left(\frac{1}{2}\operatorname{vec}\{U^{-1}Z\}'\left(Z\otimes U^{-1}\right)^{-1}\operatorname{vec}\{U^{-1}Z\}\right) \tag{161}$$

We can simplify the vectorization.

$$=\exp\left(-\frac{1}{2}\operatorname{vec}\{\bar{\beta}-ZW^{-1}\}\left(W\otimes U^{-1}\right)\operatorname{vec}\{\bar{\beta}-ZW^{-1}\}\right)\exp\left(\frac{1}{2}\operatorname{tr}\{U^{-1}ZW^{-1}Z'\}\right) \tag{162}$$

We can replace the vectorizations with traces.

$$= \exp\left(-\frac{1}{2}\operatorname{tr}\{U^{-1}\left(\bar{\beta} - ZW^{-1}\right)W\left(\bar{\beta} - ZW^{-1}\right)\}\right)\exp\left(\frac{1}{2}\operatorname{tr}\{U^{-1}ZW^{-1}Z'\}\right) \tag{163}$$

Equation (163) is the kernel of a matrix normal distribution given the covariance matrices. We substitute the definitions of W and Z back in. The row matrix covariance is U, the column posterior covariance is $(\sum_{k=1}^{K}\Sigma_k^{-1} + V^{-1})$, and the mean is $(\beta^\dagger V^{-1} + \sum_{k=1}^{K}\beta_k\Sigma_k^{-1})(\sum_{k=1}^{K}\Sigma_k^{-1} + V^{-1})^{-1}$ Note, there is no reason here that β_k cannot itself be a matrix.

To compute the distribution of U, we combine the last lines of (159) and (163). This gives

$$\begin{aligned}|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\Bigg(-\frac{1}{2}\operatorname{tr}\Bigg\{U^{-1}\Bigg(&\beta^\dagger V^{-1}\beta^{\dagger\prime} + \sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k' + \Psi_U \\ &- \left(\beta^\dagger V^{-1} + \sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)\left(\sum_{k=1}^{K}\Sigma_k^{-1} + V^{-1}\right)^{-1}\left(\beta^\dagger V^{-1} + \sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)'\Bigg)\Bigg\}\Bigg)\end{aligned} \tag{164}$$

Clearly, U is marginally inverse-Wishart. It has $\nu_U + (K+1)D$ degrees of freedom, and its scale matrix equals $\beta^\dagger V^{-1}\beta^{\dagger\prime} + \sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k' + \Psi_U - (\beta^\dagger V^{-1} + \sum_{k=1}^{K}\beta_k\Sigma_k^{-1})(\sum_{k=1}^{K}\Sigma_k^{-1} + V^{-1})^{-1}(\beta^\dagger V^{-1} + \sum_{k=1}^{K}\beta_k\Sigma_k^{-1})'$.

E.3 Derivation of the Posterior of the Innovation Covariances' Mean

The product of the relevant likelihood and prior is

$$\Omega \,|\, \{\Sigma_k,\}_{k=1}^{K} \propto \prod_{k=1}^{K}|\Omega|^{\frac{\mu_1+D-1}{2}}\exp\left(-\frac{\mu_1 - 2}{2}\operatorname{tr}\{\Omega\Sigma_k^{-1}\}\right)\cdot|\Omega|^{\frac{\mu_2-2}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{\operatorname{diag}(a_1,\ldots,a_D)^{-1}\Omega\}\right). \tag{165}$$

Since matrix multiplication distributes over matrix addition:

$$= |\Omega|^{\frac{K(\mu_1+D-1)}{2}} \exp\left(-\frac{\mu_1-2}{2}\sum_{k=1}^{K}\operatorname{tr}\{\Omega\Sigma_k^{-1}\}\right)\cdot|\Omega|^{\frac{\mu_2-2}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\{\operatorname{diag}(a_1,\ldots,a_D)^{-1}\Omega\}\right) \tag{166}$$

$$= |\Omega|^{\frac{K(\mu_1+D-1)+\mu_2-2}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left\{\left(\operatorname{diag}(a_1,\ldots,a_D)^{-1}+(\mu_1-2)\sum_{k=1}^{K}\Sigma_k^{-1}\right)\Omega\right\}\right). \tag{167}$$

This is the kernel of a Wishart distribution. That is

$$\Omega \,|\, \{\Sigma_k\}_{k=1}^{K} \sim \mathcal{W}\left(K(\mu_1+D-1)+(\mu_2+D-1), \left(\operatorname{diag}(a_1,\ldots,a_D)^{-1}+(\mu_1-2)\sum_{k=1}^{K}\Sigma_k^{-1}\right)^{-1}\right). \tag{168}$$