Bypassing the Curse of Dimensionality: Feasible Multivariate Density Estimation*

Minsu Chang & Paul Sangrey
University of Pennsylvania †

Current Version
This Version: November 11, 2018

Abstract

Most economic data are multivariate and so estimating multivariate densities is 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 if 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 and a daily financial panel. Our procedure performs well in capturing the data's stylized features such as time-varying volatility and fat-tails.

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

JEL Codes: C11, C14, C32, C55

*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.

†Department of Economics, University of Pennsylvania, The Ronald O. Perelman Center for Political Science and Economics, 133 South 36th Street, Philadelphia, PA 19104-6297. Email: minsuc@sas.upenn.edu and paul@sangrey.io, Web: minsuchang.com and sangrey.io

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 becomes infeasible if the number of series, D, is large. This phenomenon is called the *curse of dimensionality*. In particular, the number of terms required to approximate a general distribution grows exponentially quickly with D, (Stone 1980, 1982). For example, the number of terms a Taylor series approximation uses is proportional to T^D, where T is the number of periods.[1]

Leading examples of commonly used estimators that suffer from this curse of dimensionality include kernel estimation, as surveyed by Ichimura and Todd (2007), and sieve estimation, as surveyed by Chen (2007). The list of possible applications is too vast to enumerate here. In dynamic environments alone, obtaining good conditional densities is pivotal for forecasting, risk measurement, studying heterogeneous agent models, and so on, (Krusell and Smith 1998; Fan 2005; Patton, Ziegel, and Chen 2018).

Several authors have studied Bayesian density estimators, attaining these minimax rates up to a logarithmic factor without picking any tuning parameters, (Ghosal, Ghosh, and Vaart 2000; van der Vaart and van Zanten 2008; Ghosal and van der Vaart 2017). These models typically, though not exclusively, use mixture distributions to form a sieve. There is also developing literature on Bayesian estimation of conditional distributions, (Geweke and Keane 2007; Norets 2010; Pati, Dunson, and Tokdar 2013). These papers view minimax rates through the lens of information theory, characterizing the data generating process's entropy. They find that minimax rates that decline exponentially fast with D, (Yang and Barron 1999).

This exponential increase in the number of parameters makes applying these methods to datasets with more than three or four series prohibitive. However, these minimax rates, as they are known, characterize estimators' worst-case behavior. 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. The easy way to do this would be to restrict the set of data generating processes (DGPs) we consider. For example, parametric models have only a finite number of parameters regardless of T, and so the convergence rate is independent of any fixed D. However, in general, we want to guarantee a fast convergence rate without restricting the set of DGPs we allow.

To do this, we adapt the ideas in the Bayesian compression literature to density estimation by randomly binning the vectors x_t. This rapidly-growing literature in mathematics and statistics studies randomly compressing the data independently of D. It applies a random operator to some multivariate data — X_T — that significantly simplifies the problem. Various authors derive bounds that hold the vast majority of the time with respect to the randomness the compression introduces.

1. In general, the number of terms a sieve requires equals $T^{g(D)}$ for some function g that depends on the set of functions being approximated.

These bounds are useful in our case because even though we must place strong restrictions on the compression, we do not need to place them on the DGP. Johnson and Lindenstrauss (1984) introduced the idea of projecting a Gaussian process into a lower dimensional subspace. More recently, Koop, Korobilis, and Pettenuzzo (2017) introduced these ideas to econometrics using them to estimate a vector autoregression in many dimensions, while Boucheron, Lugosi, and Massart (2013) provide a book-length treatment.

We construct a novel random procedure which determines both the number of bins and which vector x_t is placed in which bin. Having constructed these bins, we bound the tail behavior of the approximation error by noting that binning and clustering refer to the same procedure. Hence, our binning procedure induces a simple mixture representation for both the unconditional and transition densities of both i.i.d. and Markov data. Given a realization of this random clustering, we construct an approximating distribution for each cluster. Since we do not know the true clustering, we average over these approximating distributions.

For any finite 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. 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 even when D is large.

Having constructed these tight bounds on the complexity of our mixture sieve, we convert them into convergence rates for the estimators by adopting the nonparametric Bayesian literature referenced above. In particular, we construct an estimator for the transition density of a Markov process that reduces to a standard Dirichlet Gaussian mixture when there are no dynamics. We then adapt the slice sampling algorithm of Walker (2007) to efficiently sample the infinite mixture approximation. Since the number of mixture components is the key determinant of our model's complexity, 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. We then analyze our estimators' performance in a monthly macroeconomic panel and a daily financial panel. The procedure performs well in capturing the data's stylized features such as time-varying volatility and fat-tails.

To summarize, we show that our estimator converges rapidly — it doe 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 as determined by our data-agnostic random clustering. Even though we cannot beat the minimax rate in general, we show that our estimators will perform well when D is large, even when the true distribution is not smooth.

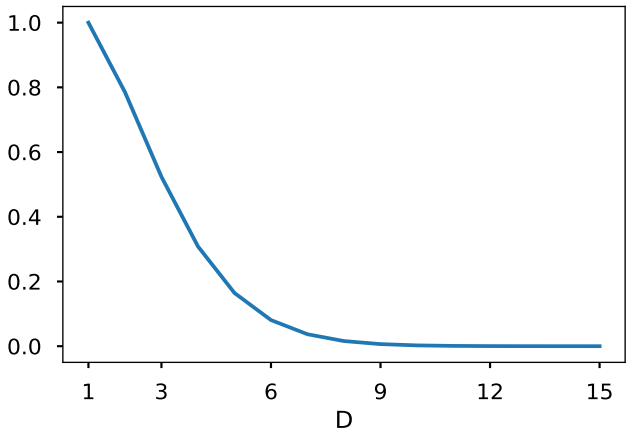
We organize the paper as follows. Section 3 surveys the literature. Section 4 describes the data generating process. Section 5 constructs the sieve and provides conditions under which it approximates the true density well. Section 6 proves our estimators converge at the rates given above. Section 7 provides our estimation strategy. Section 9 analyzes the performance of our estimator in a simulation. Section 8 introduces the data and the prior we use for the empirical analysis. Section 10 uses our method to empirically analyze a monthly macroeconomic panel and

a daily financial panel. Section 11 concludes. The appendices contain all the relevant proofs and additional empirical results.

2 Intuition

The results we mentioned so far likely seem surprising, so we now explain why they are actually reasonable. We do this by discussing the intuition driving the results in the Bayesian compression literature. The number of D-dimensional hypercubes cubes of width $\frac{1}{T}$ required to fill a large hypercube of width 1 equals T^D. This is why the number of terms used by a sieve grows exponentially with D. The compression algorithms in the Bayesian compression literature avoid requiring as many terms 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. Figure 1 shows the volume of a D-dimensional ball relative to a D-dimensional hypercube. As D grows, filling the D-dimensional ball with points becomes easier relative to filling the D-dimensional hypercube. We exploit this behavior to construct 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



The ratio between the volume of a ball of hypercube with the same diameter: $\frac{\pi^{\frac{D}{2}}}{2^D\Gamma\left(\frac{D}{2}+1\right)}$. Γ refers to the Gamma function.

Moving forward, we construct a sieve for the ball and not the hypercube. We do this by developing the first sparse discrete operator to cluster the data. This operator does not significantly perturb the data's first two sample moments. Given a process which is locally asymptotically normal, having the first two moments being close implies the associated densities are close. Since this operator is discrete, it induces a Gaussian mixture representation of a density where the number of components does not grow faster with the data's dimension.

The number of mixture components determines the complexity of a Gaussian mixture. For any

fixed sample-size T, this is a random variable. Consequently, given the data, the convergence rate itself is as a random variable with respect to prior. We consider an asymptotic experiment where T grows and D is fixed. We show with prior probability $1-\delta$, where δ is a small number chosen by the econometrician, the number of components grows logarithmically with the time dimension T. We do this by constructing finite-sample bounds that hold for all T. This differs from the literature because they study problems where the prior (if it even exists) does not affect their bounds. Consequently, they cannot exploit the prior that the smoothness gives to the posterior. At a technical level, for any fixed T our sieve is not a measurable function of the data and so the bounds derived by Stone (1980) and others do not apply.

3 Literature Review

Our paper lies at the intersection of several stands of literature. First, our theoretical results belong to literature on concentration inequalities. As mentioned above, the key idea of projecting a Gaussian process into a lower dimensional subspace dates back to Johnson and Lindenstrauss (1984). We rely heavily on Klartag and Mendelson (2005), Boucheron, Lugosi, and Massart (2013), and Talagrand (2014) to develop the random operator we use. Our main contribution to this literature is that, to the best of our knowledge, we are the first to construct a random operator that both compresses the data in a way that provides bounds that independent of the dimension and induces a mixture representation for the density.

Second, our main estimation results characterize Bayesian posteriors' concentration rates. Over the past approximately twenty years, various authors have worked extensively on these issues starting with the seminal papers Ghosal, Ghosh, and Vaart (2000) and Shen and Wasserman (2001). The most closely related paper is Nguyen (2016) which analyzes the estimation of the latent mixing measure in a hierarchical model. This paper also provides a some general conditions at which you can nonparametrically estimate multivariate densities without a curse of dimensionality. In particular, it shows if the data's generating process (DGP) can be represented as a hierarchical Dirichlet Gaussian mixture model, the posterior converges rapidly. (The particular rate depends upon various assumptions on the DGP in ways that he exposits.)

Our approximating model has a form very similar to those he considers there and in an accompanying paper, Nguyen (2013), and we rely on his work on the geometry of infinite mixture models to characterize the behavior of our model. Unlike him, we provide some general conditions under which we can represent the DGP in a way suitable to this analysis instead of assuming it directly.

Our work on transition density estimation in Bayesian contexts builds upon a number of models in that literature. Perhaps the most similar class of models is the smoothly mixing regressions started in Geweke and Keane (2007) and extended by Norets and various coauthors (Norets 2010; Norets and Pelenis 2012; Norets and Pati 2017). These papers focus on the models with finitely many mixture components and mixing weights that depend upon the conditioning variables. Conversely, our models treat the mixing variables as latent and uses infinitely many of them

asymptotically.

A few other authors have considered Bayesian nonparametric conditional density estimation. Pati, Dunson, and Tokdar (2013) directly models the mixture probabilities like Norets and coauthors do in a Gaussian mixture model. A recent work by Kalli and Griffin (2018) models the stationary and transition densities using Bayesian nonparametric methods with infinite mixtures where they use adaptive truncation to pick the number of mixture components as in Griffin (2016). However, they do not provide theoretical results regarding consistency or convergence rate. In contrast, our method endogenously determines the number of active mixtures and requires fewer tuning parameters.

Our empirical work and the practical questions that initially motivated this project concern flexible modeling of multivariate transition densities. The literature on this topic focus on the practical performance of the estimators under consideration instead of their theoretical properties. The seminal papers in this regard include work on regime-switching models, (Hamilton 1989); time-varying parameter VAR models, (Primiceri 2005); stochastic volatility models (Kim, Shephard, and Chib 1998); and many, many others that we cannot adequately survey due to space constraints. The main advantage of our model is we provide strict guarantees on when our model approximates the diverse nonlinear, non-Gaussian dynamics present in the data. In addition, our model is very parsimonious, and we provide valid credible sets for the parameters of interest.

4 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.[2] We want to estimate x_t's conditional densities for $t = 1, \ldots, T$.

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

$$p_{0,T}\left(x_t \,|\, \mathcal{F}_{t-1}\right) \coloneqq \sum_{k=1}^{\infty} \pi_{k,t-1} \phi\left(x_t \,|\, x_{t-1}\beta_{k,t}, \Sigma_{k,t}\right) \tag{1}$$

Since X_T is Gaussian process, the conditional density $p_{0,T}(x_t|\mathcal{F}_{t-1})$ has an infinite Gaussian mixture representation for each time period. Each mixture component has the associated mixture probability $\pi_{k,t-1}$ and component-specific parameters $(\beta_{k,t}, \Sigma_{k,t})$. We also assume that each x_t has finite mean and finite variance. It is worth noting that even though X_T is a Gaussian process, its conditional densities — $p_{0,T}\left(x_t \,|\, \mathcal{F}_{t-1}\right)$ — are not necessarily Gaussian. 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 $p_{0,T}$ for all T.

2. 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.

Definition 2 (Approximating Model)**.**

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

The approximating model is a Gaussian mixture with K_T components, where 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 mixing variables $(\beta_{k,t}, \Sigma_{k,t})$ across time without loss of generality. If two separate time periods have similar enough dynamics, we group them into one component with the same parameters. Now, as the clusters are defined differently in the true and approximating models, there is no simple relationship between them.

In what follows, we denote the true distribution as $P_{0,T}$ and the approximating distribution as Q_T, with associated densities $p_{0,T}$ and q_T. Throughout, we use μ_T to refer to the $T \times D$-matrix of means. In the following sections, we will also consider the rescaled data:

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

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 analogous densities of $\widetilde{X}_T$ as above — $\widetilde{p}_{0,T}$ and $\widetilde{q}_T$.

5 Sieve Construction

5.1 Setting up the Problem

In this section, we construct a sieve, a sequence of approximating models, 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 is useful because metric entropy controls the rate at which posteriors converge as shown by Ghosal, Ghosh, and Vaart (2000) and Shen and Wasserman (2001). It also controls the minimax rate at which estimators can converge (Wong and Shen 1995; Yang and Barron 1999).

The problem we are tackling is approximating a marginal density which lies 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 problems are not well-posed because there exist multiple equivalent representations for each density given X_T that satisfy some bound on the distance to $p_{0,T}$ in some metric on $\mathscr{P}\left(R^D\right)$. This implies that we can choose a representation that is particularly amenable to estimation for each T. In practice, we want to find the most parsimonious density that still approximates well.

The way we construct our sieve is as follows. Given some $\epsilon > 0$, we construct a mapping Θ_T

that takes this hypersphere and maps it onto a $K \times D$ hypersphere, where $K \ll T$. This mapping only perturbs the norms of the individual elements by at most ϵ.[3]

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

5.2 Bounding the Norm Perturbation

We construct our approximate sufficient statistic for $\widetilde{X}_T$ by borrowing ideas from Bayesian compression theory. Intuitively, we take $\widetilde{X}_T$ and "project" it onto a lower-dimensional space. The reason this intuition is not exact is because the target space is not a subspace of the original space. In particular, we want the compressed data to follow a mixture distribution. This implies that the compression operator Θ_T must be a discretization operator. A mixture distribution for some collection of data $\widetilde{X}_T$ is a 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. This does not satisfy our needs 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 the columns in the row must contain zeros, which makes the columns too dependent for our results to go through.

Fixing the first issue is relatively straightforward, we let $\theta_{t,k}$ take on values from $\{-1, 0, 1\}$ and 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 could just let each row have, potentially, as many 1's and -1's as necessary. By doing this, seeing a 1 in column k gives us no information about columns $k + 1$ through K. It does complicate the analysis slightly, though, for we are effectively letting each time period be in more than one component simultaneously. In other words, we do not just create a mixture distribution across time periods but also create one in each time period.

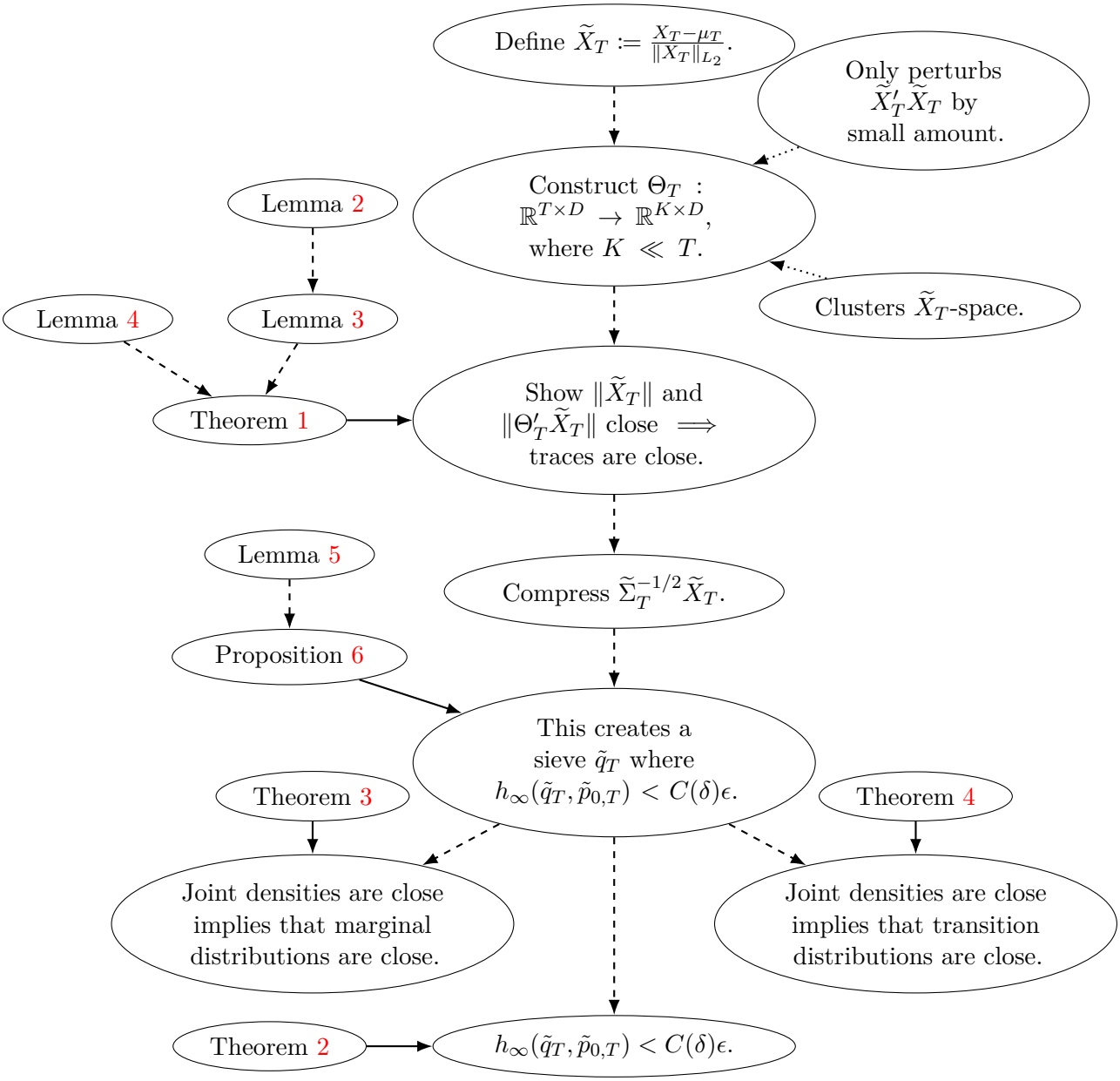
To make the discussion in the previous few paragraphs more formal, we define a random operator Θ_T. We use a stick-breaking process to construct Θ_T, adapting the form often used to construct Dirichlet processes, as proposed by Sethuraman (1994).[4]

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 $\theta := \chi \cdot b$ independent of all of the previous values, and form them into a column-vector — $\theta_{:,1}$. Form another column vector $\theta_{:,2}$ 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.

The reason we form the Θ_T operator in this manner is so that $\mathbb{E}[\theta] = 0$, but $\mathbb{V}\text{ar}(\theta) = \mathbb{E}[|\theta|] = \Pr(b = 1)$. Furthermore, its rows are independent and its columns form a martingale-difference

3. We are constructing an ϵ-isometry.
4. In Section 5.7, we will show that Θ_T can be replaced by a Dirichlet process without affecting our results.

Figure 2: Sieve Construction



This graph displays the dependencies between the various lemmas and theorems. Dashed lines are dependencies, solid lines are labels, and dotted lines are comments.

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 do create such a bound, we must understand its distribution. By Lemma 2, 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 3 that $K_T \propto \log(T)$ with high probability. This feature will be 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 each bin identically. Since $K_T \propto \log(T)$, 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 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 with probability greater than $1 - 2\delta$ with respect to Θ_T, there exists a constant c_2 such that for any $\epsilon > \sqrt{\frac{\log T}{K_T}}$*

$$\sup_t \left| \|\theta_t \widetilde{x}_t\|_{L_2} - \|\widetilde{x}_t\|_{L_2} \right| < c_2 \left(1 + \log \frac{1}{\delta}\right) \epsilon. \tag{5}$$

This means that as long as we choose Θ_T with the number of columns $K \propto \log(T)$, the norms of $\widetilde{x}_t$ are perturbed by at most ϵ by applying Θ_T. This holds with probability at least $1 - 2\delta$ with respect to the distribution over Θ_T. Since $\widetilde{X}_T \in S^{TD-1}$, Theorem 1 shows that we can map S^{TD-1} onto a smaller space $S^{K_T D-1}$, with $K \ll T$, without perturbing the individual elements significantly.

5.3 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 themselves interesting objects. Rather, they are interesting because they form a sufficient statistic for the Gaussian process. To show the densities are close, we must convert these distances between $\|\widetilde{x}_t\|_{L_2}$ into distances on $\mathscr{P}(\widetilde{X}_t)$.

Conditional on Θ_T, $\Theta_T' \widetilde{X}_T$ has some distribution. Since $\widetilde{X}_T$ is a normalized Gaussian process and Θ_T is a matrix, this process is Gaussian conditional on Θ_T. This implies there exists a distribution for $\widetilde{X}_T$ constructed by integrating out Θ_T. This provides an approximating distribution $\widetilde{Q}_T$ for $\widetilde{X}_T$. Since Θ_T is almost surely discrete, this approximating distribution is a mixture as in

Definition 2. We can represent this approximating model as an integral with respect to a latent mixing measure. Since the parameters in each components are means and covariances, this latent mixture measure is a measure over that space. Let G_t^Q be the associated mixing measure over this space of means and covariances for each t. Because Θ_T can have more than one non-zero element, we constructed a mixing distribution in each period, even conditional on Θ_T. Let G^Q be the latent mixing measure over the space of G_t^Q. In other words, for each t, we draw G_t^Q from G^Q. What this means in practice, is that since latent mixing measures are almost surely discrete, the G_t^Q share the same atoms. This regularizes the mixing measures across time, i.e., it creates a lot of "smoothness" in the approximating model. However, since the atoms of G^Q are left arbitrary, it does not restrict the set of DGPs that can be approximated well.

Then $\phi(\cdot \mid \Sigma)$ is a mean-zero multivariate Gaussian density with covariance Σ. Let δ_t^Q be the mixture identity that tells you which cluster Σ_t is in. Then $\widetilde{Q}_T$ can be expressed as

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

Likewise, $P_{0,T}$ can be written as

$$\widetilde{P}_{0,T}(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(\widetilde{x}_t \mid \delta_t^P\right) dG_t^P(\delta_t^P)\, dG^P\left(dG_t^P\right), \tag{7}$$

with its associated latent mixing measures and mixture identities. The approximating cluster identity δ^Q is different from δ^P because δ^Q's clustering is induced by Θ_T, not the underlying true clustering.

Since the densities are paramterized as mixtures in terms of their covariances, we need to convert a clustering in $\widetilde{x}_t$-space into a clustering in Σ_t-space so that we can make a statement about the densities. In general, the norms of $\widetilde{x}_t$ and $\widetilde{x}_{t^*}$ being close is insufficient to 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 $\widetilde{x}$'s are in the same cluster δ_k^Q, the following holds:[5]

$$\sup_{t,t^* \in \delta_k^Q} \left|\widetilde{x}_t \Sigma_t^{-1} \widetilde{x}_t - \widetilde{x}_{t^*}' \Sigma_{t^*}^{-1} \widetilde{x}_{t^*}\right| < \epsilon. \tag{8}$$

Here ϵ is independent of t, t^*, and the cluster identity. We can view the right-hand side of Eq. (8) as a difference on the space of covariance matrices. Accordingly, we introduce the following semimetric on the space of covariance matrices.[6]

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

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

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

$$\delta_{wl_2}(\Sigma_k, \Omega_k) \coloneqq \sup_{t,t^* \in \delta_k^Q} \left| \widetilde{x}_t' \Sigma_k^{-1} \widetilde{x}_t - \widetilde{x}_{t^*}' \Omega_k^{-1} \widetilde{x}_{t^*} \right| \tag{9}$$

It is worth noting that the δ_{wl_2} is compatible with, and weaker than, the max-norm.[7] 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, we can consider the space of covariance matrices as 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 in the space of $\widetilde{x}_t$ into bounds on the space of covariance matrices. We still have to convert this to a bound on the space of densities. The distance we use here is the Hellinger distance.

Definition 5. Hellinger Distance

$$h(p, q) \coloneqq \frac{1}{\sqrt{2}} \sqrt{\int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx} \tag{10}$$

This 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 Hellinger distance. Instead of applying this directly to the joint distribution, we take the supremum over the conditional distributions.

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

$$h_\infty^2(p, q) \coloneqq \sup_{\mathcal{F}_{t-1}^P \mathcal{F}_{t-1}^Q, 1 \le t \le T} h^2 \left(p\left(\cdot \,\middle|\, \mathcal{F}_{t-1}^P \right), q\left(\cdot \,\middle|\, \mathcal{F}_{t-1}^Q \right) \right) \tag{11}$$

The supremum Hellinger distance will be useful later 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 the divergence in terms of h_∞, we can directly deduce the other bounds that we need.

5.4 Representing the Joint Density

Here we show that the approximating distribution of $\widetilde{X}_T$ induced by Θ_T is close to the true distribution $\widetilde{P}_{0,T}$ in h_∞. This is true when the rescaled trace (sample second moment) is a locally sufficient

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

statistic for the density. Hence, we can use bounds on divergences in $\widetilde{\mathcal{X}}$ to bound divergences in the space of probability measures over $\widetilde{\mathcal{X}}$.

Theorem 2 (Representing the Joint Density)**.** *Let $\widetilde{X}_T \coloneqq \widetilde{x}_1, \dots \widetilde{x}_T$ be a D-dimensional Gaussian process with period t 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 constructed in Definition 3. Let $\widetilde{P}_{0,T}$ denote the distribution of $\widetilde{X}_T$. Then given $\epsilon > 0$ and for some $\delta \in (0,1)$, the approximating distribution Q_T, which is the mixture distribution over $\widetilde{\mathcal{X}}$ defined by the clustering induced by Θ_T satisfies the following with probability at least $1 - 2\delta$ with respect to Θ_T*

$$h_\infty \left(\widetilde{P}_{0,T}(\widetilde{\mathcal{X}}), \widetilde{Q}_T(\widetilde{\mathcal{X}}) \right) < C \log \left(\frac{1}{\delta} \right) \epsilon. \tag{12}$$

The way we represent the joint density is 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 is very similar to the various projection arguments that are made in the literature in that we are projecting 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 implies that the density of $\widetilde{x}_t$ can be represented as a process with respect to a discrete measure. That is, Q_T is a mixture distribution. In addition, we show in Section 5.7, 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. However, since $\widetilde{X}_T$ is a Gaussian process and hence locally Gaussian, we can represent $\widetilde{X}_T$ using a Gaussian mixture process whose mixing is driven by the Dirichlet process.

The main issue is that we have stated the bound of 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 5.5) and transition density (Section 5.6). We show that sample mean of the marginal and transition densities converges to those implied by Q_T, and hence those implied by $P_{0,T}$. This is feasible because sample means converge to population means. we cannot construct a sample mean of joint density because we only ever have one realization.

5.5 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 you can bound the maximum deviation of the approximating joint density by $T\epsilon^2$ using Theorem 2. Then standard arguments about convergence of means for product measures give a $\frac{1}{T}$ term. Hence, the deviation between the averages 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 from $p_{0,T}$, where $p_{0,T}$ has a product density. Let Θ_T be constructed as in Theorem 2 for each t. Let ϵ be given. We construct q_T by using the Θ_T operator to group the data, and we assume that the data are 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 in T*

$$h\left(\int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t(\delta_t^Q)\right) < C \log\left(\frac{1}{\delta}\right) \epsilon. \tag{13}$$

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 one. Consequently, using dependent 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.*

5.6 Representing the Transition Density

We show here 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 before, we can exploit the approximate product form here to get a $1/T$ term in the convergence rate and use Theorem 2 to get a $T\epsilon^2$ term. Again, the T terms cancel, and so we bound the distance by ϵ^2.

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 $p_{0,T}$. 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 such that the following holds:*

$$q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}\right) \coloneqq \sum_{k=1}^{K} \phi\left(\beta_k x_{t-1}, \Sigma_k\right) \Pr\left(\delta_t = k \,|\, \delta_{t-1}\right). \tag{14}$$

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

$$h_\infty\left(p_{0,T}\left(x_t \,|\, \mathcal{F}_{t-1}^P\right), q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right) < C\sqrt{\log \frac{1}{\delta}}\epsilon. \tag{15}$$

5.7 Replacing Θ_T with a Dirichlet Process

The previous subsections use Θ_T to construct an approximating representation that is very close to the true model. 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 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 $\Pr(\text{that subset}) = (1-2\delta)$ such that the representation above holds. Since each realization, $\Theta'_T \in \Theta'_T$-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. Consequently, we can view the sequence of clusters at each time t as an integral to a Dirichlet process without loss of generality. Furthermore, taking the union of these Dirichlet processes creates a Dirichlet process over the entire space. This is because the Dirichlet process is a normalized completely random measure (Lin, Grimson, and Fisher 2010). This implies we can view the Dirichlet process in each period as a draw from one overarching Dirichlet process. To put it the notation we used to construct Q_T, we can view D_t^Q as a draw from D^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.

6 Bayesian Nonparametrics and Convergence Rates

6.1 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. In particular, we assume the data $\{x_t\}_{t=1}^T$ are drawn from some distribution $P_{0,T}$ which is paramterized $P_{0,T}(\cdot \mid \xi)$, for $\xi \in \Xi$. This parameter set is equipped with the Borel σ-field $\mathscr{B}$ with associated prior distribution Π. We further assume that there exists a regular conditional distribution of X_T given ξ — $P(X_T \mid \xi)$ on the sample space $(\mathcal{X}, \mathscr{X})$. This implicitly defines a joint distribution over $(\mathcal{X} \times \Xi, \mathscr{X} \times \mathscr{B})$:

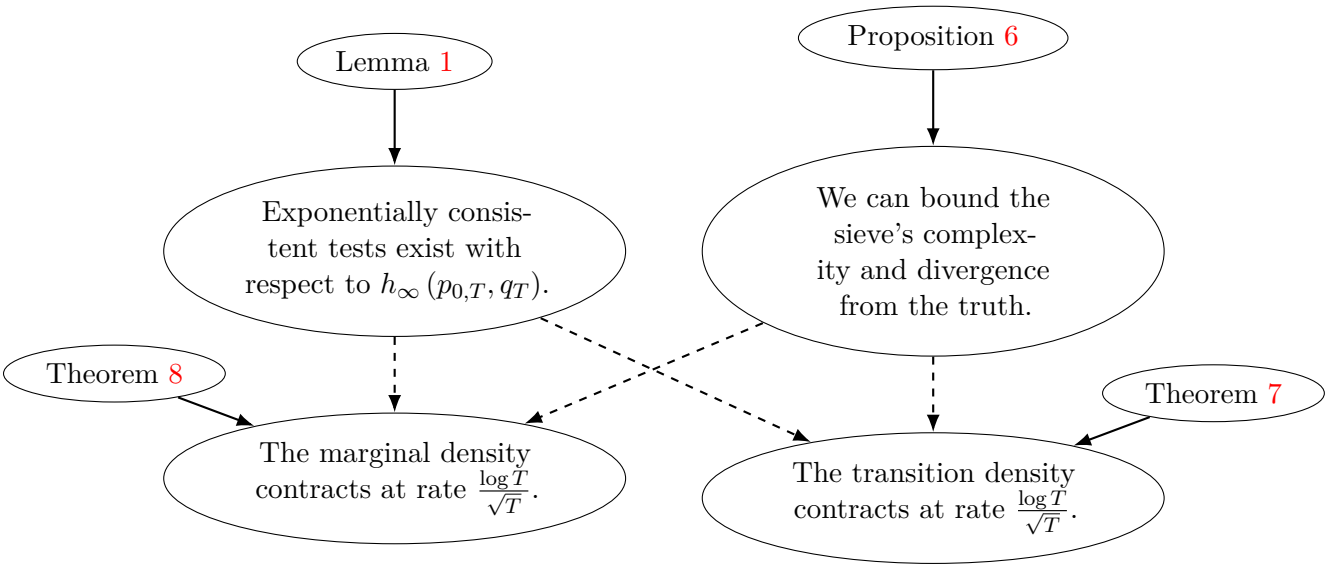
$$P(X_T \in A, \xi \in B) = \int_B P(A \mid \xi)\, d\Pi(\xi). \tag{16}$$

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

$$\Pi(B \mid X_T) \coloneqq P(\{\xi \in B\} \mid X_T),\ B \in \mathscr{B} \tag{17}$$

Figure 3: Contraction Rates



Note: This graph displays the dependencies between the various lemmas and theorems. Dashed lines are dependencies, solid lines are labels.

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_0 with respect to the semimetric d if $\Pi_T(\{\xi \mid d(\xi_0, \xi) \geq M_T \epsilon_T\} \mid X_T) \to 0$ in $P(X_T \mid \xi_0)$-probability for every $M_T \to \infty$. If all experiments share the same probability space the convergence to zero takes place almost surely $P(X_\infty \mid \xi_0)$, then ϵ_T is a *posterior contraction rate in the strong sense.*

To bound the asymptotic behavior of ϵ_T, we must simultaneously bound two separate quantities. First, we must show our model is close to the true data generating process in an appropriate distance. This is what we did 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 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)**.** $N(C\epsilon, d_T(\xi, \xi_0), e_T)$ is the function whose value for $\epsilon > 0$ is the minimum number of balls of radius $C\epsilon$ with respect to 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_0.

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 of related divergence measures. We 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:

$$V_{k,0}(f,g) \coloneqq \int \left|\log(f/g) - \mathrm{D}_{\mathrm{KL}}\left(f \,||\, g\right)\right|^k dF. \tag{18}$$

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

$$B_T(\xi_0,\epsilon,k) \coloneqq \left\{\xi \in \Xi \,\middle|\, \mathrm{D}_{\mathrm{KL}}\left(f\left(X_T \,|\, \xi_0\right) || \, f\left(X_T \,|\, \xi\right)\right) \leq T\epsilon^2, V_{k,0}(f\left(X_T \,|\, \xi_0\right), f\left(X_T \,|\, \xi_0\right)) \leq T\epsilon^2\right\}. \tag{19}$$

Having defined the relevant notations 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 theorems in Ghosal, Ghosh, and Vaart (2000), which is the most common way to derive convergence rates in the literature, to the dynamic case.

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 1 with respect to d_T.*

2. $$\sup_{\epsilon_T > \epsilon} \log N\left(\frac{C_2}{2}\epsilon, \left\{\xi \in \Xi_T \,|\, d_T(\xi,\xi_0) \leq \epsilon\right\}, e_n\right) \leq T\epsilon_T^2 \tag{20}$$

3. $$\frac{\Pi_T\left(\left\{\xi \in \Xi_T \,|\, n\epsilon_T < d_T(\xi,\xi_0) \leq 2n\epsilon_T\right\} \,|\, X\right)}{\Pi_T\left(B_T(\xi_0,\epsilon_T,C_1) \,|\, X\right)} \leq \exp\left(\frac{C_2 T\epsilon_T^2 n^2}{2}\right) \tag{21}$$

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

$$\Pr_T\left(\Pi_T\left(\left\{\xi \in \Xi_T \,|\, d_T(\xi,\xi_0) \geq M_T\epsilon_T\right\} \,|\, X\right) \,|\, \xi_0\right) \to 0 \tag{22}$$

6.2 Contraction Rates

We now show that tests exists with respect to the semimetric that we use: h_∞. It is stronger than the divergences usually used in the Bayesian nonparametric estimation of Markov transition densities: the average squared Hellinger distance (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

$$f_T(X) = f\left(x_T \,|\, \mathcal{F}_{T-1}\right) \cdot f\left(x_{T-1} \,|\, \mathcal{F}_{T-2}\right) \cdots f\left(x_2 \,|\, \mathcal{F}_1\right) \cdot f\left(x_1 \,|\, \mathcal{F}_0\right), \tag{23}$$

where I use $\mathcal{F}_0$ to refer to the 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 the consistent tests to separate two distributions in relevant semimetric exist. To do so, we provide the following lemma.

Lemma 1 (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_0 with $h_\infty(\xi_1, \xi_0)$:*

1. $$\Pr_T \left(\Upsilon_T \,|\, \xi_0\right) \leq \exp(-C_2 T \epsilon^2) \tag{24}$$

2. $$\sup_{\xi \in \Xi, e_n(\xi_1, \xi) < \epsilon C_3} \Pr_T \left(1 - \Upsilon_T \,|\, \xi_0\right) \leq \exp(-C_2 T \epsilon^2) \tag{25}$$

Then the following two conditions hold with probability $1 - 2\delta_T$ with respect to the prior:

$$\sup_{\epsilon_T > \epsilon} \log N\left((\epsilon, \left\{\xi \in \Xi_T \,|\, h_\infty(\xi, \xi_0) \leq \epsilon\right\}, h_\infty\right) \leq T\epsilon_T^2 \tag{26}$$

and

$$\Pi_T \left(B_T(\xi_0, \epsilon_T, C_1) \,|\, X\right) \geq C \exp\left(-C_0 T \epsilon_T^2\right). \tag{27}$$

Having done that we show that Eq. (20) and Eq. (21) 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 will.

Proposition 6 (Bounding the Posterior Divergence)**.** *Let p_0 be a uniformly ergodic Hidden Markov Gaussian process, i.e. $p_0 \coloneqq \sum_k p_k \phi\left(\cdot \,|\, \mu_t, \Sigma_t\right)$ with finite mean and finite variance. Let $\Xi_T \subset \Xi$ and $T \to \infty$. Let the following condition hold with probability $1 - 2\delta$ for $\delta > 0$ and constants C and $n \in \mathbb{N}$*

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_{0,T}\left(x_t \,|\, \mathcal{F}_{t-1}^P\right)\right) < C\eta_T. \tag{28}$$

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

$$\sup_{\epsilon_{T,n} > \epsilon_n} \log N\left((\epsilon_n, \left\{\xi \in \Xi_T \,|\, h_\infty(\xi, \xi_0) \leq \epsilon_n\right\}, h_\infty\right) \leq T\epsilon_{T,n}^2, \tag{29}$$

and

$$\Pi_T \left(B_T\left(\xi_0, \epsilon_{T,n}, 2\right) \,|\, X\right) \geq C \exp\left(-C_0 T \epsilon_{T,n}^2\right). \tag{30}$$

As a consequence, by Theorem 5, we have the following result.

Theorem 7 (Contraction Rate of the Transition Density)**.** *Let p_0 be a uniformly ergodic Hidden Markov Gaussian process, i.e. $p_0 \coloneqq \sum_k \pi_{t,k} \phi\left(\cdot \,|\, \mu_t, \Sigma_t\right)$ with finite mean and finite variance. Let $T \to \infty$, then the following holds with $\epsilon_T = \sqrt{\frac{\log(T)^2}{T}}$ 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

$$P_0\left(\Pi_T\left(\sup_{\mathcal{F}_{t-1}^P,\mathcal{F}_{t-1}^Q} h\left(p_{0,T}\left(x_t \,|\, \mathcal{F}_{t-1}^P\right), q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right) \geq C\epsilon_T \,\middle|\, X_T\right)\right) \to 0. \tag{31}$$

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 uniformly ergodic Markov processes, we cover the i.i.d. case as well.

Theorem 8 (Contraction Rate of the Marginal Density)**.** *Let p_0 be a uniformly ergodic Hidden Markov Gaussian process, i.e. $p_0 \coloneqq \sum_k p_k \phi\left(\cdot \,|\, \mu_t, \Sigma_t\right)$ with finite mean and finite variance. Let $T \to \infty$, then the following holds with $\epsilon_T = \sqrt{\frac{\log(T)}{T}}$ 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*

$$P_0\left(\Pi_T\left(h\left(p_{0,T}\left(x_t\right), q_T\left(x_t\right)\right) \geq C\epsilon_T \,|\, X\right)\right) \to 0. \tag{32}$$

7 Estimation Strategy

As discussed all along, we use Bayesian methods to estimate our model. So far, the discussion has been rather abstract, and we have focused on providing theoretical results concerning our general estimation strategy. We construct a Gibbs sampler to estimate the model in this section.

Recall the definition of the approximating model:

$$q_T\left(x_t \,|\, \mathcal{F}_{t-1}\right) \coloneqq \sum_{k=1}^{K_T} \pi\left(k = \delta_t \,|\, \delta_{t-1}\right) \phi\left(\beta_k x_{t-1}, \Sigma_k\right). \tag{33}$$

We need to place a prior on each of the components in this model. We start by using a Dirichlet process to construct a prior on $\pi_{k,t-1} = \pi\left(k = \delta_t \,|\, \delta_{t-1}\right)$, and hence implicitly on K_T. We then construct priors for β_k and Σ_k. Equation (33) is almost a standard Gaussian mixture model.[8] Given $\delta_t = k$, we apply the standard Bayesian regression to obtain β_k and Σ_k with normal-inverse-Wishart prior. However, the predictive distributions approximated by the mixtures are time-varying. Hence, we must construct a prior for the transition matrix — Π — of the cluster identities — δ_t.

8. Conditional on δ_{t-1} it is.

7.1 Posterior for the Cluster Identities

In each period, the density is a Dirichlet mixture model, as is the marginal density. Consequently, we can draw the cluster identities by adapting algorithms from the literature. Essentially, we are in the standard situation, except our prior varies from period to period.

There are two difficult issues with sampling Dirichlet mixtures. First, because the prior allows for infinitely many clusters, we cannot sum the probabilities, and hence cannot compute the marginal cluster probabilities. In other words, we cannot compute the probability of cluster k, p_k, by using $1 - \sum_{\kappa \neq k} p_\kappa$. This is true in any Dirichlet mixture model, and several authors have developed ingenious ways of dealing 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 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

$$f(x_t) = \sum_{k=0}^{\infty} \pi_{t,k} \phi\left(x_t \,|\, \beta_k, \Sigma_k\right). \tag{34}$$

As predicted, we introduce a latent variable $u_t \sim U(0, \pi_{t,k})$ so we can rewrite Eq. (34) as

$$f(x_t) = \sum_{k=0}^{\infty} \mathbf{1}\left(u_t < \pi_{t,k}\right) \phi\left(x_t \,|\, \beta_k, \Sigma_k\right) = \sum_{k=0}^{\infty} \pi_{t,k} U\left(u_t \,|\, 0, \pi_{t,k}\right) \phi\left(x_t \,|\, \beta_k, \Sigma_k\right). \tag{35}$$

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

$$f\left(u_t \,|\, \pi_t\right) = \sum_{k=0}^{\infty} \pi_{t,k} U\left(u_t \,|\, 0, \pi_{t,k}\right) = \sum_{k=0}^{\infty} \mathbf{1}\left(u_t < \pi_{t,k}\right). \tag{36}$$

Hence, conditional on $\pi_{t,k=\delta_t}$, we can draw u_t from $U(0, \pi_{t,k=\delta_t})$. Then we can condition on $\{u\}_{t=1}^T$ as a vector, but not on $\pi_{k,t}$:

$$f\left(\{v_k\}_{k=0}^{\infty} \,|\, \{\delta_t\}_{t=1}^T\right) = \text{prior}\left(\{v_k\}_{k=0}^{\infty}\right) \prod_{t=1}^{T} \mathbf{1}\left(v_{k=\delta_t} \prod_{\kappa<\delta_t} (1 - v_\kappa) > u_{k=\delta_t}\right), \tag{37}$$

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

The dependence between the u_t does not affect Eq. (37) because the v_k do not depend upon t. Hence, the v_t are conditionally independent given $\{u_t\}_{t=1}^T$. Exploiting this independence and the stick-breaking representation of the prior, we can draw v_t from the above, since 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 Eq. (37) to draw v_k. As shown by Papaspiliopoulos and Roberts (2008), this

implies the sticks for $k = 0, 1, \ldots$ are distributed:

$$v_k \sim Beta\left(\sum_{t=1}^{T} \mathbf{1}(\delta_t = k) + 1, T - \sum_{\kappa=1}^{k}\sum_{t=1}^{T} \mathbf{1}(\delta_t = \kappa) + \alpha\right). \tag{38}$$

We only need to do this for v_k, such that $k \leq \max(\delta_t)$. They are the only v_k that appear in the likelihood. We can calculate the marginal cluster probabilities π_k

$$\pi_k = v_k \prod_{\kappa=1}^{k}(1 - v_\kappa). \tag{39}$$

The tricky part is sampling the indicator variables. 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. However, the data are not i.i.d., i.e., the π depend on δ_{t-1}. This is what is done in the references above. The question at hand is how do you change the underlying marginal distribution (which is what we have computed) to the conditional distribution while conditioning on the dependence structure embedded in the transition matrix.

What we must do is to construct a probability matrix such that relationship between two clusters, say κ_1 and κ_2, remain the same as they did in the old case, but the marginal distributions are correct. Consider two transition matrices Π and $\widetilde{\Pi}$ and associated marginal distributions π and $\widetilde{\pi}$. We know that Markov transition matrices and marginal distributions have the following relationship for all k:[9]

$$\pi_k = \sum_{\kappa=0}^{\infty} \Pi_{\kappa,k}\pi_\kappa. \tag{40}$$

Define $\widetilde{\Pi}^*$ so that $\widetilde{\Pi}^*_{\kappa,k} = \Pi_{\kappa,k}\frac{\tilde{\pi}_k}{\pi_k}\frac{\pi_\kappa}{\tilde{\pi}_\kappa}$.

$$\tilde{\pi}_k = \frac{\tilde{\pi}_k}{\pi_k}\pi_k = \frac{\tilde{\pi}_k}{\pi_k}\sum_{\kappa=0}^{\infty}\Pi_{\kappa,k}\pi_\kappa = \sum_{\kappa=0}^{\infty}\frac{\tilde{\pi}_k}{\pi_k}\Pi_{\kappa,k}\pi_\kappa = \sum_{\kappa=0}^{\infty}\frac{\tilde{\pi}_k}{\pi_k}\Pi_{\kappa,k}\frac{\pi_\kappa}{\tilde{\pi}_\kappa}\tilde{\pi}_\kappa = \sum_{\kappa=0}^{\infty}\widetilde{\Pi}^*_{\kappa,k}\tilde{\pi}_\kappa. \tag{41}$$

In other words, Π^* has the same marginals as $\widetilde{\Pi}$. In addition, since we only multiplied and divided through by elements of the marginal distribution, we did not alter the dependence structure embedded in Π. Consequently, $\widetilde{\Pi}^* = \widetilde{\Pi}$. More rigorously, we condition on all but the first left eigenvector of the transition matrix, $\widetilde{\Pi}$, and replace that left eigenvector with the one associated with the new stationary distribution. Then we calculate the resulting transition matrix. Since the transition matrices associated with irreducible Markov chains have exactly one stationary distribution and that stationary distribution is the first left eigenvector (the one associated with the eigenvector 1), this new transition matrix is the $\widetilde{\Pi}$ we derived in Eq. (41). If our new stationary distribution $\widetilde{\pi}$ has more clusters than the old one π did, we use the prior for Π to compute them. We do not have to perturb them any because we have no datapoints in them and so they are the same between Π and $\tilde{\Pi}$ as they have the same prior. Likelihoods are irrelevant in sections of the space without any

9. This is the standard condition that a stationary distribution is a left-eigenvector of the transition matrix.

observations.

From $\widetilde{\Pi}$ can compute $\pi_{t,k}$ for each t by using the stationary distribution for $\pi_{t,0}$ and 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

$$\Pr\left(\delta_t = k \,|\, \cdots\right) \propto \mathbf{1}\left(k \in \{k : \pi_{t,k} > u_t\}\right) \phi\left(x_t \,|\, \beta_k x_{t-1}, \Sigma_k\right). \tag{42}$$

Since this is a finite set with known probabilities, we can easily sample from it. The δ_t are categorical variables.

7.2 Posterior for the Coefficient Parameters

We use a standard normal inverse-Wishart prior for the component coefficients:

$$\beta_k \sim \Phi(\beta_\beta, \Sigma_\beta). \tag{43}$$

Since this is a conjugate prior, we can directly use the standard formulas to estimate it.

As mentioned above, we adopt an inverse-Wishart prior for the covariance matrices. We can write the inverse-Wishart prior in the following form for the Inverse-Wishart density with inverse scale matrix Ψ, and prior degrees of freedom ν:

$$\Sigma_k \,|\, \Psi, \nu \sim \text{Inverse-Wishart}\left(\Sigma_k \,|\, \Psi, \nu\right). \tag{44}$$

We have several covariance matrices to estimate, one for each k. Standard Bayesian intuition implies that our estimators will be more efficient if we specify a hierarchical model for them. Then we can estimate the hyperparameters ν and Ψ. This will be particularly useful in our case because we have some clusters without many datapoints in them. Consequently, we need to shrink them a great deal. By estimating the hyperparameters, we shrink them to precisely the right place.

We adapt the hierarchical prior Huang and Wand (2013) construct. We deviate from them to allow the hyper-covariance matrix scales to have off-diagonal elements. In other words, our covariance matrices are i.i.d. before we see any of the data, but the prior for a new covariance matrix is not necessarily i.i.d. In addition, Huang and Wand's (2013) model does not necessarily have a density with respect to Lebesgue measure for the covariance matrix itself.

We parameterize our model as follows. We have two degree of freedom parameters — μ_1 and μ_2. We use $\Omega \coloneqq \mathbb{E}[\Sigma_k]$ to parameterize the scale.[10]

10. This is implied by the Definition 10 by the formula for the mean of an Inverse-Wishart random variable. $\mathbb{E}[\Sigma_k] = \frac{\text{Scale}}{\text{Degrees of Freedom} - D - 1} = \frac{(\mu_1 - 2)\Omega}{\mu_1 + D - 1 - D - 1} = \Omega$.

Definition 10 (Prior for the Covariances)**.**

$$\Sigma_1, \Sigma_2, \dots \Sigma_K \,|\, \Omega \overset{i.i.d.}{\sim} \text{Inverse-Wishart}\left(\mu_1 + D - 1, (\mu_1 - 2)\Omega\right) \tag{45}$$

$$\Omega \overset{i.i.d.}{\sim} \text{Wishart}\left(\mu_2 + D - 1, \frac{\text{diag}(A_1, \dots, A_D)}{\mu_2 + D - 1}\right) \tag{46}$$

If we send $\mu_2 \to \infty$ the implied prior for the prior for Ω becomes fully dogmatic. Setting $\nu_2 = 1/2$ implies the $\sqrt{\Sigma_k}$ have half-t distributions if $D = 1$. In general, the $(\Sigma_k)_{dd}$ have appropriately scaled F-distributions.[11] 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 Eq. (45) 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. So the two questions of interest are as follows. First, what is the posterior distribution of Ω? Second, what is the posterior distribution of Σ_k given $\Omega, \{x_t \,|\, \delta_t = k\}$?

The answer the second question is entirely standard. You draw Σ_k from its Inverse-Wishart posterior in the standard fashion. The answer to the first question is slightly non-standard but still not particularly difficult. We derive the posterior below. Let $\mathcal{V}$ be the prior scale, i.e. $\mathcal{V} \coloneqq \text{diag}\left(A_1, \dots A_P\right)$, then

$$\begin{aligned} &p\left(\Omega \,|\, \Sigma_1, \dots \Sigma_K, A_1, \dots, A_D\right) \\ &\propto \prod_{k=1}^{K} |\Omega|^{\frac{\mu_1 + D - 1}{2}} \exp\left(-\frac{\mu_1 - 2}{2} \text{tr}\left(\Omega \Sigma_k^{-1}\right)\right) \cdot |\Omega|^{\frac{\mu_2 - 2}{2}} \exp\left(-\frac{1}{2} \text{tr}\left(\mathcal{V}^{-1}\Omega\right)\right) \end{aligned} \tag{47}$$

Since matrix multiplication distributes over matrix addition.

$$= |\Omega|^{\frac{K(\mu_1 + D - 1)}{2}} \exp\left(-\frac{\mu_1 - 2}{2} \sum_{k=1}^{K} \text{tr}\left(\Omega \Sigma_k^{-1}\right)\right) \cdot |\Omega|^{\frac{\mu_2 - 2}{2}} \exp\left(-\frac{1}{2} \text{tr}\left(\mathcal{V}^{-1}\Omega\right)\right) \tag{48}$$

$$= |\Omega|^{\frac{K(\mu_1 + D - 1) + \mu_2 - 2}{2}} \exp\left(-\frac{1}{2} \text{tr}\left(\left(\mathcal{V}^{-1} + (\mu_1 - 2) \sum_{k=1}^{K} \Sigma_k^{-1}\right) \Omega\right)\right) \tag{49}$$

This is the kernel of a Wishart distribution. That is

$$\Omega \,|\, \Sigma_1 \dots \Sigma_K, A_1, \dots A_D \sim \text{Wishart}\left(K(\mu_1 + D - 1) + (\mu_2 + D - 1), \left(\mathcal{V}^{-1} + (\mu_1 - 2) \sum_{k=1}^{K} \Sigma_k^{-1}\right)^{-1}\right). \tag{50}$$

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

11. $\sigma^2 \sim F(1, \nu) \implies \sigma \sim Half\text{-}t(\nu)$. In the one dimensional case, $\nu_2 = 1/2$ implies that $\sigma^2 \sim F(1, \nu)$. 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.

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.

If in the course of the algorithm, we need to add a new component, we draw the parameters from the prior. If we have a component with data already in it, then we draw from the posterior.

7.3 Posterior on the Transition Matrix

We place 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 a Π as a Dirichlet-distributed infinite-dimensional square matrix as noted by Lin, Grimson, and Fisher (2010).

Given the cluster identities δ_t which we drew in Section 7.1, we draw the transition matrices. We do this by noting that the prior probability of a transition is the product of the unconditional probability appropriately normalized. We can update this by counting the proportion of realized transitions:

$$(\Pi_T)_{kj} = \frac{\Pr(\delta_{t-1} = k)\Pr(\delta_t = j) + \#(\text{transitions } k \to j)}{\Pr(\delta_{t-1} = k) + \sum_j \#(\text{transitions } k \to j)}. \tag{51}$$

Each element, $(\Pi_T)_{kj}$, reflects that the belief over (δ_{t-1}, δ_t) and is updated by counting the number of transitions from k to j. Lastly, once we condition on a cluster k, we can simply run a Bayesian regression to estimate (β_k, Σ_k) in a standard manner. We iterate our posterior Gibbs sampler to draw from the joint posterior.

7.4 Identification Strategy and Cluster Labeling Problem

As mentioned in the introduction to this section, the other problem endemic to mixture models is that the cluster identities are not 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 is particularly problematic in i.i.d. environments because there is no natural way to order the clusters.

In a time series environment, 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 cluster one. This has two nice features relative to the existing methods of ordering the clusters such as by their probabilities. First, it imposes a strict order of the clusters. We have no ties, such as occur in a weighting by probability when the two probabilities are equal. Second, the ordering is invariant to estimation uncertainty. I do not have to estimate which datapoint comes first in the time series, and so it is easy to maintain the same ordering over time.

Pursuant to this identification restriction, we re-order the cluster identities immediately before returning a posterior draw so that they always arrive in time order. This does not solve the problem

that estimating these cluster identity for each period is difficult, but it does reduce the amount of multi-modality in our posterior.

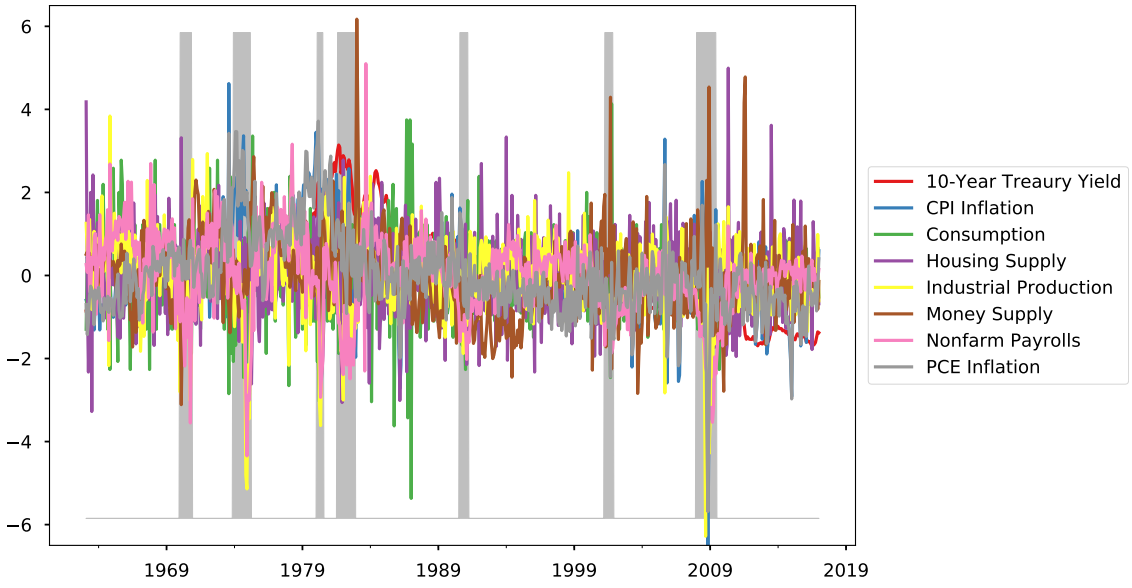
8 Data and Prior

We downloaded monthly data on real consumption (DPCERAM1M225NBEA), the consumer price index (CPIAUCSL), the personal consumption expenditures price index (PCEPI), industrial production (INDPRO), housing supply (MSACSR), the M2 measure of money supply (M2), total nonfarm payrolls (PAYEMS), and 10-year Government bond yields 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 were seasonally-adjusted by FRED. We converted to percent changes by log-differencing all of the data except for the consumption measure, which was already measured in percent changes and the long-term interest rate. We then demeaned the data and rescaled 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 January 2017. The time dimension is 649, and the cross-sectional dimension is 8. Figure 4 shows the standardized monthly macroeconomic data used in this subsection. The gray bars are the NBER recessions.

Figure 4: Monthly Macroeconomic Series

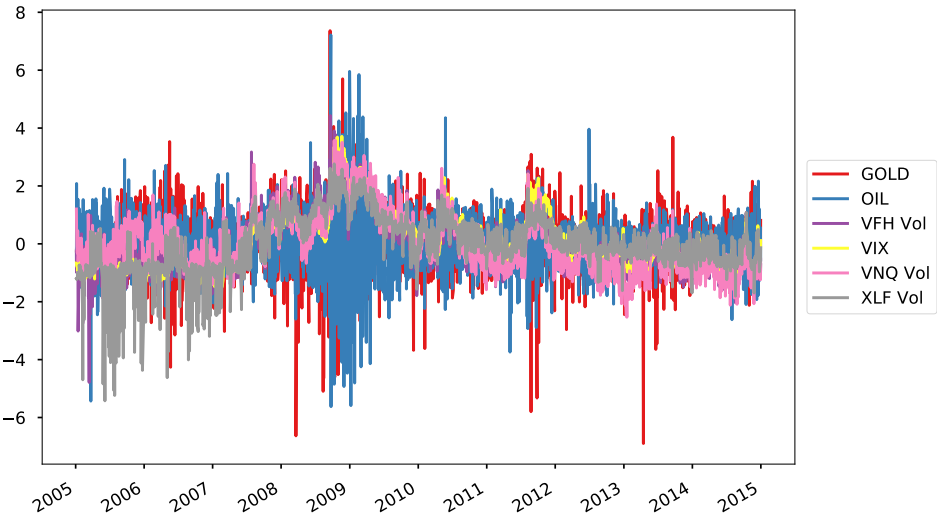


The next dataset we use is financial. There are two types of data used here. The first are volatility measures. We downloaded high-frequency data from Wharton Research Data Services (WRDS)'s Trade and Quote (TAQ) database. The three series we used were the Financial Select

Sector SPDR ETF (XLF), Vanguard Financials ETF (VFH), and the Vanguard Real Estate ETF (VNQ). We then computed the 5 min realized volatility, we then took the logarithm of this volatility. The second dataset we downloaded from FRED. The first is the price of West Texas Intermediate (DCOILWTICO). The second is the price of Gold Bullion, (GOLDAMGBD228NLBM), and the third is the CBOE Volatilty Index / VIX (VIXCLS), of which we took the logarithm. Again, we standardize the data so that it is mean zero and all of the series have standard deviation equal to one.

The time periods covered range from January 3rd, 2005 to December 31st, 2014, a total of 2350 periods. This empirical application is useful because these series are also of great interest to practitioners and policymakers, and they are known to have highly non-Gaussian dynamics. The literature studying stochastic volatility in financial data is one of the largest literatures in economics. They also exhibit time-varying fat tails and often have complex tail dependencies (Patton 2012).

Figure 5: Daily Financial Series



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.

9 Simulation Results

Having characterized our estimators' theoretical properties, we now consider their behavior in practice. We analyze the performance in simulations to better better understand how the approximation works when we know what the true DGP is. The data generating process (DGP) we consider is

Table 1: Prior

Expected Number of Components	20
Component Coefficients	
Intercept	0
Expected Diagonal Autocorrelation	1
Expected Off-Diagonal Autocorrelation	0
Component Covariances	
Mean	$.25^2\mathbb{I}_D$
μ_1	3
μ_2	3

the smooth transition autoregressive model (STAR). These models introduce nonlinearity into the linear autoregressive DGP by letting the parameters depend upon the data x_{t-1}, (Teräsvirta 1994; van Dijk, Teräsvirta, and Franses 2002):[12]

$$x_t = (1 - \text{Logistic}(x_{t-1}))\,\zeta_1 x_{t-1} + \text{Logistic}(x_{t-1})\zeta_2 x_{t-1} + \epsilon_t, \quad \epsilon_t \sim \text{Student's } t\left(0, \sigma_\epsilon^2, df = 5\right). \tag{52}$$

This is useful because it satisfies our nonparametric restrictions, (The density is smooth and has finite mean and variance), but is not a finite Gaussian mixture. It is not a special case of our sieve.

We parameterize the logistic transition function as

$$\text{Logistic}(x_{t-1}) \coloneqq \left[1 + \exp\left(-\left(x_{t-1} - \mu\right)\right)\right]^{-1}. \tag{53}$$

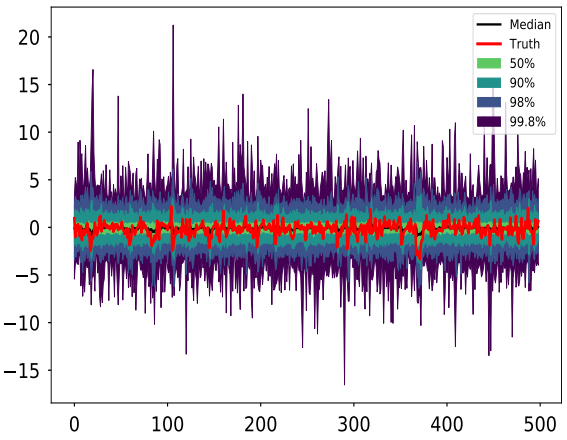
We parameterize the DGP by setting $\zeta_1 = 0.8$, $\zeta_2 = -0.3$ and $\mu = 0$. We use the Student's t-distributed innovations to give the innovations distribution fat tails. As we can see in Fig. 6 our method based on the Gaussian mixtures approximates the dynamics of the STAR model quite closely.

Figure 6a shows the in-sample predictive posterior density of x_t given x_{t-1}. The colored intervals shows the credible set drawn with posterior draws for with the labeled percentages. The red line shows the true x_t. The black solid line is the posterior median. We can see that the posterior transition density closely captures the true dynamics of x_t. Figure 6b shows the cluster identities over time. The Fig. 6c shows the probability integral transition (PIT) histogram. The PIT is the cumulative density of the random variable x_{T+1} evaluated at the true realization. Figure 6d shows the PIT autocorrelation function (ACF). If the predictive distribution is correctly conditionally calibrated, the PIT histogram should be distributed as Uniform[0,1] and ACF should not show any serial dependence. The gray area is credible set drawn using Barlett's formula. From Fig. 6c and Fig. 6d, we see that this is the case.

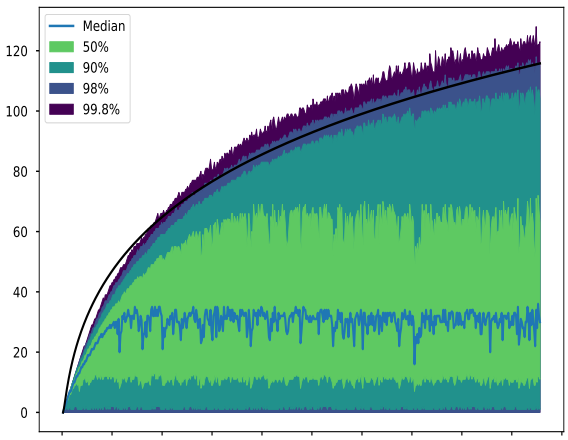
12. We also conducted simulation experiments with other specifications including univariate structural break model and vector autoregressive model with regime-switching correlation. These results are available upon request.

Figure 6: STAR Simulation Results

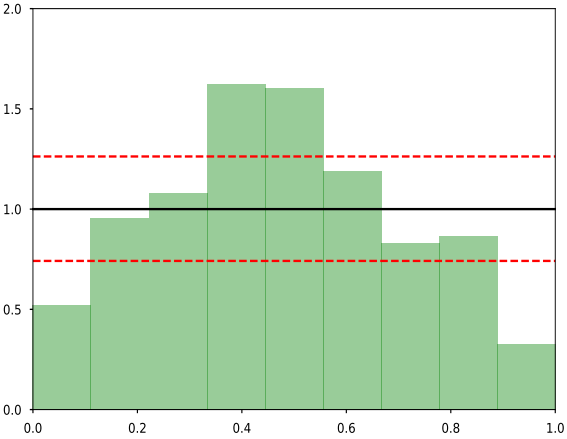
(a) Posterior Density of $x_t \mid x_{t-1}$



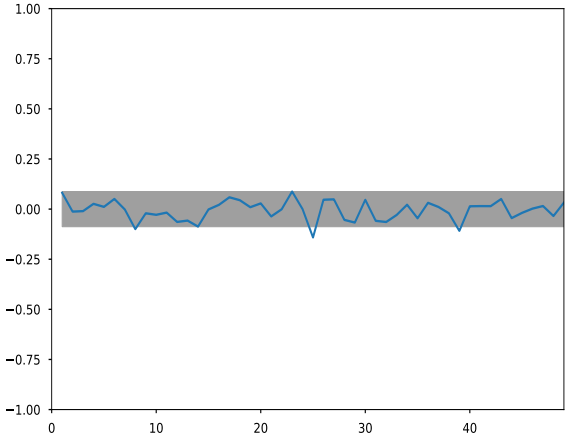
(b) Cluster Identities



(c) The PIT Histogram



(d) The PIT's Autocorrelation Function (ACF)



In addition, we can see from Fig. 6b that we use more clusters as time progresses. This is almost by construction. The bands in that graph are confidence intervals with coverage given in the left. In addition, since the Student' t-distribution has fatter tails than the normal we use at least three clusters in all of the periods.

However, not only do they increase, they increase at the predicted rate. The black line in Fig. 6b is $C \log(T)^2$, where we pick C to match the number of components used. The rate of increase is very close to $\log(T)^2$, exactly as theory predicts.

10 Empirical Results

10.1 Monthly Macroeconomic Series

Using the macroeconomic data, we obtain the posterior draws from our sampler which are summarized in Fig. 7. In Fig. 7a, 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:

$$\mathbb{V}\text{ar}\left(x_t \mid \mathcal{F}_{t-1}\right) = \mathbb{V}\text{ar}\left(\mathbb{E}\left[x_t \mid \delta_t\right] \mid \mathcal{F}_{t-1}\right) + \mathbb{E}\left[\mathbb{V}\text{ar}\left(x_t \mid \delta_t\right) \mid \mathcal{F}_{t-1}\right]. \tag{54}$$

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 7d shows the volatility associated with autoregressive coefficients, whereas Fig. 7b shows the volatility associated with innovations. The total volatility, which we graph for consumption in Fig. 9a, is the sum of the two. Interestingly, most of the variation arises from the variation in the conditional means, not the variation in the conditional variances.

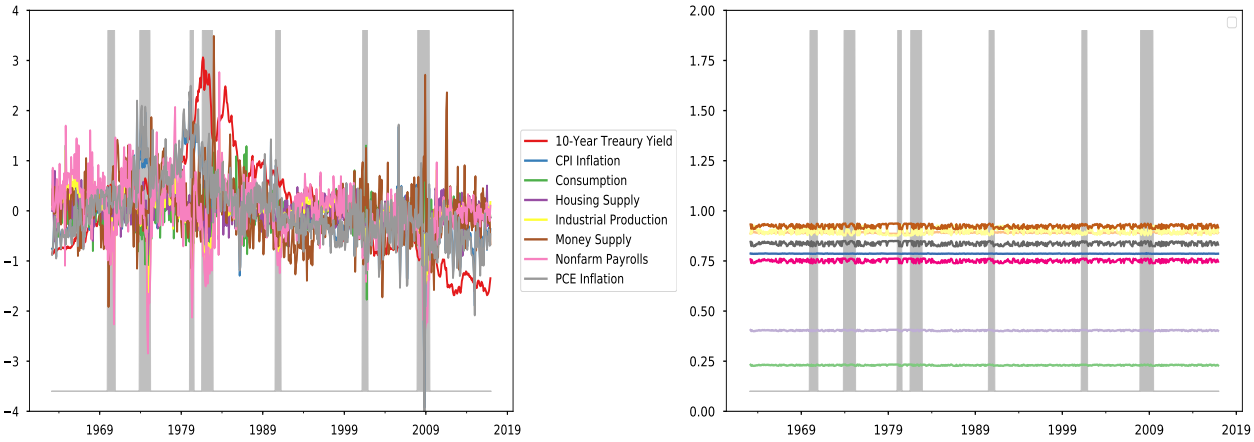
Comparing these two volatilities, we observe bigger changes in dynamics for the coefficient volatility. Figure 7c shows the number of active clusters in each period. This implies that the stochastic volatility in macroeconomic data studied in papers such as Fernández-Villaverde and Rubio-Ramírez (2010) and Fernández-Villaverde et al. (2015) can be more parsimoniously modeled using variation in the conditional mean than by using stochastic volatility.

Examining the time-variation in the cluster identities in Fig. 7c, we can see that we only use 2 clusters in the vast majority of cases. Hence, our model is very parsimonious. We did not impose this. The prior mean for the number of clusters (20) was higher than this, and in several of the other examples, the algorithm used more. We can also see that the cluster identity fluctuates at a very high frequency. Unlike many regime-switching models, we do not have a "recession" regime and a "normal-times" regimes.

Figure 7: Empirical Results with Monthly Macroeconomic Series

(a) Mean

(b) Innovation Volatility



To show that our algorithm works reasonably well in practice, we display the conditional density forecast for consumption in Fig. 8.[13] 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. The main caveat is this is an in-sample fit.

The dynamics of the data in Fig. 8a are not obviously non-Gaussian or non-linear. Are we

13. Predictive Densities and PIT's for the other series are provided in Section D.

(c) Cluster Identity

(d) Coefficient Volatility

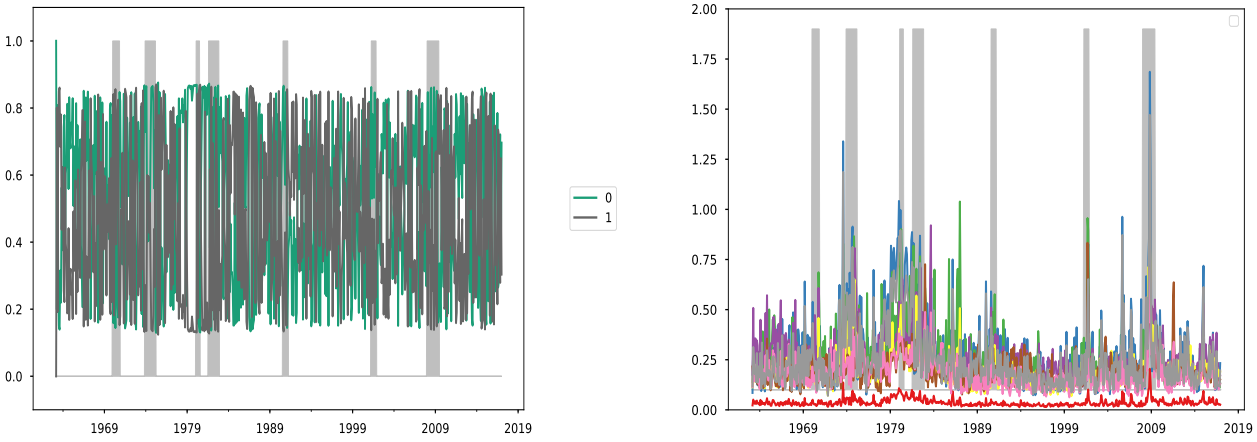
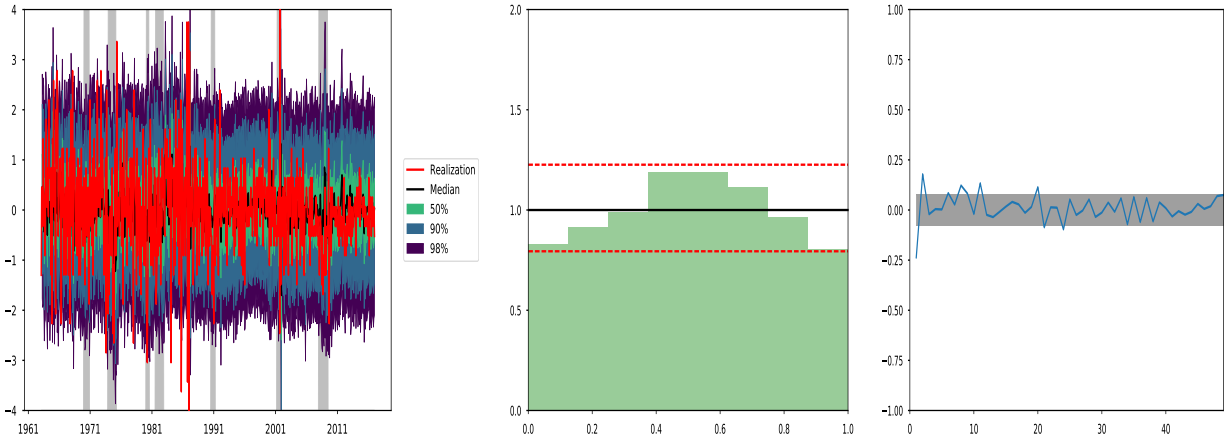


Figure 8: 1-Period Ahead Conditional Forecasts: Consumption Expenditure

(a) Posterior Density

(b) PIT Histogram

(c) PIT ACF



effectively just estimating a VAR? No, if we examine the conditional variance, (Fig. 9a), we see that it spikes a great deal in recessions. In other words, we can see stochastic volatility in consumption data that varies with the business cycle. Interestingly, neither skewness, (Fig. 9b), nor kurtosis, (Fig. 9c), varies dramatically. The data actually become more positively skewed in recessions.

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, 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.

Our model estimates a large number of parameters, and so we cannot show you all of them. In each of the K_T components we have a $D \times (D+1)$ coefficient matrix and $D \times D$ covariance matrix. We also have a $K_T \times K_T$ transition matrix. One parameter that is particularly interesting is the transition matrix, and so we report the mean transition matrix draw, and associated stationary distribution. In general, this is an infinite matrix, but this estimation only required two components.

Figure 9: Consumption Variability

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

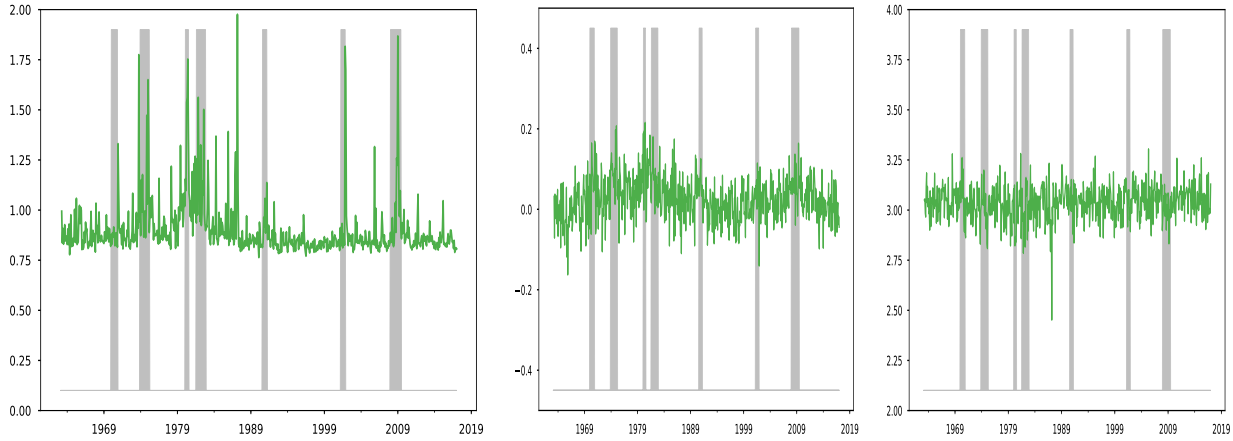


Table 2: Macroeconomic Transition Matrix

	0	1	Marginal
0	0.60	0.40	0.53
1	0.47	0.53	0.48

As can clearly be seen in Table 2, both components are very common, taking up around one-half of the periods each. In addition, they are not very persistent. The diagonal of this matrix governs the probability of remaining in the same component. It is larger in magnitude than the off-diagonal, but not by a large amount. This is also apparent from Fig. 7c, where we see the more likely a posteriori cluster switching between the two very frequently.

10.2 Daily Financial Series

Using the various financial series, we obtain the posterior draws from our sampler which are summarized in Fig. 10. In Fig. 10a, we can see the conditional mean dynamics. Again, the mean tracks the data relatively well, and we see a sharp increase in the volatility during crises.

Similar to the macroeconomic dataset, we see that almost all of the increase in volatility is coming from volatility in the conditional means. The Fig. 10d shows the volatility associated with autoregressive coefficients, whereas the Fig. 10b show the volatility associated with innovations.

Figure 10c shows the number of active clusters in each period. The number of clusters is much smaller than the length of the time series, as our theory predicts. It is much larger than the macroeconomic case though. We see some time-variation in the number, but not a great deal. Most of the time is spent in the clusters 0 – 2. Also, recall our identification scheme labels the clusters by when they first occur. This is why the cluster probability for cluster 0 starts at 1.

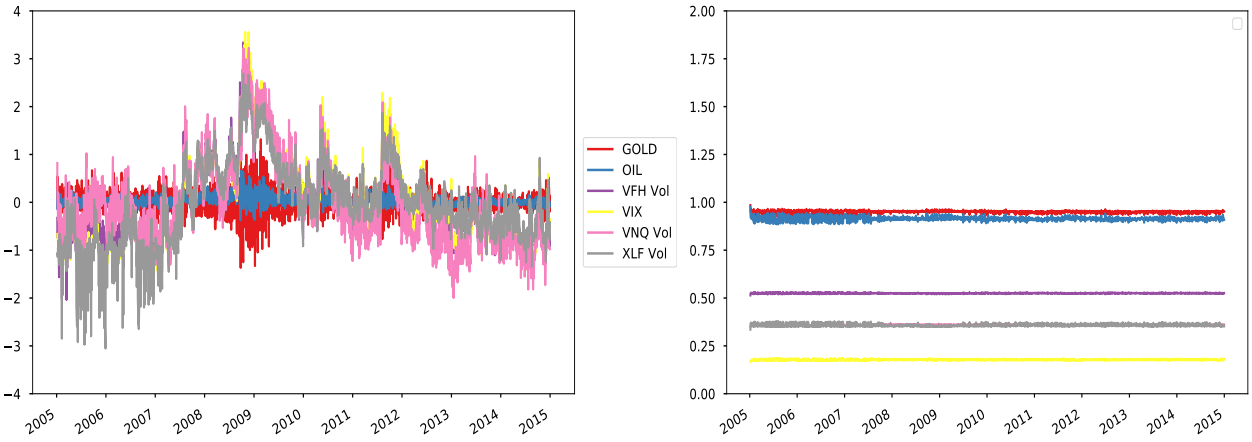
Figure 11 shows the 1-period ahead conditional forecasts for the price of oil.[14] Figure 11a shows

14. The other series' forecasts are in Section E.

Figure 10: Empirical Results with Daily Financial Series

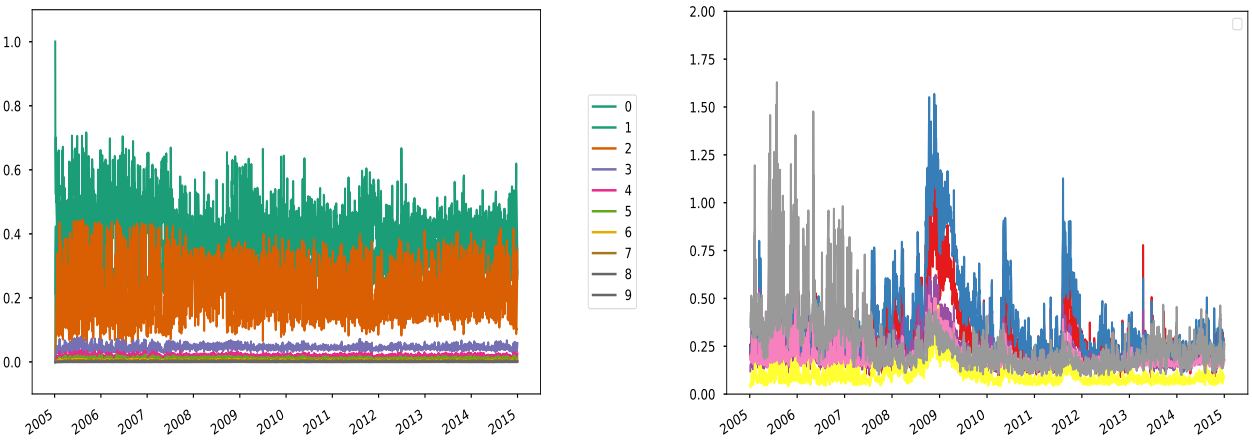
(a) Mean

(b) Innovation Volatility



(c) Number of clusters

(d) Volatility Coming from the Coefficients



that the posterior transition density tracks the dynamics of oil price somewhat well. In particular, Fig. 11b is only roughly uniformly distributed, the tails are not estimated that well. The algorithm seems to struggle in finite samples when there is different levels of fat-tails across the various series. This is likely because the current prior specification for the component coefficients does not have a hierarchical structure, and so we cannot shrink all of the intercepts as far to zero as we would like. Estimating intercepts (expected returns) for financial data is known to be quite difficult, (Lettau and Ludvigson 2010), and we are doing it in each component. Conversely, Fig. 11c shows the autocorrelation function of PIT. It displays no serial correlation.

As with the macroeconomic data, we estimated far to many parameters to report them all here. Again, we consider the transition matrix. Unlike the previous case, we need several more clusters to approximate the financial data well. In Table 3, we report the first 6 components of the infinite-dimensional transition matrix. Here we see that the probability of being in each cluster declines rapidly. Again, we see the diagonal being slightly, but not significantly larger than the off-diagonal components. There is some persistence in the clusters, but not a large amount.

Figure 11: 1-Period Ahead Conditional Forecasts: OIL

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

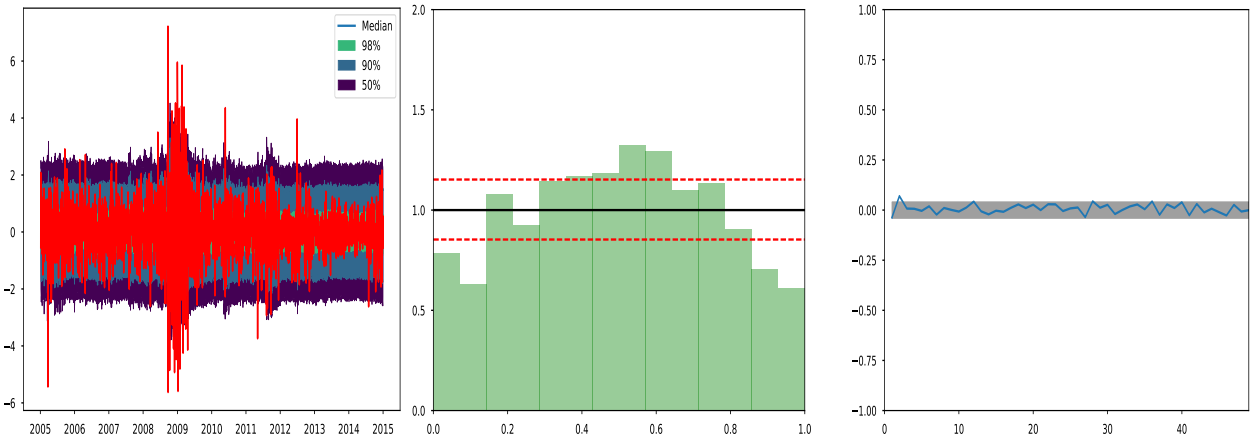


Table 3: Financial Transition Matrix

	0	1	2	3	4	5	Marginal
0	0.48	0.26	0.22	0.14	0.12	0.11	0.41
1	0.37	0.34	0.28	0.11	0.11	0.09	0.27
2	0.31	0.24	0.35	0.12	0.11	0.09	0.23
3	0.35	0.13	0.14	0.24	0.12	0.09	0.05
4	0.25	0.14	0.13	0.12	0.23	0.10	0.02
5	0.26	0.12	0.12	0.10	0.10	0.20	0.01

11 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 has been shown previously. 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. This rate is noteworthy because it is the parametric rate up to a logarithmic term. It is remarkable because these rates do not depend on the number of series.

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

Birgé, Lucien. 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,* edited by M. Banerjee, F. Bunea, J. Huang, M. Koltchinksii, and M. H. Maathius, 9:47–68. IMS Collections. Institute of Mathematical Statistics.

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart. 2013. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford, UK: Oxford University Press.

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." In Heckman and Leamer 2007, chap. 76.

Fan, Jianqing. 2005. "A Selective Overview of Nonparametric Methods in Financial Econometrics." *Statistical Science* 20 (4): 317–337.

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

Fernández-Villaverde, Jesús, and Juan Rubio-Ramírez. 2010. *Macroeconomics and Volatility: Data, Models, and Estimation.* Working Paper 16618. National Bureau of Economic Research, December.

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

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

Ghosal, Subhashis, and Aad van der Vaart. 2007. "Convergence Rates of Posteior Distributions for Non-i.i.d. Observations." *The Annals of Statitics* 35:192–223.

———. 2017. *Fundamentals of Nonparametric Bayesian Inference.* Vol. 44. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Griffin, Jim E.. 2016. "An Adaptive Truncation Method for Inference in Bayesian Nonparametric Models." *Statistics and Computing* 26, no. 1 (January): 423–441.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57 (2): 357–384.

Heckman, James J., and Edward E. Leamer, eds. 2007. Vol. 6, Part B. Handbook of Econometrics. Elsevier.

Huang, Alan, and Matt P. Wand. 2013. "Simple Marginally Noninformative Prior Distributions for Covariance Matrices." *Bayesian Analysis* 8, no. 2 (June): 439–452.

Ichimura, Hidehiko, and Petra E. Todd. 2007. "Implementing Nonparametric and Semiparametric Estimators." In Heckman and Leamer 2007, chap. 74.

Johnson, William B., and Joram Lindenstrauss. 1984. "Extensions of Lipschitz Maps into a Hilbert space." *Contemporary Mathematics* 26:189–206.

Kalli, Maria, and Jim E. Griffin. 2018. "Bayesian Nonparametric Vector Autoregressive Models." *Journal of Econmetrics* 203 (2): 267–282.

Kim, Sangjoon, Neil Shephard, and Siddhartha Chib. 1998. "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models." *The Review of Economic Studies* 65 (3): 361–393.

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

Koop, Gary, Dimitris Korobilis, and Davide Pettenuzzo. 2017. *Bayesian Compressed Vector Autoregressions.* Working Paper 2754241. Social Science Research Network, June.

Krusell, Per, and Anthony A. Smith Jr. 1998. "Income and Wealth Heterogeneity in the Macroeconomy." *Journal of Political Economy* 106 (5): 867–896.

Lettau, Martin, and Sydney C. Ludvigson. 2010. "Measuring and Modeling Variation in the Risk-Return Trade-off." Chap. 11 in *Handbook of Financial Econometrics: Tools and Techniques,* edited by Yacine A it-Sahalia and Lars Peter Hansen, 1:617–690. Handbooks in Finance. San Diego: North-Holland.

Lin, Dashua, Eric Grimson, and John Fisher. 2010. "Construction of Dependent Dirichlet Processes Based on Poisson Processes." In *Advances in Neural Information Processing Systems,* edited by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, 23:1396–1404. Curran Associates, Inc.

Nguyen, XuanLong. 2013. "Convergence of Latent Mixing Measures in Finite and Infinite Mixture Models." *The Annals of Statistics* 41, no. 1 (February): 370–400.

———. 2016. "Borrowing Strengh in Hierarchical Bayes: Posterior Concentration of the Dirichlet Base Measure." *Bernoulli* 22, no. 3 (August): 1535–1571.

Norets, Andriy. 2010. "Approximation of Conditional Densities by Smooth Mixtures of Regressions." *The Annals of Statistics* 38, no. 3 (June): 1733–1766.

Norets, Andriy, and Debdeep Pati. 2017. "Adaptive Bayesian Estimation of Conditional Densities." *Econometric Theory* 33 (4): 980–1012.

Norets, Andriy, and Justinas Pelenis. 2012. "Bayesian Modeling of Joint and Conditional Distributions." *Journal of Econometrics* 168, no. 2 (June): 332–346.

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

Pati, Debdeep, David B. Dunson, and Surya T. Tokdar. 2013. "Posterior Consistency in Conditional Distribution Estimation." *Journal of Multivariate Analysis* 116:456–472.

Patton, Andrew J.. 2012. "A Review of Copula Models for Economic Time Series." *Journal of Multivariate Analysis* 110 (September): 4–18.

Patton, Andrew J., Johanna F. Ziegel, and Rui Chen. 2018. *Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk).* Working Paper. Duke University, April.

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

Primiceri, Giorgio E.. 2005. "Time Varying Structural Vector Autoregressions and Monetary Policy." *The Review of Economic Studies* 72 (3): 821–852.

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

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

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

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

———. 1982. "Optimal Global Rates of Convergence for Nonparametric Regression." *The Annals of Statistics* 10 (4): 1040–1053.

Talagrand, Michel. 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.

Teräsvirta, Timo. 1994. "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models." *Journal of the American Statistical Association* 89 (425): 208–218.

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

van der Vaart, Aad W., and Harry J. van Zanten. 2008. "Rates of Contraction of Posterior Distributions based on Gaussian Process Priors." *The Annals of Statistics* 36, no. 3 (June): 1435–1463.

van Dijk, Dick, Timo Teräsvirta, and Philip Hans Franses. 2002. "Smooth Transition Autoregressive Models – A Survey of Recent Developments." *Econometric Reviews* 21 (1): 1–47.

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

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

Yang, Yuhong, and Andrew 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.

$\gamma_\alpha(\mathcal{X}, d) = \inf \sup_{x\in\mathcal{X}} \sum_{s=0}^{\infty} 2^{s/\alpha} d(s, \mathcal{X}_s)$

$\gamma_2(\mathcal{X}, d)$ 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 p. This is useful because a standard argument shows if X has a bounded ψ_p norm then the tail of X decays faster than $2\exp\left(-\frac{x^p}{\|x\|_{\psi_p}^p}\right)$. Hence, if x has a finite ψ_2 norm, it is subgaussain,,

$\psi_p \coloneqq \inf_{C>0} \mathbb{E}\left[\exp\left(\frac{|X|^p}{C^p} - 1\right) \leq 1\right]$

A.2 Definition and Properties of the Θ_T-operator

Lemma 2. *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 $\mu \coloneqq \Pr(b = 1)$.*

$$\Pr(K \leq \widetilde{K}) = \left(1 - (1-\mu)^{\widetilde{K}}\right)^T \tag{55}$$

Proof.

$$\begin{aligned}
\Pr(K \leq \widetilde{K}) &= \Pr(\text{All of the rows contain at least one one}) && (56)\\
&= \Pr(\text{Row } t \text{ contains at least one one})^T && (57)\\
&= (1 - \Pr(\text{row } t \text{ contains all zeros.}))^T && (58)\\
&= \left(1 - (1-\mu)^{\widetilde{K}}\right)^T && (59)
\end{aligned}$$

□

Lemma 3. *There exists a constant $\gamma \in (0,1)$ and constants c_1, c_2, such that with probability at least γ, the following holds.*

$$c_1 \log(T) \leq K \leq c_2 \log(T) \tag{60}$$

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.

$$\begin{aligned}
(1-\gamma) &= (1 - (1-\mu)^{\widetilde{K}})^T && (61)\\
\implies \log(1-\gamma)/T &= \log(1 - (1-\mu)^{\tilde{K}}) && (62)
\end{aligned}$$

We use the bases in this way so that the change of base constant is positive.

$$\implies \log(1-\gamma)/T = \log\left(1 - \left(\frac{1}{1-\mu}\right)^{-\log B}\right) \tag{63}$$

Using a change-of-base formula.

$$\begin{aligned}
\implies \log(1-\gamma)/T &= \log\left(1 - \left(\frac{1}{B}\right)^{-\log(1-\mu)}\right) && (64)\\
\implies \log(1-\gamma)/T &= \log\left(1 - B^{\log(1-\mu)}\right) && (65)
\end{aligned}$$

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

$$\begin{aligned}
&\implies -\log(1-\gamma)/T \approx B^{\log(1-\mu)} && (66)\\
&\implies T \propto B^{-\log(1-\mu)} && (67)\\
&\implies B \propto T^{-1/\log(1-\mu)} && (68)\\
&\implies K \propto -\frac{1}{\log(1-\mu)} \log(T) && (69)\\
&\implies K \propto \log(T) && (70)
\end{aligned}$$

We can bound this in the opposite direction by replacing $1-\gamma$ with γ.

□

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 4. *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. The root of the proof follows from realizing that Θ_T is a generalized selection matrix, and covariances are dominated by variances.

$$\mathbb{E}_\Theta\left[X'\Theta_k\Theta_k'X\right] = \mathbb{E}_\Theta\left[\sum_{t=1}^T x_t\theta_{t,k}\theta_{t,k}x_t'\right] = \mathbb{E}_{\Theta_k}\left[\sum_{t=1}^T|\theta_{t,k}|x_tx_t'\right] \tag{71}$$

Simplifying this using independence of the rows of Θ_k.

$$= \frac{1}{K}\sum_{t=1}^T x_tx_t' \tag{72}$$

Note, this implies that $\mathbb{E}_\Theta\{\theta_kX\} = \|x\|_{L_2}^2$ because they are both sums over all of the elements considered.

Now, 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.

$$\mathbb{E}_\Theta\left[X'\Theta\Theta'X\right] = \sum_{k=1}^K \mathbb{E}_{\Theta_k}\left[X'\Theta_k\Theta_k'X\right] = \sum_{t=1}^T x_tx_t' \tag{73}$$

Part 8.2. Now, that we have shown Item 1, we need to 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.

$$\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{|\langle\Theta_k,x\rangle|^2}{C^2}\right)\right]-1\leq 1\right\} \tag{74}$$

$$=\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{\sum_{t=1}^{T}|\delta_{t,k}|x_t'x_t+2\sum_{t,\tau\neq t}\delta_{t,k}\delta_{\tau,k}x_t'x_\tau}{C^2}\right)\right]\leq 2\right\} \tag{75}$$

Since the cross-terms are proportional to squares, and the Θ_k are generalized selection vectors.

$$\leq\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{2\sum_{t=1}^{T}|\delta_{t,k}|x_t'x_t}{C^2}\right)\right]\leq 2\right\} \tag{76}$$

By the definition of the exponential function and $|\delta_{t,k}|\in\{0,1\}$.

$$=\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\sum_{h=0}^{\infty}\frac{2^h\left(\sum_{t=1}^{T}|\delta_{t,k}|x_t'x_t\right)^h}{C^{2h}h!}\right]\leq 2\right\} \tag{77}$$

$$=\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\sum_{h=0}^{\infty}\frac{2^h\sum_{\sum k_t=h}\binom{h}{k_1,k_2,\dots k_T}\prod_{t=1}^{T}|\delta_{t,k}|(x_t'x_t)^{k_t}}{C^{2h}h!}\right]\leq 2\right\} \tag{78}$$

Since everything is absolutely convergent, we can interchange expectations and infinite sums.

$$=\inf\left\{C>0\,\middle|\,\sum_{h=0}^{\infty}\frac{2^h\sum_{\sum k_t=h}\binom{h}{k_1,k_2,\dots k_T}\prod_{t=1}^{T}\frac{1}{K}(x_t'x_t)^{k_t}}{C^{2h}h!}\leq 2\right\} \tag{79}$$

Then we can use the multinomial theorem and the formula for the exponential function in the opposite direction.

$$=\inf\left\{C>0\,\middle|\,\frac{1}{K}\exp\left(\frac{2\|x\|_{L_2}^2}{C^2}\right)\leq 2\right\} \tag{80}$$

$$=\inf\left\{C>0\,\middle|\,\frac{2\|x\|_{L_2}^2}{C^2}=\log(2K)\right\} \tag{81}$$

Since $K\geq 1$.

$$\leq\frac{\sqrt{2}\|x\|_{L_2}}{\sqrt{\log(2)}} \tag{82}$$

Therefore, if we set $\beta\propto\sqrt{\frac{\log(2)}{2}}$, 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 $W_x \coloneqq \Upsilon(|Z_x|)$ and $\epsilon \coloneqq \frac{\gamma_2(\mathcal{X},d)}{\sqrt{K}}$. 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 $u < \delta_0 \coloneqq \frac{4}{\eta}\log\frac{1}{\delta}$,*

$$\Pr\left(|Z_x - Z_y| > ud(x,y)\right) < \exp\left(-\frac{\eta}{\delta_0}Ku^2\right) \tag{83}$$

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

$$\Pr\left(|W_x - W_y| > ud(x,y)\right) < \exp\left(-\eta K u^2\right) \tag{84}$$

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 $1-2\delta$, where $C(\Upsilon,\delta,\eta) \coloneqq \left(c(\Upsilon)c(\eta)(\frac{2}{\eta}(\log\frac{1}{\delta}+1))\right) > 0$ where $c(\Upsilon)$ and $c(\eta)$ depend solely on their arguments.

$$\sup_{x\in\mathcal{X}}|Z_x| < C(\Upsilon,\delta,\eta)\epsilon. \tag{85}$$

Here we quote a version of Bernstein's inequality for martingales due to (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 $x, y > 0$,*

$$\Pr\left(\left\{\left|\sum_{i=1}^k x_i\right| \ge u, v_k \le y \text{ for some } k\right\}\right) \ge 2\exp\left(-\frac{u^2}{2(y+uM)}\right) \tag{86}$$

If we choose c small enough, this implies the following.

$$\Pr\left(\left\{\left|\frac{1}{k}\sum_{i=1}^k x_i\right| \ge u, v_k \le y \text{ for some } k\right\}\right) \ge 2\exp\left(-c\min\left\{\frac{u^2k^2}{v}, \frac{uk}{M}\right\}\right) \tag{87}$$

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 $\epsilon > \sqrt{\frac{\log T}{K_T}}$*

$$\sup_t |\|\theta_t \widetilde{x}_t\|_{L_2} - \|\widetilde{x}_t\|_{L_2}| < c_2 \left(1 + \log \frac{1}{\delta}\right) \epsilon. \tag{5}$$

Proof. We mimic the proof of Klartag and Mendelson 2005, Theorem 3.1, verifying the conditions of Proposition 9. Similar to them we use $\Upsilon(t) = \sqrt{1 = t}$. Our conclusion is stated in terms of the logarithm of the sample size — T. This is a weaker conclusion 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, (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 $\sum_{d=1}^{D} \theta_k' x_d$.

$$Z_x^K := \frac{1}{K} \sum_{k=1}^{K} x^2(\theta_k) - \|x\|_{L_2}^2 \tag{88}$$

Consider $Z_x^K - Z_y^K$.

$$Z_x^K - Z_y^K = \frac{1}{K} \sum_{k=1}^{K} x^2(\theta_k) - y^2(\theta_k) = \frac{1}{K} \sum_{k=1}^{K} (x - y)(\theta_k)(x + y)(\theta_k) \tag{89}$$

Part 10.1. Let $Y_k := x^2(\theta_k) - y^2(\theta_k)$.

$$\begin{aligned} &\Pr(|Y_k| > 4u\|x - y\|_{\psi_2}\|x + y\|_{\psi_2}) && (90)\\ &\quad \leq \Pr(|Y_k| > 2\sqrt{u}\|x - y\|_{\psi_2}) + \Pr(|Y_k| > 2\sqrt{u}\|x + y\|_{\psi_2}) && (91)\\ &\quad \leq 2\exp(-u) && (92) \end{aligned}$$

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 $v := 4K\|\theta\|_{\psi_1}^2$ and $M = \|\theta\|_{\psi_1}$:

$$\Pr\left(\left\{\left|\frac{1}{K}\sum_{k=1}^{K}\theta_k\right| > u\right\}\right) \leq 2\exp\left(-cK\min\left\{\frac{u^2}{\|\theta\|_{\psi_1}^2}, \frac{u}{\|\theta\|_{\psi_1}}\right\}\right) \tag{93}$$

Then by applying Eq. (93) to $\Pr(|z_x^k - z_y^k| > u)$, we have the following.

$$\Pr\left(\left|Z_x^k - Z_y^k\right| > u\right) \leq 2\exp\left(-c\min\left\{\frac{u^2}{\|x-y\|_{L_2}^2}, \frac{u}{\|x-y\|_{L_2}}\right\}\right) \tag{94}$$

The estimate for $\Pr\left(\left|Z_x^k\right| > u\right)$ follows from the same method but we define $Y_k := x^2(\theta_k) - 1$, and use the fact that $\|x(\theta)\|_{\psi_2} \leq 1$, which we verified in Lemma 4. 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

$$\Pr\left(\left|Z_x^k - Z_y^k\right| > 2\|x-y\|_{L_2}\right) \leq 2\exp\left(\eta K\min\left\{u, u^2\right\}\right) < \exp\left(-\eta K\frac{u^2}{\delta_0}\right). \tag{95}$$

Part 10.2.

$$W_x - W_y = \left(\frac{1}{K}\sum_{k=1}^{K}x^2(\theta_i)\right)^{1/2} - \left(\frac{1}{K}\sum_{k=1}^{K}y^2(\theta_i)\right)^{1/2} \leq \left(\frac{1}{K}\sum_{k=1}^{K}(x-y)^2(\theta_i)\right)^{1/2} \tag{96}$$

Applying Eq. (93) for $u > 1$.

$$\begin{aligned}
\Pr\left(|W_x - W_y| > u\|x-y\|_{\psi_2}\right) &\leq \Pr\left(\frac{1}{K}\sum_{k=1}^{K}(x-y)^2(\theta_k) > u^2\|x-y\|_{\psi_2}^2\right) &(97)\\
&\leq \Pr\left(\frac{1}{K}\sum_{k=1}^{K}(x-y)^2(\theta_k) > u^2\|x-y\|_{\psi_1}^2\right) &(98)\\
&< \exp\left(-cku^2\right) &(99)
\end{aligned}$$

Since $\eta < c$.

$$\leq \exp\left(-\eta K u^2\right) \tag{100}$$

Part 10.3. For any $x \in \mathcal{X}$ by Eq. (93),

$$\Pr(|Z_x| > \epsilon) < \exp(-\eta K\epsilon^2) < \delta \tag{101}$$

Part 10.4.

$$\Upsilon'(0) = 1/2 > 0 \tag{102}$$

□

Appendix B Representation Theory

B.1 The Joint Density

Lemma 5. *Consider the ratio of the densities between $p_{0,T}$ 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 $\mu_k^q = \mathbb{E}_{P_{0,T}}\left[x_t \,\middle|\, t \in \delta_k^q\right]$ and $\Sigma_k^q = \mathbb{C}\mathrm{ov}_{P_{0,T}}\left[x_t \,|\, x_t \in \delta_t^q\right]$:*

$$\sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - \left(x_t - \mu_k^q\right)' \left(\Sigma_k^q\right)^{-1} \left(x_t - \mu_k^q\right) \right| < C(\delta)\epsilon. \tag{103}$$

Then the log-divergence satisfies

$$\sup_{x_t, x_t^*} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - (x_{t^*} - \mu_{t^*})' \Sigma_{t^*}^{-1} (x_{t^*} - \mu_{t^*}) \right| < \epsilon \implies \sup_{x_t, x_t^*} \left| \log \frac{p_0(x_t)}{p_0(x_{t^*})} \right| \propto \epsilon. \tag{104}$$

Proof. Consider the log ratio of Gaussian kernels:

$$-\frac{1}{2} \sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - \left(x_t - \mu_k^q\right)' \left(\Sigma_k^q\right)^{-1} \left(x_t - \mu_k^q\right) \right| \tag{105}$$

(The right-hand side is bounded by assumption.)

$$\leq -\frac{1}{2}\epsilon.$$

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

$$\frac{1}{\chi^p} = \int_{\mathcal{X}} k^p(x)\, dx, \quad \frac{1}{\chi^q} = \int_{\mathcal{X}} k^q(x)\, dx. \tag{106}$$

By the definition of proportionality constant, we can write

$$\left| \frac{\chi^q}{\chi^p} \right| = \frac{\int k^q(x)\, dx}{\int k^p(y)\, dy}. \tag{107}$$

Since $\frac{1}{x}$ is a convex function, by Jensen's inequality:

$$\leq \int \frac{\int k_2(x)\, dx}{k_1(y)}\, dy \tag{108}$$

$$= \int \int \frac{k^q(x)}{k^p(y)}\, dx\, dy. \tag{109}$$

(We can change measures to $P_{0,T}$. The Jacobian terms for each of the two densities cancel.)

$$= \int \int \frac{k^q(x)}{k^p(y)} \, dP_{0,T}(x) \, dP_{0,T}(y). \tag{110}$$

That will be less than the maximum ratio deviation integrated appropriately:

$$\leq \int \sup_{\delta_k^q} \sup_{x_t \in \delta_t^q} \frac{k^q(x_t)}{k^p(x_t)} \, dP_{0,T}(x). \tag{111}$$

We can rewrite the integral with respect to $P_{0,T}$ as an average:

$$= \frac{1}{T} \sum_t \sup_{\delta_k^q} \sup_{x_t \in \delta_t^q} \frac{k^q(x_t)}{k^p(x_t)}. \tag{112}$$

We can split the sum up into a sum within the groups and a sum over the groups:

$$= \frac{1}{T} \sum_{\delta_k^q} \sup_{\delta_k^q} \sum_{x_t \in \delta_k^q} \sup_{x_t \in \delta_t^q} \frac{k^q(x_t)}{k^p(x_t)}. \tag{113}$$

By Eq. (105), the kernel ratio is bounded above, and the second part is just a density integrated over its entire domain:

$$\leq \frac{1}{T} T \exp\left(-\frac{1}{2} c(\delta)\epsilon\right) \tag{114}$$

$$\leq \exp\left(-\frac{1}{2} c(\delta)\epsilon\right). \tag{115}$$

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 the following.

$$\left|\log \frac{\mu_1}{\mu_2}\right| = \frac{1}{2} c(\delta)\epsilon \tag{116}$$

Hence the log ratio of the densities is the sum of the log-ratio of the kernels and the log-ratio of the proportionality constants for some global constants C_1 and C_2.

$$\left|\log \frac{p_{0,T}(x_t)}{q_T(x_t)}\right| = C_1 C(\delta)\epsilon, \quad \left|\log \frac{q_T(x_t)}{p_{0,T}(x_t)}\right| = C_2 C(\delta)\epsilon \tag{117}$$

□

Proposition 11. *Let $\widetilde{X} \coloneqq \widetilde{x}_1, \ldots \widetilde{x}_T$ be a D-dimensional Gaussian process with 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 $\widetilde{P}_{0,T}$ denote the distribution of $\widetilde{X}$. Then given $\epsilon > 0$ and for some $\delta \in (0,1)$, the approximating distribution Q_T, which is the mixture distribution over $\widetilde{X}$ defined*

by the clustering induced by Θ_T satisfies the following with probability at least $1-2\delta$ with respect to Θ_T.

$$\sup_t h^2\left(\int_{G_t}\phi\left(\widetilde{x}_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P), \int_{G_t}\phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\right) < c\left(\log\frac{1}{\delta}\right)^2\epsilon^2 \tag{118}$$

Proof. In this proof, we drop the tilde's over the x_t because all of the terms have them.

$$\sup_t h^2\left(\int_{G_t}\phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P), \int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\right) \tag{119}$$

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 squaured Hellinger distance gives:

$$\leq \int_{G_t\times G_t}\sup_t h^2\left(\phi\left(x_t \,|\, \delta_t^P\right), \phi\left(x_t \,\middle|\, \delta_t^Q\right)\right)\mathcal{K}(G_t^P, G_t^Q). \tag{120}$$

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

$$\leq \int_{G_t\times G_t}\sup_t\int_{\mathbb{R}^D}\left|\left(\frac{\phi\left(x_t \,|\, \delta_t^P\right)}{\phi\left(x_t \,\middle|\, \delta_t^Q\right)}\right)^{1/2} - 1\right|^2 d\Phi\left(x_t \,\middle|\, \delta_t^Q\right)\mathcal{K}(G_t^P, G_t^Q). \tag{121}$$

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

$$= \int_{G_t\times G_t}\int_{\mathbb{R}^D}\sup_t\left|\left(\frac{\phi\left(x_t \,|\, \delta_t^P\right)}{\phi\left(x_t \,\middle|\, \delta_t^Q\right)}\right)^{1/2} - 1\right|^2 d\Phi\left(x_t \,\middle|\, \delta_t^Q\right)\mathcal{K}(G_t^P, G_t^Q). \tag{122}$$

By Lemma 5 and a first-order Taylor series of the exponential function after pulling the square-root inside

$$\leq \int_{G_t\times G_t}\int_{\mathbb{R}^D}\sup_t\left|(x_t-\mu_t^P)'\Sigma_t^P(x_t-\mu_t^P) - (x_t-\mu_t^Q)\Sigma_t^Q(x_t-\mu_t^Q)\right| d\Phi\left(x_t \,\middle|\, \delta_t^Q\right)\mathcal{K}(G_t^P, G_t^Q). \tag{123}$$

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

$$\leq C\left(\log\frac{1}{\delta}\right)^2\int_{G_t\times G_t}\int_{\mathbb{R}^D}|\epsilon|^2\, d\Phi\left(x_t \,\middle|\, \delta_t^Q\right)\mathcal{K}(G_t^P, G_t^Q). \tag{124}$$

All of the integrals integrate to 1:

$$= C \left(\log \frac{1}{\delta} \right)^2 \epsilon^2. \tag{125}$$

□

Theorem 2 (Representing the Joint Density)**.** *Let $\widetilde{X}_T \coloneqq \widetilde{x}_1, \dots \widetilde{x}_T$ be a D-dimensional Gaussian process with period t 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 constructed in Definition 3. Let $\widetilde{P}_{0,T}$ denote the distribution of $\widetilde{X}_T$. Then given $\epsilon > 0$ and for some $\delta \in (0, 1)$, the approximating distribution Q_T, which is the mixture distribution over $\widetilde{\mathcal{X}}$ defined by the clustering induced by Θ_T satisfies the following with probability at least $1 - 2\delta$ with respect to Θ_T*

$$h_\infty \left(\widetilde{P}_{0,T}(\widetilde{\mathcal{X}}), \widetilde{Q}_T(\widetilde{\mathcal{X}}) \right) < C \log \left(\frac{1}{\delta} \right) \epsilon. \tag{12}$$

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 $P_{0,T}$ 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. Then we can represent both densities succinctly as follows. Importantly, we do not require that the G^P_t are independent.

$$p_{0,T}(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi \left(x_t \,|\, \delta^P_t \right) dG^P_t \left(\delta^P_t \right) dG^P \left(dG^P_t \right) \tag{126}$$

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

$$q_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi \left(x_t \,\middle|\, \delta^Q_t \right) dG^Q_t \left(\delta^Q_t \right) dG^Q \left(dG^Q_t \right) \tag{127}$$

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

$$h^2_\infty \left(p_{0,T}(\widetilde{\mathcal{X}}), q_T(\widetilde{\mathcal{X}}) \right) \tag{128}$$

$$= h^2_\infty \left(\int \int \phi \left(x_t \,|\, \delta^P_t \right) dG^P_t(\delta^P_t) \, dG^P(dG^P), \int \int \phi \left(x_t \,\middle|\, \delta^Q_t \right) dG^Q_t(\delta^Q_t) \, dG^Q(dG^Q_t) \right) \tag{129}$$

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.

$$= h^2_\infty \left(\int_G \int_{G_t} \phi \left(x_t \,|\, \delta^P_t \right) dG^P_t(\delta^P_t) \, d\mathcal{K}(dG^P_t, dG^Q_t), \int_G \int_{G_t} \phi \left(x_t \,\middle|\, \delta^Q_t \right) dG^Q_t(\delta^Q_t) \mathcal{K}(dG^P_t, dG^Q_t) \right) \tag{130}$$

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

$$\leq \int_{G\times G} \sup_t h^2 \left(\int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q) \right) d\mathcal{K}(dG_t^P, dG_t^Q) \tag{131}$$

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 Eq. (131). 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 $1 - 2\delta$.

$$h_\infty^2(P_{0,T}(\widetilde{\mathcal{X}}), Q_T(\widetilde{\mathcal{X}})) < C\left(\log\frac{1}{\delta}\right)^2 \epsilon^2 \tag{132}$$

□

Lemma 6. *Let f, g be two densities of locally asymptotically normal (LAN) processes.*[15] *Squared Hellinger distance and Kullback-Leibler divergence are equivalent.*

Proof. Consider the following decomposition.

$$\int (\sqrt{f/g} - 1)\, dG \tag{133}$$

$$= \int \left(\exp\left(\frac{1}{2}(\log f - \log g) \right) - 1 \right) dG \tag{134}$$

Taking a mean-value expansion of the exponential function.

$$= \int \left(1 + \frac{1}{2}\log\frac{f}{g} + O\left(\log\left(\frac{f}{g}\right)^2 \right) - 1 \right) dG \tag{135}$$

$$= \int \frac{1}{2}\log\frac{f}{g}\, dG + \int O\left(\log\left(\frac{f}{g}\right)^2 \right) dG \tag{136}$$

most By the locally asymptotically normal assumption $\log f(x) \propto (x-\mu_f)'\Sigma_f^{-1}(x-\mu_f) + o(T)$ Choose $\epsilon \propto \frac{1}{T}$.

$$\log(f/g)^2 \tag{137}$$

By the convexity of the square function.

$$\leq \left| (x-\mu_f)'\Sigma_f^{-1}(x-\mu_f) - (x-\mu_g)'\Sigma_g^{-1}(x-\mu_g) \right| + O(\epsilon)O(\epsilon) \tag{138}$$

15. This trivially covers all Gaussian processes.

We can bring back in the log deviation at the expense of at most a ϵ term.

$$\leq \left|(x-\mu_f)'\Sigma_f^{-1}(x-\mu_f) - (x-\mu_g)'\Sigma_g^{-1}(x-\mu_g)\right|^2 + O(\epsilon^2) \tag{139}$$

The first term in the above expansion is clearly bounded by $\log(f/g)$. Consequently, $\mathrm{D}_{\mathrm{KL}}(f \,||\, g)$ bounds squared Hellinger. We can see from the Taylor series expansion in Eq. (135) that it is also bounded by the squared Hellinger distance as well.

□

B.2 Representing the Marginal Density

Theorem 3 (Representing the Marginal Density)**.** *Let $x_1, \ldots x_T$ be drawn from $p_{0,T}$, where $p_{0,T}$ has a product density. Let Θ_T be constructed as in Theorem 2 for each t. Let ϵ be given. We construct q_T by using the Θ_T operator to group the data, and we assume that the data are 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 in T*

$$h\left(\int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t(\delta_t^P), \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t(\delta_t^Q)\right) < C \log\left(\frac{1}{\delta}\right)\epsilon. \tag{13}$$

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 strange 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 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.

$$\prod_{t=1}^{T} \int_{G_t} \phi\left(x_t \,|\, \delta_t\right) dG_t(\delta_t) = \prod_{t=1}^{T} \int_{G_m} \phi\left(x_t \,|\, \delta_t\right) dG_m(\delta_t) \tag{140}$$

All it is saying is that the joint T independent draws from the marginal are the same as T independent draws from a sequence of $G_1, \ldots G_T$, drawn from G. Note, by assumption G has a product form, else this would not hold. The Kullback-Leibler divergence between the two joint distributions is

$$\mathrm{D_{KL}}\left(q_T \,||\, p_{0,T}\right) = \int_{\mathbb{R}^{T\times D}} \log\left(\frac{q_T}{p_{0,T}}\right) dP_{0,T} \tag{141}$$

$$= \int_{\mathbb{R}^{T\times D}} \log\left(\frac{\prod_{t=1}^{T}\int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\prod_{t=1}^{T}\int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) dP_{0,T} \tag{142}$$

Ratios of products are products of ratios, and logs of products are sums of logs.

$$= \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_t}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) dP_{0,T} \tag{143}$$

As noted in the definition of the marginal distribution, Eq. (140). Noting that both the $P_{0,T}$ and Q_T are product distributions.

$$= \int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_m^Q(\delta_t^Q)}{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_m^P(\delta_t^P)}\right) dP_{0,T} \tag{144}$$

We can rewrite $P_{0,T}$ in terms of its mixture representation. Note, the δ_t with respect to $P_{0,t}$ have no superscript because they are different variables.

$$= \int_{G_t}\int_{\mathbb{R}^{T\times D}} \sum_{t=1}^{T} \log\left(\frac{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right) \prod_{t=1}^{T}\phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t) \tag{145}$$

The only interaction between the two terms is x_t.

$$= \sum_{t=1}^{T}\left(\left(\int_{G_t}\int_{\mathbb{R}^D} \log\left(\frac{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}\right)\phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t)\right)\right. \tag{146}$$
$$\left.\left(\int_{\mathbb{R}^{(T-1)\times D}}\prod_{\tau\neq t}\phi\left(x_\tau \,|\, \delta_\tau\right) dx\, dG_m^P(\delta_\tau)\right)\right)$$

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

$$= \sum_{t=1}^{T}\left(\int_{G_t}\int_{\mathbb{R}^D} \log\left(\frac{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P)}{\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)}\right)\phi\left(x_t \,|\, \delta_t\right) dx\, dG_m^P(\delta_t)\right) \tag{147}$$

The term inside the sum is the Kullback-Leibler divergence between the two marginal distributions.

$$= \sum_{t=1}^{T} \mathrm{D_{KL}}\left(\int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_m^Q(\delta_t^Q) \,\middle\|\, \int_{G_m}\phi\left(x_t \,\middle|\, \delta_t^P\right) dG_m^P(\delta_t^P)\right) \tag{148}$$

The marginal distributions does not depend upon T.

$$= T\mathrm{D}_{\mathrm{KL}}\left(\int_{G_m}\phi\left(x_t\,\middle|\,\delta_t^Q\right)dG_m^Q(\delta_t^Q)\,\middle\|\,\int_{G_m}\phi\left(x_t\,|\,\delta_t^P\right)dG_m^P(\delta_t^P)\right) \tag{149}$$

In other words, the distance between the joint densities is at least T times the distance between the distance marginal densities. Also, by Lemma 6 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 $\|X\|^2 \propto T$.

$$h^2\left(\int_{G_m}\phi\left(x_t\,\middle|\,\delta_t^Q\right)dG_m^Q(\delta_t^Q),\int_{G_m}\phi\left(x_t\,|\,\delta_t^P\right)dG_m^P(\delta_t^P)\right)\leq\frac{1}{T}h^2(q_T,p_{0,T})\leq C\frac{T}{T}\epsilon^2=C\epsilon^2 \tag{150}$$

□

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 $p_{0,T}$. 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 such that the following holds:*

$$q_T\left(x_t\,|\,x_{t-1},\delta_{t-1}\right) := \sum_{k=1}^{K}\phi\left(\beta_k x_{t-1},\Sigma_k\right)\Pr\left(\delta_t = k\,|\,\delta_{t-1}\right). \tag{14}$$

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

$$h_\infty\left(p_{0,T}\left(x_t\,|\,\mathcal{F}_{t-1}^P\right),q_T\left(x_t\,\middle|\,\mathcal{F}_{t-1}^Q\right)\right) < C\sqrt{\log\frac{1}{\delta}}\epsilon. \tag{15}$$

Proof. We need the conditional density of $\widetilde{x}_t \mid \widetilde{x}_{t-1}, \delta_{t-1}$. 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.)

$$\theta_t \widetilde{x}_t \mid \widetilde{x}_{t-1}, \theta_t, \theta_{t-1} \stackrel{\mathcal{L}}{=} \theta_t \widetilde{x}_t \mid \theta_{t-1}\widetilde{x}_{t-1}, \theta_t, \theta_{t-1} \tag{151}$$

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

$$\stackrel{\mathcal{L}}{=} \phi(\beta_{k,k'}\theta_{t-1}\widetilde{x}_{t-1}, \Sigma_{k,k'}) \tag{152}$$

Then since 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 will take the weighted averages over them in the end.

$$\stackrel{\mathcal{L}}{=} \phi(\beta_{k,k'}\widetilde{x}_{t-1}, \Sigma_{k,k'}) \tag{153}$$

In fact, 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 $\widetilde{x}$.

$$\widetilde{x}_t \mid \theta_{t-1} = k, \widetilde{x}_{t-1} \sim \sum_{k'} \phi(\beta_{k,k'}\widetilde{x}_{t-1}, \Sigma_{k,k'}) \Pr\left(\theta_t = k'\right) \tag{154}$$

Define a set of clusters in $(\widetilde{x}_t, \widetilde{x}_{t-1})$ space by grouping the ones whose associated β's are equal. In other words, take the Cartesian product of the clusters used in Eq. (154) and denote the cluster identities by δ_t's. Integrating out the cluster identities gives

$$\widetilde{x}_t \mid \widetilde{x}_{t-1}, \delta_{t-1} \sim \sum_{j} \phi(\beta_j\widetilde{x}_{t-1}, \Sigma_j) \Pr\left(\delta_t = j \mid \delta_{t-1}\right) \tag{155}$$

Clearly, there are $\log(T)^2 = K_T^2$ different clusters.[16]

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.

Again, we start by considering the Kullback-Leibler divergence between the joint distributions. We assumed that $p_{0,T}$ 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 the vector of cluster identities with respect to $P_{0,T}$. Consider the supremum of the deviations in

16. The number of clusters used here is of the same asymptotic order as in the prior. Also, this bound may no longer be tight.

each period:

$$\sup_t \mathrm{D}_{\mathrm{KL}}\left(\int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P) \,\middle\|\, \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\right). \tag{156}$$

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:

$$= \sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \sup_t \mathrm{D}_{\mathrm{KL}}\left(\int_{G_M} \phi\left(x_t \,\middle|\, \delta_t^P\right) dG_M^P\left(\delta_t^P \,\middle|\, \mathcal{F}_{t-1}^P\right) \,\middle\|\, \int_{G_M} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_M^Q\left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right). \tag{157}$$

The goal is to show that the integral of Eq. (157) with respect to $P_{0,T}$ can be rewritten as thee the sum of the individual conditionals.

Again, we start by considering the Kullback-Leibler divergence between the joint distributions. We assumed $p_{0,T}$ 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 $P_{0,T}$.

$$\mathrm{D}_{\mathrm{KL}}\left(P_{0,T} \,||\, Q_T\right) = \mathrm{D}_{\mathrm{KL}}\left(\prod_{t=1}^T p_{0,T}\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right) \,\middle\|\, \prod_{t=1}^T q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right)\right) \tag{158}$$

Since the Kullback-Leibler divergence of product densities is the sum of the Kullback-Leibler divergences.

$$= \sum_{t=1}^T \mathrm{D}_{\mathrm{KL}}\left(p_{0,T}\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right) \,\middle\|\, q_T(x_t \,\middle|\, \mathcal{F}_{t-1}^Q)\right) \tag{159}$$

$$\leq T \sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \mathrm{D}_{\mathrm{KL}}\left(p_{0,T}\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right) \,\middle\|\, q_T(x_t \,\middle|\, \mathcal{F}_{t-1}^Q)\right) \tag{160}$$

$$\int_{\mathbb{R}^{T\times D}} \sup_t \log \frac{q_T(X)}{p_{0,T}(X)} \, dP_{0,T} \tag{161}$$

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 $x_1, \ldots x_{t-1}$, we can rewrite this as follows.

$$= \int_{\mathbb{R}^{T\times D}} \left(\sup_{x_t, \mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} \log \frac{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^Q\right) dG_M^Q\left(\delta_t^Q \,\middle|\, \mathcal{F}_{t-1}^Q\right)}{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta_t^P\right) dG_P\left(\delta_t \,\middle|\, \mathcal{F}_{t-1}^P\right)}\right) dP_{0,T} \tag{162}$$

Clearly, the supremum with respect to x_t is greater than the average with respect to the x_t.

$$\geq \int_{\mathbb{R}^{T\times D}} \left(\sup_{\mathcal{F}^P_{t-1},\mathcal{F}^Q_{t-1}} \log \frac{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right) dG_P\left(\delta_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} \right) dP_{0,T} \tag{163}$$

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^P_t and δ^Q_t.

$$= \int_{\mathbb{R}^{T\times D}} \int_{G^P\times G^Q} \left(\sup_{\mathcal{F}^P_{t-1},\mathcal{F}^Q_{t-1}} \log \frac{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right) dG_P\left(\delta_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} \right) \tag{164}$$
$$d\mathcal{K}\left(dG^P(\Delta^P), dG^Q(\Delta^Q)\right) dP_{0,T}$$

Conditional on Δ^P, Δ^Q, $\mathcal{F}^P_{t-1}$ and $\mathcal{F}^Q_{t-1}$ contain no information regarding δ^P_t and δ^Q_t. By the law of iterated expectations, we can rewrite this as integral with respect to the joint distribution as we did above.

$$= \int_{\mathbb{R}^{T\times D}} \left(\int_{G^P\times G^Q} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \log \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} d\mathcal{K}(dG^P(\Delta^P), dG^Q(\Delta^Q) \right) dP_{0,T} \tag{165}$$

Factoring $\mathcal{K}$.

$$= \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \log \left(\prod_{t=1}^{T} \int_{G^P_t\times G^Q_t} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} d\mathcal{K}\left(dG^P_t(\delta^P_t), dG^Q_t(\delta^Q_t) \,\middle|\, \mathcal{F}^Q_{t-1}, \mathcal{F}^P_{t-1}\right) \right) dP_{0,T} \tag{166}$$

The Markov assumption on x_t, z_t implies that the δ^P_t and δ^Q_t will be 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.

$$= \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \sum_{t=1}^{T} \log \left(\int_{G^P_M\times G^Q_M} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} \right. \tag{167}$$
$$\left. d\mathcal{K}\left(dG^P_M(\delta^P_t), dG^Q_M(\delta^Q_t) \,\middle|\, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) \right) dP_{0,T} \tag{168}$$

We can break the joint distribution into a conditional and a marginal.

$$= \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \sum_{t=1}^{T} \log \left(\int_{G^Q_M} \int_{G^P_M} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} \right. \tag{169}$$
$$\left. dG^P_M\left(\delta^P_t \,\middle|\, \delta^Q_t, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) \right) dP_{0,T}$$

Since the logarithm is a concave function, we can use Jensen's inequality.

$$\geq \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \sum_{t=1}^{T} \int_{G^Q_M} \log \left(\int_{G^P_M} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} dG^P_M\left(\delta^P_t \,\middle|\, \delta^Q_t, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) \right) \tag{170}$$
$$dG^Q_M\left(\delta^Q_t \,\middle|\, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) dP_{0,T}$$

We can distribute the sum into the sum of the groups and the sum within each group.

$$\geq T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \frac{1}{T} \sum_{\delta^Q_k} \int_{G^Q_M} \sum_{t\in\delta^Q_k} \log \left(\int_{G^P_M} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} dG^P_M\left(\delta^P_t \,\middle|\, \delta^Q_t, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) \right) \tag{171}$$
$$dG^Q_M\left(\delta^Q_t \,\middle|\, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) dP_{0,T}$$

Within each group δ^Q_K the log of the integral with respect to δ^P_t is a constant. Let T^Q_k be the number of t in group δ^Q_k

$$= T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \frac{1}{T} \sum_{\delta^Q_k} \int_{G^Q_M} T^Q_k \log \left(\int_{G^P_M} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} dG^P_M\left(\delta^P_t \,\middle|\, \delta^Q_t, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) \right) \tag{172}$$
$$dG^Q_M\left(\delta^Q_t \,\middle|\, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) dP_{0,T}$$

Since multiplication is a convex operation, we can use Jensen's inequality.

$$\geq T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \left(\frac{1}{T} \sum_{\delta^Q_k} T^Q_k \right) \left(\int_{G^Q_M} \log \left(\int_{G^P_M} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} dG^P_M\left(\delta^P_t \,\middle|\, \delta^Q_t, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) \right) \right) \tag{173}$$
$$dG^Q_M\left(\delta^Q_t \,\middle|\, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) dP_{0,T}$$

The first sum is the sample size, and the second sum is the integral over all of the periods. Since we are only considering the supremum over all $t-1$ and each t can be in at most cluster with respect to Q, this mean does not change the value.

$$= T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \int_{G^Q_M} \log \left(\int_{G^P_M} \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} dG^P_M\left(\delta^P_t \,\middle|\, \delta^Q_t, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) \right) \tag{174}$$
$$dG^Q_M\left(\delta^Q_t \,\middle|\, \delta^Q_{t-1}, \delta^P_{t-1}, x_{t-1}, z_{t-1}\right) dP_{0,T}$$

We can use Jensen's inequality to pull the logarithm inside the second integral, and then combine the successive integrals into an integral with respect to the joint.

$$= T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \int_{G^Q_M\times G^P_M} \log \frac{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right)}{\phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right)} d\mathcal{K}\left(dG^Q_M(\delta^Q_t), dG^Q_M(\delta^Q_t) \,\middle|\, \mathcal{F}^Q_{t-1}\mathcal{F}^Q_{t-1}\right) dP_{0,T} \tag{175}$$

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

$$= T \int_{\mathbb{R}^{T\times D}} \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} dP_{0,T} \tag{176}$$

We can factor the $P_{0,T}$, and pull the supremum outside of the expectation, because we have finitely many terms.

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \int_{\mathbb{R}^D} \cdots \int_{R^D} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} \prod_{t=1}^{T} p_{0,T}\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right) dx_t \tag{177}$$

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.

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \int_{\mathbb{R}^D} \log \frac{\int_{G^Q_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^Q_t\right) dG^Q_M\left(\delta^Q_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)}{\int_{G^P_M} \phi\left(x_t \,\middle|\, x_{t-1}, \delta^P_t\right) dG^P_M\left(\delta^P_t \,\middle|\, \mathcal{F}^P_{t-1}\right)} p_{0,T}\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right) dx_t \tag{178}$$

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

$$= T \sup_{\mathcal{F}^Q_{t-1},\mathcal{F}^P_{t-1}} \mathrm{D}_{\mathrm{KL}}\left(q_T\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right) \,\middle\|\, p_{0,T}\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right)\right) \tag{179}$$

By Lemma 6 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.)

$$\sup_t h^2\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_{0,T}\left(x_t \,|\, \mathcal{F}_{t-1}^P\right)\right) \leq \sup_t \frac{1}{T} h^2(q_T(X), p_{0,T}(X)) \leq C\frac{T}{T}\epsilon^2 = C\epsilon^2 \tag{180}$$

□

Appendix C Contraction Rates

C.1 Constructing Exponentially Consistent Tests with Respect toh_∞

Lemma 1 (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_0 with $h_\infty(\xi_1, \xi_0)$:*

1. $$\Pr_T\left(\Upsilon_T \,|\, \xi_0\right) \leq \exp(-C_2 T\epsilon^2) \tag{24}$$

2. $$\sup_{\xi\in\Xi, e_n(\xi_1,\xi)<\epsilon C_3} \Pr_T\left(1-\Upsilon_T \,|\, \xi_0\right) \leq \exp(-C_2 T\epsilon^2) \tag{25}$$

Then the following two conditions hold with probability $1 - 2\delta_T$ with respect to the prior:

$$\sup_{\epsilon_T>\epsilon} \log N\left((\epsilon, \{\xi \in \Xi_T \,|\, h_\infty(\xi, \xi_0) \leq \epsilon\}, h_\infty\right) \leq T\epsilon_T^2 \tag{26}$$

and

$$\Pi_T\left(B_T(\xi_0, \epsilon_T, C_1) \,|\, X\right) \geq C\exp\left(-C_0 T\epsilon_T^2\right). \tag{27}$$

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. Since we are letting G_t differ every period, we can do this for both Q_T and $P_{0,T}$.

$$f\left(X \,|\, G_1, \ldots G_T\right) = \prod_{t=1}^T \int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f) \tag{181}$$

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.

$$h_{avg}^2\left(f\left(X \,\middle|\, \{G_t^f\}\right), g\left(X \,|\, \{G_t^g\}\right)\right) := \frac{1}{T}\sum_{t=1}^T h^2\left(\int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi\left(x_t \,|\, \delta_t^g\right) dG_t^g(\delta_t^g)\right) \tag{182}$$

Then by (Birgé 2013, Corollary 2), there exists a test ϕ_T that satisfies the following.[17]

17. To map his notation into ours, take his $z = 0$, and take his measure R equal to P. Eq. (183) is obvious then, and Eq. (184) follows by taking the exponential of both sides in the inequality inside the probability and rearranging.

$$\Pr_T\left(\phi_T(X)\,\middle|\,\left\{G_t^f, G_t^g\right\}\right) \leq \exp\left(-\frac{1}{3}Th_{avg}^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f), \int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right)\right) \tag{183}$$

$$\Pr_T\left(1-\phi_T(X)\,\middle|\,\left\{G_t^f, G_t^g\right\}\right) \leq \exp\left(-\frac{1}{3}Th_{avg}^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f), \int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right)\right) \tag{184}$$

Now, 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 by the following.

$$\frac{T}{3}\sup_{\{(G_t^f,G_t^g)\}} h_{avg}^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f), \int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right) \tag{185}$$

For any length T sequence, this equals the least favorable G_f^t and G_g^t repeated T times. This joint distribution exists in our set because we are not placing any restrictions on the dynamics besides ergodicity but a stationary distribution is clearly ergodic.

$$= \frac{T}{3}\frac{1}{T}\sum_{t=1}^{T} h^2\left(\int_{G_{sup}^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_{sup}^f(\delta_t^f), \int_{G_{sup}^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_{sup}^g(\delta_t^g)\right) \tag{186}$$

The terms inside the sum are all the same.

$$= \frac{T}{3}h^2\left(\int_{G_{sup}^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_{sup}^f(\delta_t^f), \int_{G_{sup}^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_{sup}^g(\delta_t^g)\right) \tag{187}$$

$$= \frac{T}{3}\sup_{(G_t^f,G_t^g)} h^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f), \int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right) \tag{188}$$

$$= \frac{T}{3}h_\infty^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f), \int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right) \tag{189}$$

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. Essentially, they both span the same sets of information.

$$= \frac{T}{3}h_\infty^2\left(\int_{G_t^f}\phi\left(x_t\,\middle|\,\delta_t^f\right)dG_t^f(\delta_t^f), \int_{G_t^g}\phi\left(x_t\,|\,\delta_t^g\right)dG_t^g(\delta_t^g)\right) \tag{190}$$

Since we can bound the error probabilities in both directions, using exponentially consistent

tests, we have shown that Item 1 holds. □

C.2 Bounding the Posterior Divergence

Proposition 6 (Bounding the Posterior Divergence)**.** *Let p_0 be a uniformly ergodic Hidden Markov Gaussian process, i.e. $p_0 \coloneqq \sum_k p_k \phi(\cdot \mid \mu_t, \Sigma_t)$ with finite mean and finite variance. Let $\Xi_T \subset \Xi$ and $T \to \infty$. Let the following condition hold with probability $1 - 2\delta$ for $\delta > 0$ and constants C and $n \in \mathbb{N}$*

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_{0,T}\left(x_t \,|\, \mathcal{F}_{t-1}^P\right)\right) < C\eta_T. \tag{28}$$

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

$$\sup_{\epsilon_{T,n} > \epsilon_n} \log N\left((\epsilon_n, \{\xi \in \Xi_T \mid h_\infty(\xi, \xi_0) \leq \epsilon_n\}, h_\infty\right) \leq T\epsilon_{T,n}^2, \tag{29}$$

and

$$\Pi_T\left(B_T\left(\xi_0, \epsilon_{T,n}, 2\right) \mid X\right) \geq C \exp\left(-C_0 T \epsilon_{T,n}^2\right). \tag{30}$$

Proof. We are looking at locally asymptotically normal models, as discussed in Lemma 6, 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 h_∞ 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 $\left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi, \xi_0) < T\epsilon_T^2\right\}$. 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 $\epsilon_T \to 0$ that satisfies the following two conditions.

$$\sup_{\epsilon_{T,n} > \epsilon_n} \log \phi\left((\epsilon_n, \{\xi \in \Xi_T \mid h_\infty(\xi, \xi_0) \leq \epsilon_n\}, h_\infty\right) \leq T\epsilon_{T,n}^2 \tag{191}$$

$$\Pi_T\left(\left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi, \xi_0) < \epsilon_{T,n}\right\}\right) \geq C \exp\left(-T\epsilon_{T,n}^2\right) \tag{192}$$

These two conditions work in opposite directions. The first criterion is easier to satisfy the larger η, but to get a fast rate of convergence we want a small η_T in the second condition.

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

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_{0,T}\left(x_t \,|\, \mathcal{F}_{t-1}^P\right)\right) < C\sqrt{\log\frac{1}{\delta}}\eta_T \tag{193}$$

Equation (192) is satisfied if $\eta_T^2 \geq \frac{C_0}{|\log\delta|} \exp\left(-T\epsilon_{T,n}^2\right)$.

$$\eta_T^2 = \exp\left(-T\frac{\log(T)^n}{T}\right) = \frac{1}{T^n} \tag{194}$$

We need h_∞^2 to 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 $K_T = \frac{\log(T)^n}{\eta_T^*}$ components.

$$K_T = \frac{\log(T)^n}{\eta_T^*} = \frac{\log(T)^n}{\log(T)^n / T^n} = T^n \tag{195}$$

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

$$N\left(\epsilon_n, \left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi, \xi_0)\right\} \leq \epsilon_n, h_\infty^2\right) \leq T^n = \exp\left(\log\left(T^n\right)\right) = \exp\left(T\epsilon_{T,n}^2\right) \tag{196}$$

Taking logarithms of both sides of Eq. (196) finishes the proof.

□

C.3 Bounding the Hellinger Distance and Kullback-Leibler Divergence

Lemma 7 (van der Vaart and van Zanten (2008) Lemma 3.1)**.** *For any measurable functions $v, w : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a metric space we have the following:*

1. $h(p_v, p_w) \leq \|v - w\|_\infty \exp\left(\frac{1}{2}\|v - w\|_\infty\right)$
2. $\mathrm{D}_{\mathrm{KL}}\left(p_v \,||\, p_w\right) \leq \|v - w\|_\infty^2 \exp\left(\|v - w\|_\infty\right)\left(1 + \|v - w\|_\infty\right)$
3. $V_{2,0}(p_v, p_w) \leq \|v - w\|_\infty^2 \exp\left(\|v - w\|_\infty\right)\left(1 + \|v - w\|_\infty\right)^2$

In particular, we can set $\mathcal{X} = \mathbb{R}^{T \times D}$, and then the statement relates bounds with respect to the sup-norm of X to the other divergence measures.

C.4 Contraction Rate of the Marginal Density

Theorem 8 (Contraction Rate of the Marginal Density)**.** *Let p_0 be a uniformly ergodic Hidden Markov Gaussian process, i.e. $p_0 \coloneqq \sum_k p_k \phi\left(\cdot \,|\, \mu_t, \Sigma_t\right)$ with finite mean and finite variance. Let $T \to \infty$, then the following holds with $\epsilon_T = \sqrt{\frac{\log(T)}{T}}$ 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*

$$P_0\left(\Pi_T\left(h\left(p_{0,T}\left(x_t\right), q_T\left(x_t\right)\right) \geq C\epsilon_T \,|\, X\right)\right) \to 0. \tag{32}$$

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). Again, through a sampling argument, 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 $n = 1$ in Eq. (28), Theorem 3 implies the necessary bound on the sieve complexity exists. In addition, since h_∞ is bounded above by the normal 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.5 Contraction Rate of the Transition Density

Theorem 7 (Contraction Rate of the Transition Density)**.** *Let p_0 be a uniformly ergodic Hidden Markov Gaussian process, i.e. $p_0 \coloneqq \sum_k \pi_{t,k}\phi\left(\cdot \mid \mu_t, \Sigma_t\right)$ with finite mean and finite variance. Let $T \to \infty$, then the following holds with $\epsilon_T = \sqrt{\frac{\log(T)^2}{T}}$ 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*

$$P_0\left(\Pi_T\left(\sup_{\mathcal{F}^P_{t-1},\mathcal{F}^Q_{t-1}} h\left(p_{0,T}\left(x_t \mid \mathcal{F}^P_{t-1}\right), q_T\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)\right) \geq C\epsilon_T \,\middle|\, X_T\right)\right) \to 0. \tag{31}$$

Proof. The proof of this is essentially identical to the marginal density case mutatis mutandis. Lemma 1 implies the that h_∞ has the required exponentially consistent tests.

Having done that we can verify the conditions in Proposition 6. If we take $n = 2$ in Eq. (28), Theorem 4 implies the necessary bound on the sieve complexity exists.

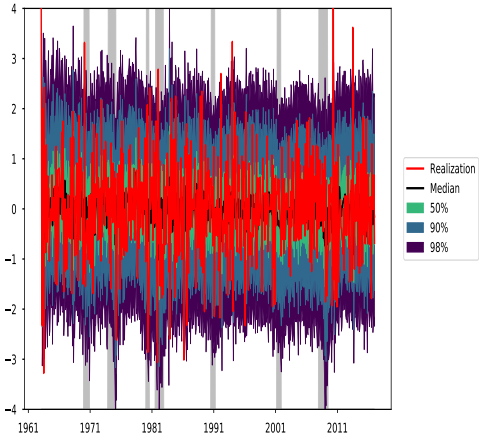
This verifies the three conditions in Theorem 5 on a set with with probability $1-2\delta$ with respect to the prior. This then gives us the posterior contraction rate $\epsilon_T = \sqrt{\frac{\log(T)^2}{T}}$.

□

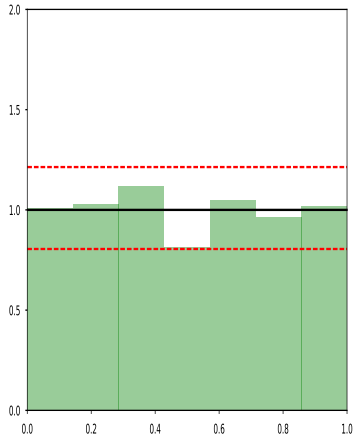
Appendix D Macroeconomic Empirical Results

Figure 12: Housing Supply

(a) Posterior Density



(b) PIT Histogram



(c) PIT ACF

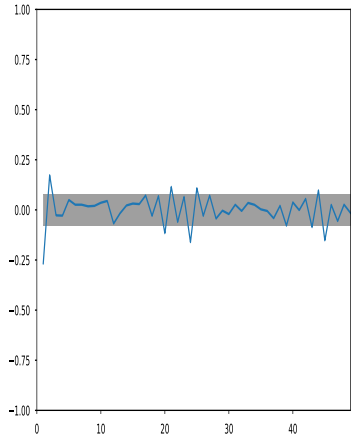


Figure 13: Industrial Production

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

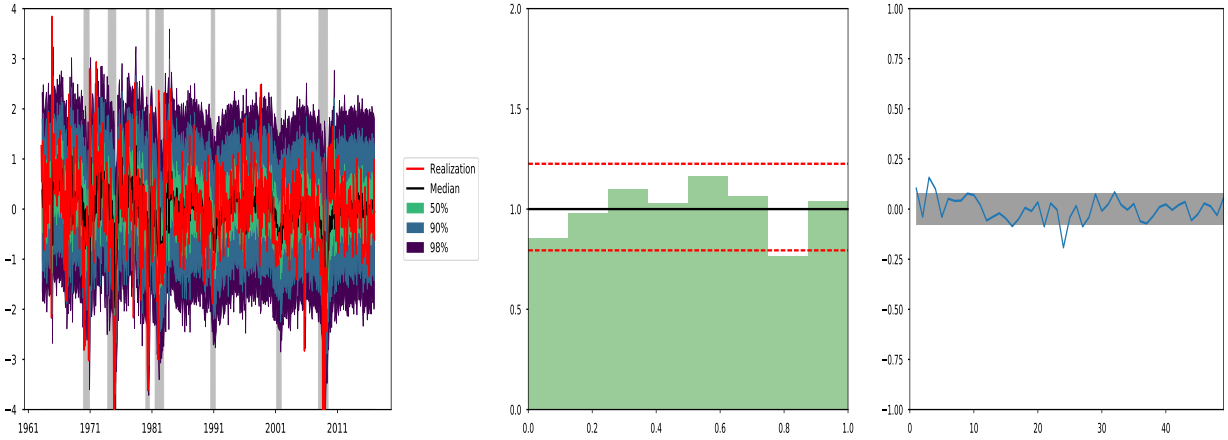
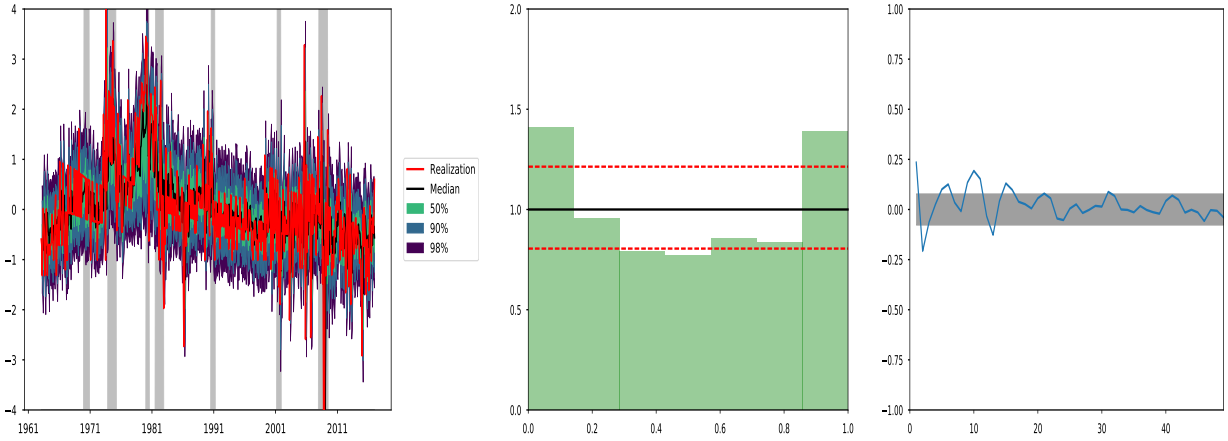


Figure 14: CPI gInflation

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



Appendix E Financial Empirical Results

Figure 15: Long-Term Interest Rate

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

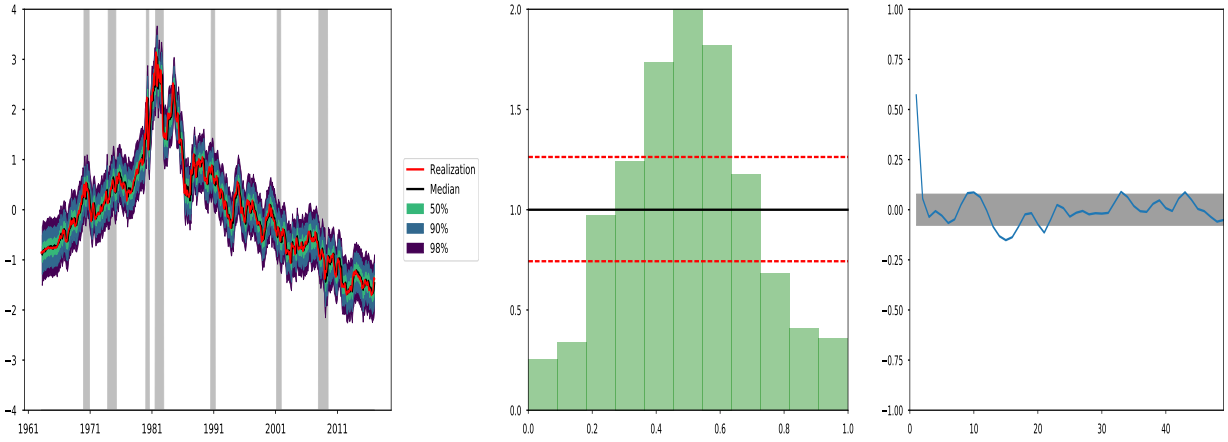


Figure 16: M2

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

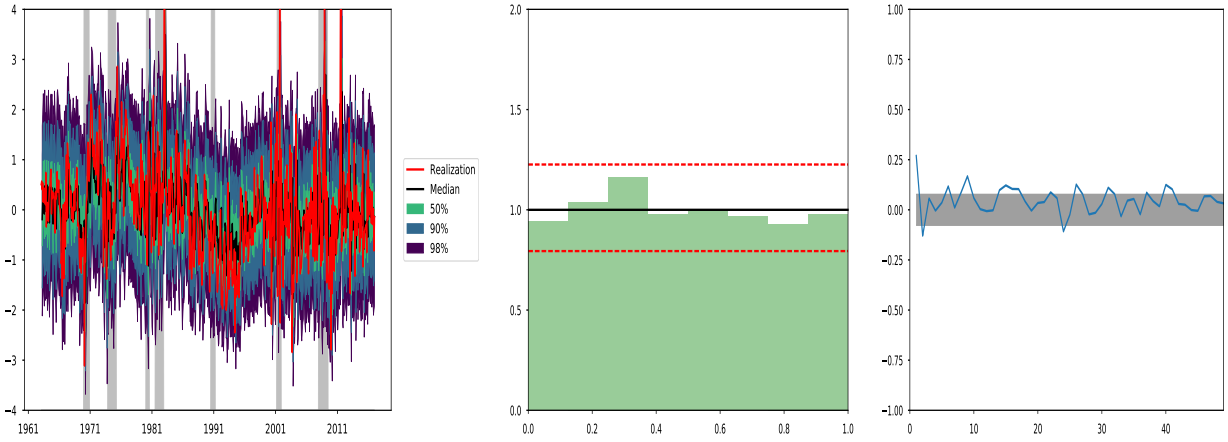


Figure 17: Personal Consumption Expenditures (PCE) Inflation

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

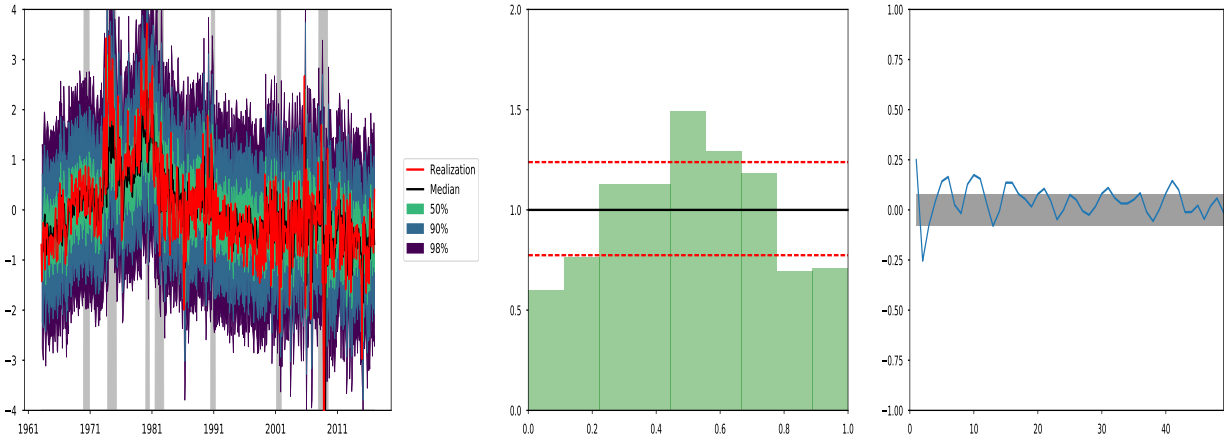


Figure 18: GOLD

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

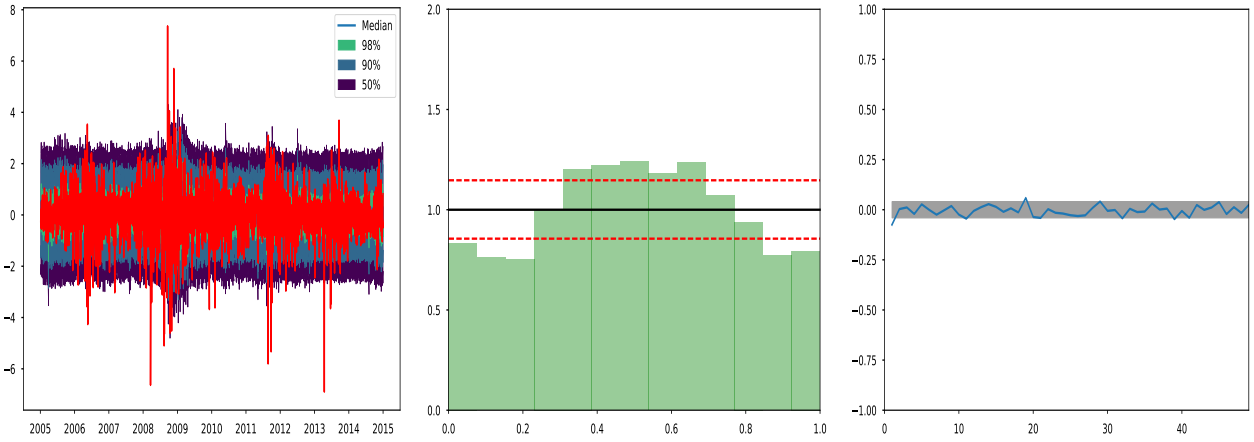


Figure 19: VFH Volatility

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

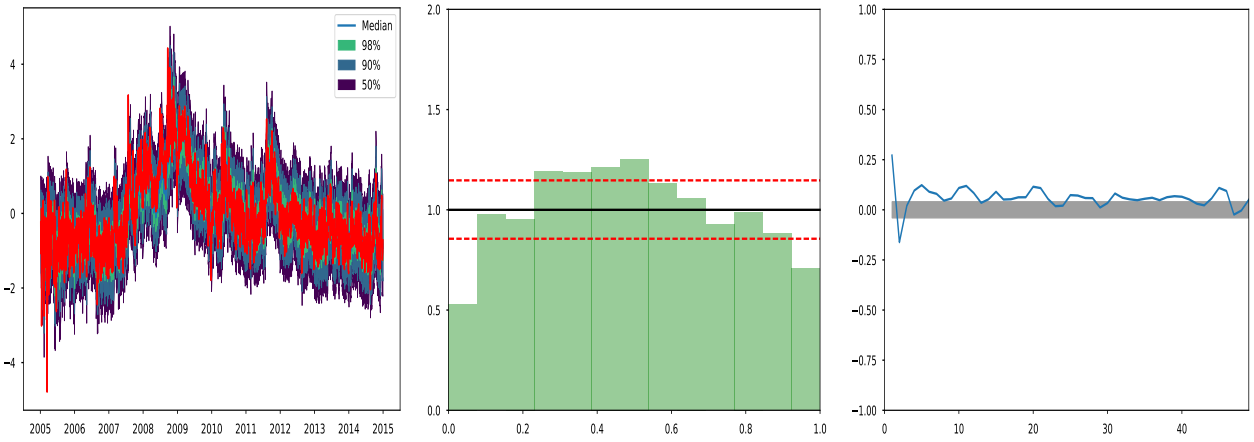


Figure 20: VIX

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

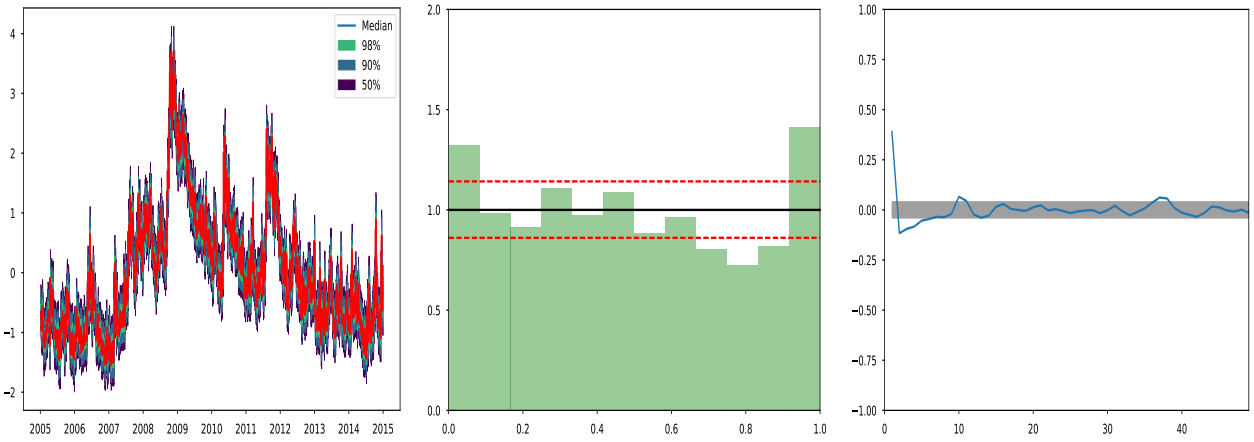
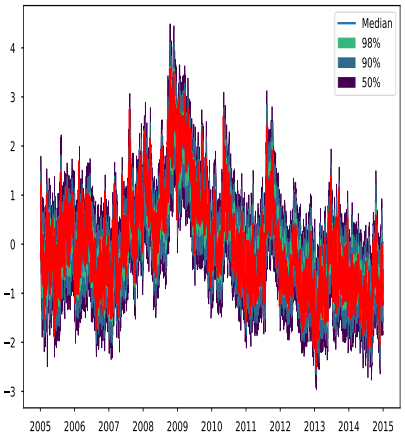
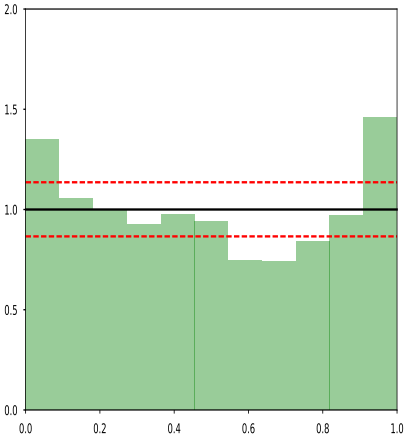


Figure 21: VNQ Volatility

(a) Posterior Density



(b) PIT Histogram



(c) The PIT ACF

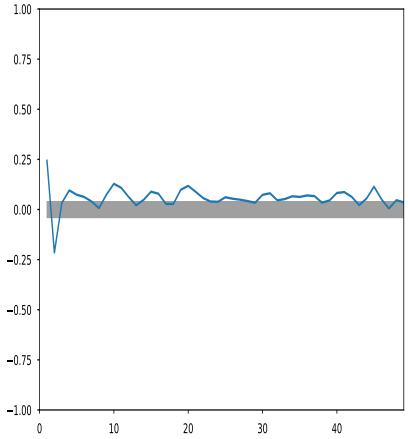
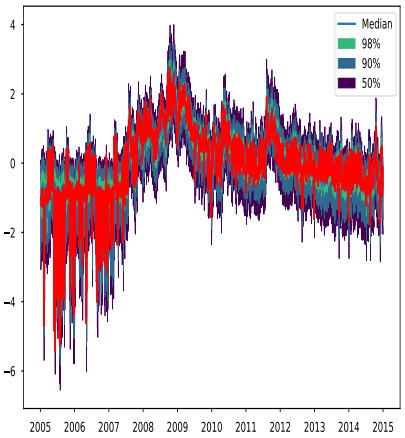
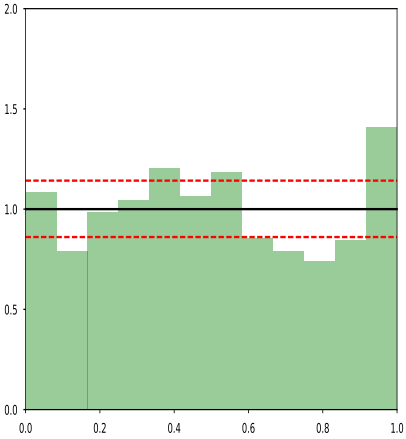


Figure 22: XLF Volatility

(a) Posterior Density



(b) PIT Histogram



(c) The PIT ACF