Feasible Multivariate Density Estimation Using Random Compression[*]

Minsu Chang [†] Paul Sangrey [‡]

December 6, 2020

Abstract

Given vector-valued data $x_t \in \mathbb{R}^D$ for $t = 1, \dots, T$, nonparametric density estimators typically converge slowly when the number of series D is large. We extend ideas from the random compression literature to nonparametric density estimation, constructing an estimator that, with high probability, converges rapidly even when applied to a large, fixed number of series. We devise a discrete random operator to compress the data so that the density of the compressed data can be represented as a parsimonious mixture of Gaussians. We show that this mixture representation closely approximates the true distribution. Then we provide a computationally efficient Gibbs sampler to construct our Bayesian density estimator using Dirichlet mixture models. We estimate both marginal and transition densities for both i.i.d. and Markov data. With high probability with respect to the randomness of the compression, our estimators' convergence rate — $\sqrt{\log(T)}/\sqrt{T}$ — depends on D only through the constant term. Our procedure produces a well-calibrated joint predictive density for a macroeconomic panel.

Keywords: Bayesian nonparametrics, curse of dimensionality, density forecasting, infinite Gaussian mixtures, random compression

[*] We are indebted to Francis X. Diebold, Jesús Fernández-Villaverde, and Frank Schorfheide. We have also benefited greatly from conversations with Karun Adusumilli, Mike Bedard, Ben Connault, Frank DiTraglia, Domenico Giannone, Ivana Komunjer, Laura Liu, Andriy Norets, Davide Pettenuzzo, Alexandre Poirier, Allan Timmermann, our colleagues at Amazon and Georgetown University, and seminar participants at Amazon, the University of Pennsylvania, the 2018 NBER-NSF SBIES conference at Stanford University, the 2019 Canadian Econometric Study Group in Montreal, and the 2019 Midwest Econometrics Group at Ohio State University. All remaining errors are our own.

[†] Georgetown University, Email: minsu.chang@georgetown.edu, Web: minsuchang.com

[‡] Amazon, Email: paul@sangrey.io, Web: 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 $x_t \in \mathbb{R}^D$ for $t = 1, \dots, T$, nonparametric estimators for the data's joint or conditional density converge very slowly when the number of series D is large. Given that we cannot avoid this *curse of dimensionality*, we construct a nonparametric density estimator that, with high probability, converges faster than the minimax rates for a large but fixed number of series by extending ideas from the random compression literature. This paper will be useful for practitioners, especially those who do not want a priori to choose which parametric model to use, because the procedure produces parsimonious nonparametric density estimates for a large number of time series with a computationally efficient algorithm.

Various authors have studied random compression which transforms the high-dimensional data to a much lower dimensional space while approximately preserving the distances between data points with high probability (e.g., Johnson and Lindenstrauss (1984); Klartag and Mendelson (2005); Boucheron et al. (2013); Talagrand (2014), Koop et al. (2019)). Based on this idea, we devise a discrete random compression operator that induces the compressed data's distribution to be close to the data's true distribution. To the best of our knowledge, this paper is the first to extend this idea of randomly compressing the data to the space of densities. An advantage of our random operator is that the compressed data's distribution can be represented as a parsimonious mixture of Gaussians. We then develop a computationally efficient estimator for this representation. Given $\delta > 0$, high probability $1 - 2\delta$ with respect to the random compression, our estimators converge rapidly and their convergence rate — $\sqrt{\log(T)}/\sqrt{T}$ — depends on D only through the constant term. In contrast, minimax rates bound the worst-case behavior of the estimator. By only requiring the estimators converge at the given rate with probability $1 - 2\delta$ with respect to randomness in the compression instead of with probability 1 (as minimax rates do), substantially faster rates are obtained.[1]

We consider a data generating process for a sequence of conditional densities $p(x_t|\mathcal{F}_{t-1})$ given

[1] Building upon existing probabilistic guarantees in the data compression literature, we show that fast rates of density estimation can be recovered, at least probabilistically.

filtrations $\mathcal{F}_{t-1}$ for $t = 1, \ldots, T$. We assume that each of the $p(x_t|\mathcal{F}_{t-1})$ is an infinite Gaussian mixture.[2] We use random compression that operates on the length of the time series T to construct an approximating distribution in a series of steps. First, we construct a discrete random operator that clusters each vector x_t into a Gaussian component of the mixture representation and also endogenously determines the number of mixture components. This random operator is a compression device that approximately preserves the norms of the individual data points that are originally from an infinite Gaussian mixture. The induced distribution after compression has a Gaussian mixture representation whose mixture probabilities do not depend on the data. In this regard, our random compression is data-agnostic.[3] We then show that, with high probability with respect to this data-agnostic measure, the number of mixture components only depends logarithmically on T.

Second, we show that this approximating mixture is close to the data's true density with high probability. To build intuition, suppose the data $\{x_t\}_{t=1}^T$ are from the mean-zero normal distribution with covariance $\{\Sigma_t\}_{t=1}^T$. The value of the probability density function evaluated at x_t is determined by the data's weighted norm, $x_t'\Sigma_t^{-1}x_t$. If a random compression operator preserves this norm of the data for all x_t, the induced density after compression will be close to the true density of x_t. Generalizing this intuition, we show uniform bounds on the local divergence between the norms of the data points imply bounds on the divergence between the densities such as Hellinger and Kullback-Leibler. We devise a random compression operator that preserves the relevant norms of the data and we show the induced Gaussian mixture representation closely approximates the data's true density as well.

We then relate our approximating Gaussian mixture representation to a Gaussian mixture whose latent mixing measure is Dirichlet. This lets us develop a Gibbs sampler based on Dirichlet mixture models to estimate marginal and transition densities for both i.i.d. and Markov data. Our principal contribution in this part lies in placing the Dirichlet process prior on the mixture identities' transitions instead of placing the prior period-by-period. We then adapt state-of-the-art computational samplers for the static case to this dynamic case. The resulting

[2] This is a very general assumption that, for example, is implied by the existence of any absolute moment (Tokdar (2006)). In particular, DGPs we study do not need to be supersmooth as assumed in previous papers that find similar rates of convergence, e.g.,Ghosal and van der Vaart (2007b).

[3] Each element of our random compression matrix is drawn in a way that does not depend upon the data. Several papers such as Achlioptas (2003) and Guhaniyogi and Dunson (2015) study random compression in this data oblivious way.

sampler only takes a few minutes on a standard computer to make thousands of draws from the posterior.

This paper is closely related to the literature using mixture distributions to estimate unconditional densities (Ghosal and van der Vaart, 2017). It provides a dynamic generalization of the infinite-mixture representation. This paper complements alternative methods to construct Bayesian conditional density estimators (e.g., Geweke and Keane (2007); Norets (2010); Pati et al. (2013)). Furthermore, in contrast to our random compression approach, much of the nonparametric literature indexes the functions it approximates by some smoothness class. These papers show that requiring the estimators to be consistent forces the estimator and the approximation, which is deterministic, to use the same number of terms asymptotically. For example, Stone (1980)'s minimax estimation procedure constructs a deterministic approximation that requires $T^{g(D)}$ terms for some g that depends upon the smoothness class under consideration (Yang and Barron (1999); Ichimura and Todd (2007)).

Instead, we construct a bound for the number of mixture components as a function of T that holds with high probability. This probability is with respect to the data-agnostic random compression procedure that determines the number of mixture components, similar to the results in the random compression literature. In particular, we consider an asymptotic experiment where D is medium to large but fixed and T grows. We then convert the bound on the number of mixture components into convergence rates for the density estimators. With high probability, our estimators' convergence rate — $\sqrt{\log(T)}/\sqrt{T}$ for both the marginal density and the transition density — depend on D only through the constant term, instead of decaying exponentially fast in D as minimax rates do. Even though we cannot beat the minimax rate in general (that is impossible), we show our estimators usually perform well even when D is large. We only need to tolerate an arbitrarily small probability of the estimators converging slowly. In particular, we show that the distance between the induced mixture representation and the data's true distribution is small even when we take the supremum over the set of true data generating processes and D is a large constant.[4]

We organize the paper as follows. Section 2 describes the data generating process. Section 3 constructs the sieve and provides conditions under which it approximates the true density well.

[4] We show our estimators converge rapidly with high probability. We make no claim that these are the only estimators that converge rapidly with high probability.

Section 4 proves our estimators converge at the rate given above with high probability. Section 5 provides a computationally efficient Gibbs sampling algorithm to estimate our sieve. Section 6 analyzes the model's performance in a simulation with Student's t-distributed shocks. Section 7 empirically analyzes a monthly macroeconomic panel showing our method works well in practice. Section 8 concludes. The appendices contain the proofs.

2 Data Generating Process

We now specify the set of data generating processes (DGPs) that we allow.

Definition 1 (Data Generating Process)**.** The data $X_T \in \mathbb{R}^{T \times D}$ is obtained by stacking the vector $x_t \in \mathbb{R}^D$ over $t = 1, \ldots, T$. $X_T's$ conditional densities given filtration $\mathcal{F}_{t-1}$ for each time period are

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

where $\Pi^p_{t-1,k}$ is the mixture probability of the k^{th} component and $\phi\left(x_t \,|\, x_{t-1}\beta_{k,t}, \Sigma_{k,t}\right)$ is the probability density function of normal distribution having mean $x_{t-1}\beta_{k,t}$ and covariance $\Sigma_{k,t}$. The $\Pi^p_{t-1,k}$ must be a valid Markov transition matrices whose entries are nonnegative real numbers.

The data's conditional densities — $p_T(x_t \,|\, \mathcal{F}_{t-1})$ — have an infinite Gaussian mixture representation for each time period. Each mixture component has an associated mixture probability, $\Pi^p_{t-1,k}$ and component-specific parameters, $\beta_{k,t}$, and $\Sigma_{k,t}$. We let the true DGP depend upon T because at this point we are only approximating the density for a fixed T.

We now define the approximating model in Definition 2 that is a Gaussian mixture with K_T components. The number of components K_T governs the complexity of the model and so grows with T. Dirichlet mixture models have been used in a wide variety of environments. For example, Fox (2009) studies similar Markov processes. Our contribution does not lie in the modeling front, but rather in relating Dirichlet processes to a discrete random compression operator that induces the mixture representation as in Definition 2.

Definition 2 (Approximating Model)**.**

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

We use the terms mixture and cluster interchangeably. Each cluster's (mixture's) components, (β_k, Σ_k), no longer have time t subscripts. The idea is that we can reuse the latent variables $(\beta_{k,t}, \Sigma_{k,t})$ across time. Since the clusters are defined differently in Definition 1 and Definition 2, no simple relationship between the parameters exists. This is also the case with the mixture probabilities. We use the notation $\Pi^p_{t-1,k}$ and $\Pi^q_{t-1,k}$ above to highlight that they are different objects. In the paper, we drop the superscripts p and q when it is clear what object we refer to. Throughout, we use $\mu_T = E[X_T]$ to refer to the TD-dimensional mean vector. We also consider the rescaled data that lie on the unit hypersphere:

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

where $\|\cdot\|_{L_2}$ is the L_2-norm. Since $\widetilde{X}_T$ is on the unit hypersphere, it is a compact space for any fixed T. Since $X_T - \mu_T$ is a zero-mean Gaussian conditional on a mixture component, its $TD \times TD$ covariance matrix completely determines its component-wise distributions. We define the densities of $\widetilde{X}_T$ as we did for X_T above and denote them $\widetilde{p}_T$ and $\widetilde{q}_T$.

We impose the following assumptions to derive our results:

Assumption 1. *Assume the conditional densities $p_T(x_t \,|\, \mathcal{F}_{t-1})$ given filtration $\mathcal{F}_{t-1}$ can be represented as infinite Gaussian mixtures for all $t = 1, \ldots, T$ as in Equation (1). Further assume that the $x_t \in \mathbb{R}^D$ have uniformly bounded means μ_t and covariances Σ_t where the Σ_t are positive-definite. That is, $\sup_{t\geq 1}\|\mu_t\|_{L_1} < C_1 < \infty$ and the minimum and the maximum eigenvalues of Σ_t, denoted by $\lambda_{\min}(\Sigma_t)$ and $\lambda_{\max}(\Sigma_t)$, satisfy $0 < \lambda_{\min}(\Sigma_t)$ and $\lambda_{\max}(\Sigma_t) < C_2 < \infty$ for some constants C_1 and C_2.*

Because Assumption 1 allows for infinitely many components and does not uniformly bound the variances from below, it is a very general assumption. For example, Tokdar (2006) shows that any density with a finite absolute moment could be approximated by a mixture of normals.[5]

[5] In other words, if there exists an $\eta > 0$ where the true density p_0 satisfies $\int |x|^\eta \, dP_0(x) < \infty$, then we can represent p_0 as an infinite Gaussian mixture as in the first part of Assumption 1.

Very few interesting densities do not have any absolute moments. Also, this does not place restrictions on the smoothness class of the distributions. For instance, non-differentiable densities can still have finite moments. It is a very weak tail condition.

Now, Assumption 1 does not impose any structure on the relationship between the $p(x_t \mid \mathcal{F}_{t-1})$ over different time periods. The positive-definite assumption rules out perfect correlation between the various components in the vector x_t. Our results on the transition densities require the data are sufficiently regular across time.

Assumption 2. *There exists a latent state $z_t \in \mathbb{R}^N$ such that $w_t := (x_t', z_t')'$ is a countably-generated geometrically ergodic Markov chain.*[6]

Note, if the x_t form a Markov sequence, then this holds automatically; we can take z_t to be a constant. In the following sections, we sometimes specialize to the case where the x_t are independent across t.

3 Sieve Construction

3.1 Setting up the Problem

We construct a sieve that approximates the wide variety of DGPs our paper considers. Given the rescaled data $\widetilde{X}_T$, $\epsilon > 0$, and $\delta \in (0, 1/2)$, we construct a discrete random operator that takes a TD-dimensional hypersphere and maps it onto a KD-dimensional hypersphere, where $K \ll T$. We show in Theorem 1 that this mapping only perturbs the norms of the individual elements in $\widetilde{X}_T$ by at most ϵ with high probability $1 - 2\delta$.

We then show the joint densities across $t = 1, \ldots, T$ are also not perturbed significantly in Theorem 2. This result holds whenever the value of the joint density function evaluated at a datapoint is determined by the data's norm. Since our random compression operator does not perturb the norms of the individual data points significantly, the induced density after compression is still close to the data's true density. In other words, we link bounds on the divergences between the norms to bounds on the divergences between the densities.

[6] For a sequence $w \in \mathcal{W}$, let P denote the associated Markov kernel, π denote the associated stationary distribution, and $\|\cdot\|_{TV}$ denote the total variation norm. Then $P(w, \cdot)$ is a geometrically ergodic Markov chain if for π-almost-everywhere $w \in \mathcal{W}$ there exists constants $\rho_w < 1$ and $C_w < \infty$ such that $\|P^n(w, \cdot) - \pi(\cdot)\|_{TV} < C_w \rho_w^n$ for $n \in \mathbb{N}$.

Lastly, we approximate both the marginal density of X_T in the space of densities over $\mathbb{R}^D$ and its conditional/transition density, which lies in the associated product space. Note, we are interested in X_T's marginal and transition density, not $\widetilde{X}_T's$. We show that the difference between the marginal densities is $1/T$ times the difference between the joint densities by exploiting the product form of joint densities of independent or Markov data. Extending this argument, we further show that the difference between the joint densities can be used to bound the difference between the transition densities.

3.2 Bounding the Norm Perturbation

We compress the data with a random operator so that the induced distribution has a mixture distribution. A mixture distribution can be treated as a random clustering of the data where the data in each cluster has the same parametric distribution. For example, one can cluster the data with K bins using a $T \times K$ discretization operator where each row of the operator contains exactly one 1 and the rest of the elements equal zero. A variable x_t is in bin k if and only if the operator has a 1 in row t and column k.

We first show that when the operator satisfies certain conditions, it preserves the norms of the datapoints. This will be used to bound the divergence between densities.

Theorem 1 (Bounding the Norm Perturbation)**.** *Let $\widetilde{X}_T$ be in the unit hypersphere in $\mathbb{R}^{TD-1}$. Let $\epsilon > 0$ and $0 < \delta < 1/2$ be given. Construct Θ_T as a $T \times K_T$ operator comprised of each element $\theta_{t,k}$ taking a value from {-1, 0, 1} such that 1) the rows of Θ_T are i.i.d., 2) the columns of Θ_T form a martingale difference sequence, and 3) the number of columns K_T of Θ_T satisfies $K_T > \max\left\{\frac{\log(1/\delta)}{C_1\epsilon^2}, \frac{\log(T)}{\epsilon^2}\right\}$ for some universal constant C_1 with arbitrarily high probability. Then with probability greater than $1 - 2\delta$ with respect to the randomness in Θ_T, there exists a universal constant C_2 such that*

$$\sup_t \left| \frac{1}{K_T} \sum_{k=1}^{K_T} \left(\theta_{k,t} \sum_{d=1}^{D} x_{t,d} \right)^2 - \|x_t\|_{L_2}^2 \right| < C_2 \left(1 + \log\left(\frac{1}{\delta}\right)\right) \epsilon.$$

Theorem 1 implies that when the number K_T of Θ_T's columns grows logarithmically with T, applying Θ_T perturbs the norms of $\widetilde{x}_t$ by at most ϵ. This result holds with probability at least $1 - 2\delta$ with respect to the distribution over Θ_T. Since $\widetilde{X}_T \in \mathbb{R}^{TD-1}$, we can map $\mathbb{R}^{TD-1}$

onto a smaller space $\mathbb{R}^{K_T D-1}$, with $K_T \ll T$, without perturbing the individual elements' norms significantly. This does not affect the mean or the variance. This increased randomness induced by Θ_T "smooths" the data.

Given a fixed $\epsilon > 0$, a smaller value of δ makes the lower bound on K_T larger. Hence, Θ_T must have more columns, i.e. clusters. Furthermore, the right hand side of the last inequality in Theorem 1 is larger and the probability $1 - 2\delta$ is larger with a smaller value of δ. Conversely, given a fixed $\delta > 0$, a smaller value of ϵ makes the lower bound on K_T larger but the right hand side of the last inequality smaller. To summarize, the result's dependence on tolerances ϵ and δ leads to trade-offs between the number of clusters we need and the tightness of the last inequality.

We now construct an explicit Θ_T operator that satisfies the conditions in Theorem 1. This Θ_T operator differs in two ways from a standard discretization operator to make the columns of Θ_T be a martingale difference sequence. First, we let $\theta_{t,k}$ take on values from $\{-1, 0, 1\}$. Each x_t is in bin k if $\theta_{t,k} = 1$ and in bin $K_T + k$ if $\theta_{t,k} = -1$. If Θ_T has K_T columns, there are $2K_T$ possible clusters given that $\theta_{t,k}$ could be 1 or -1. Second, we let each row of Θ_T have as many 1's and -1's as necessary. Then realizing 1 in column k does not change the distribution of columns $k + 1$ through K_T. In a standard discretization operator, once 1 realizes the remaining columns in the row equal zero. This property makes the columns too dependent to form a martingale difference sequence.

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

The Θ_T operator satisfies $\mathbb{E}[\theta_{t,k}] = 0$ and $\mathbb{V}\text{ar}(\theta_{t,k}) = \mathbb{E}[|\theta_{t,k}|] = \Pr(\zeta = 1)$. Since each row of Θ_T can have multiple nonzero elements, each datapoint may be in multiple components simultaneously. In other words, we do not just create a mixture distribution across periods but also create one in each period. In addition, Θ_T is independent of $\widetilde{X}_T$. Since Θ_T is discrete, Θ_T with K_T columns implicitly clusters $\widetilde{X}_T$ with $2K_T$ possible clusters.

To use Θ_T in Theorem 1, it must satisfy the relevant conditions. Clearly, $\theta_{t,k} \in \{-1, 0, 1\}$. Also, Θ_T's rows are independent and its columns form a martingale difference sequence. The only dependence between the columns of Θ_T arises through the stopping rule, and stopped martingales are still martingales. Theorem 1 requires $K_T > \max\left\{\frac{\log(1/\delta)}{C_1\epsilon^2}, \frac{\log T}{\epsilon^2}\right\}$ holds with arbitrarily high probability. Given fixed ϵ, δ the first number is a constant, so we need to satisfy $K_T > \log(T)/\epsilon^2$ as T increases. By setting $b = \Pr(\theta_{t,k} \neq 0)$ appropriately with respect to ϵ, we could satisfy this lower bound.[7] To wrap up, we constructed an operator Θ_T that we rely on extensively in the remainder of the paper. In particular, we relate this Θ_T operator to the Dirichlet process. This operator compresses the data by clustering $\widetilde{X}_T$. Since the number of clusters K_T grows logarithmically in T, this compression substantially reduces the complexity in contrast to considering each of the T values of x_t separately.

3.3 Distances on the Space of Densities

In the previous section, we showed that Θ_T does not perturb the rescaled data $\widetilde{x}_t$'s norms significantly. Our goal is to convert bounds on the sequence of norms of the data into bounds on the densities, which requires us to decide on which distances to use on the space of densities. We use the Hellinger distance and the supremum Hellinger distance.

Definition 4. (Hellinger Distance).

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

Definition 5 (Supremum Hellinger Distance)**.**

$$h^2_\infty(p, q) := \sup_{\mathcal{F}^P_{t-1}, \mathcal{F}^Q_{t-1}, 1 \leq t \leq T} h^2\left(p\left(x_t \,\middle|\, \mathcal{F}^P_{t-1}\right), q\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)\right).$$

The Hellinger distance is a valid norm on the space of densities. Instead of applying this to joint densities, we take the supremum over the conditional densities. The supremum Hellinger distance will prove useful because it is stronger than both the Hellinger distance and the

[7] Lemma 3, which is proven in the online appendix, shows that $\frac{C_1 \log(T)}{|\log(1-b)|} \leq K_T \leq \frac{C_2 \log(T)}{|\log(1-b)|}$ for some constants C_1, C_2. If we choose $b = 1 - \exp(-C_1\epsilon^2))$, the lower bound of K_T in Theorem 1 is satisfied.

Kullback-Leibler divergence applied to joint densities. As a consequence, once we bound h_∞, we can directly deduce other bounds as necessary.

3.4 Representing the Joint Density

We now show that the approximating joint density of the rescaled data induced by Θ_T is close to the true joint density using h_∞. Since Θ_T is discrete, the approximating density denoted by $\widetilde{q}_T$ is a mixture of Gaussians. We rely on the properties of latent mixing measures to show that two distributions $\widetilde{q}_T$ and $\widetilde{p}_T$ are close, even though they likely have different mixture components. We do this by using techniques for controlling the behavior of latent random measures similar to those developed in the literature, for example Nguyen (2016) and Guha et al. (2020).

We represent a mixture of Gaussians as an integral with respect to a latent mixing measure — G_t^Q — for each period t. The parameters in each component are means and covariances, and so the G_t^Q measure is over the space of means and covariances. Let G^Q be the latent mixing measure over the space of G_t^Q. That is, each G_t^Q is a draw from G^Q. Let δ_t^Q denote the mixture identity induced by Θ_T. Let $\phi(\cdot \mid \delta_t^Q)$ denote the multivariate Gaussian density with the mean and the covariance associated with the δ_t^Q mixture. Then $\widetilde{q}_T$ can be expressed as

$$\widetilde{q}_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\, dG^Q(dG_t^Q). \tag{4}$$

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

$$\widetilde{p}_T(\widetilde{\mathcal{X}}) = \int_G \int_{G_t} \phi\left(\widetilde{x}_t \mid \delta_t^P\right) dG_t^P(\delta_t^P)\, dG^P\left(dG_t^P\right), \tag{5}$$

with its associated latent mixing measures and mixture identities. Note the approximating cluster identities $\{\delta_t^Q\}_{t=1}^T$ are different from the true cluster identities $\{\delta_t^P\}_{t=1}^T$, because Θ_T induces Q's clustering. It is not induced by the underlying true clustering.[8]

Theorem 1 shows that Θ_T does not perturb the norms of the rescaled data significantly with high probability. Hence, under the context of a mixture of Gaussians, $(\widetilde{x}_t - \widetilde{\mu}_t^P)'(\widetilde{\Sigma}_t^P)^{-1}(\widetilde{x}_t - \widetilde{\mu}_t^P)$ and $(\widetilde{x}_t - \widetilde{\mu}_t^Q)'(\widetilde{\Sigma}_t^Q)^{-1}(\widetilde{x}_t - \widetilde{\mu}_t^Q)$ are close with $\widetilde{\mu}_t^P := \mathbb{E}[\widetilde{x}_t \mid \delta_t^P]$, $\widetilde{\Sigma}_t^P := \mathbb{C}\text{ov}[\widetilde{x}_t \mid \delta_t^P]$, $\widetilde{\mu}_t^Q = \mathbb{E}[\widetilde{x}_t \mid \delta_t^Q]$,

[8] Equations (4) and (5) are immediate consequences of Definition 1 and Definition 2 applied to the rescaled data because we can create hierarchies of the G_t by expanding the probability space.

and $\widetilde{\Sigma}_t^Q = \mathbb{C}\mathrm{ov}[\widetilde{x}_t \,|\, \delta_t^Q]$. Since this holds for all $\widetilde{x}_t$ and $(\widetilde{x}_t - \widetilde{\mu}_t^P)'(\widetilde{\Sigma}_t^P)^{-1}(\widetilde{x}_t - \widetilde{\mu}_t^P)$ determines the value of $\phi\left(\widetilde{x}_t \,\middle|\, \delta_t^P\right)$, the densities $\widetilde{p}_T$ and $\widetilde{q}_T$ are close. This lets us convert bounds on divergences in the space of $\widetilde{X}_T$ — $\widetilde{\mathcal{X}}$ — to bound on divergences in the space of densities.

Theorem 2 (Representing the Joint Density)**.** *Let $\widetilde{X}_T \coloneqq \frac{X_T - \mu_T}{\|X_T - \mu_T\|_{L_2}}$ where X_T satisfies Assumption 1. Let $\epsilon > 0$ and $\delta \in (0, 1/2)$ be given. Construct the random operator Θ_T using the procedure in Definition 3. Let $\widetilde{p}_T$ denote the density of $\widetilde{X}_T$. Then the approximating density $\widetilde{q}_T$, which is the mixture of Gaussians over $\widetilde{\mathcal{X}}$ that Θ_T induces, satisfies the following with probability at least $1 - 2\delta$ with respect to Θ_T for some universal constant C:*

$$h_\infty^2\left(\widetilde{p}_T(\widetilde{\mathcal{X}}), \widetilde{q}_T(\widetilde{\mathcal{X}})\right) < C\left(1 + \log\left(\frac{1}{\delta}\right)\right)^2 \epsilon^2.$$

We represent the joint density as follows. Unlike previous compression operators in the literature, Θ_T is discrete; hence, it clusters $\widetilde{x}_t$. This property implies that the density of $\widetilde{x}_t$ is a process with respect to a discrete measure. That is, $\widetilde{Q}_T$ is a mixture distribution. Also, we show in Section 3.6, that we can assume that this latent measure is Dirichlet without loss of generality. In other words, we can represent $\widetilde{X}_T$ using a Gaussian mixture process whose latent mixing measure is a Dirichlet process.

The remaining issue is that Theorem 2 bounds the rescaled data $\widetilde{X}_T$, not X_T. Since $\widetilde{X}_T$ is rescaled with the square root of the sum of the squares over T periods and Theorem 2 is based on Theorem 1 about the data's norm, the bound from Theorem 2 is of the order $T\epsilon^2$ when applied to X_T. As T is increasing, this bound is not useful. In Section 3.5, we estimate X_T's transition density. We exploit the joint distributions' product structure implied by Assumption 2 to remove this T term in the bound.

3.5 Representing the Transition Density

We now show our model approximates transition densities well. The intuition behind the proof is as follows. Theorem 2 implies that $T\epsilon^2$ bounds the maximum deviation of the approximating joint density from the true density. Since the data are Markov, we construct the sample transition density as an average of the transitions in the data. This represents the joint distribution as a product of conditional densities. Thus the deviation of joint densities is the proportional to the sum of deviations of transition densities over T periods. Hence, we can use the bound for joint densities divided by T to bound the divergence of transition densities.

Theorem 3 (Transition Density Representation)**.** *Let X_T satisfy Assumption 1 and Assumption 2. Let p_T denote the true density. Let $\epsilon > 0$ and $\delta \in (0, 1/2)$ be given. Let Θ_T be constructed as in Definition 3. Let K_T be C_1(number of columns of $\Theta_T)^2$ for some constant C_1. Let δ_t be the cluster identity at time t. Then there exists a mixture density q_T with K_T clusters with the following form:*

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

Construct $q_T\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)$ from $q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}\right)$ by integrating out δ_{t-1} using $\Pr(\delta_{t-1} \,|\, X_T)$. Then with probability $1 - 2\delta$ with respect to Θ_T, there exists a universal constant C_2 such that the following holds uniformly:

$$h^2_\infty\left(p_T\left(x_t \,|\, \mathcal{F}^P_{t-1}\right), q_T\left(x_t \,\middle|\, \mathcal{F}^Q_{t-1}\right)\right) < C_2\left(1 + \log\left(\frac{1}{\delta}\right)\right)^2 \epsilon^2.$$

3.6 Replacing Θ_T with a Dirichlet Process

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

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

Lemma 1 (Replacing Θ_T with a Dirichlet Process)**.** *Let Q_T be a mixture distribution representable as an integral with respect to the Θ_T process defined in Definition 3. Then Q_T has a mixture representation as an integral with respect to the Dirichlet process.*

The intuition behind Lemma 1 is as follows. Both the Θ_T process and the Dirichlet process are based on the Chinese restaurant problem (seating customers at tables in a Chinese restaurant with an infinite number of circular tables). Since they have almost identical structure, we can replace Θ_T with the Dirichlet process without affecting our theoretical results.

4 Bayesian Nonparametrics and Convergence Rates

4.1 Problem Setup

We use the bounds constructed in the previous section to derive the convergence rates of the associated density estimators. We adopt the standard Bayesian nonparametric framework and show how fast the posteriors contract to the truth. Our definition of posterior contraction rate comes from (Ghosal and van der Vaart, 2017, Theorem 8.2). We assume the data X_T are drawn from some distribution P_T which is parameterized $P_T(\cdot \,|\, \xi)$, for $\xi \in \Xi$.

Definition 6. (Contraction Rate) A sequence ϵ_T is a *posterior contraction rate* at parameter ξ^P with respect to the semimetric d if $\mathcal{Q}_T\left(\{\xi \,\middle|\, d(\xi^P, \xi) \geq M_T\epsilon_T\} \,\middle|\, X_T\right) \to 0$ in $P_T\left(X_T \,\middle|\, \xi^P\right)$-probability for every $M_T \to \infty$.

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

We start by defining some notation that we use in deriving our theorems for the contraction rates. The concepts we use here are standard in the Bayesian nonparametrics literature. First, we define the metric (Kolmogorov) entropy for some small distance ϵ, some set Ξ, and some semimetrics, d_T and e_T. (One can use the same semimetric for both d_T and e_T.)

Definition 7. (Metric Entropy). $N(C\epsilon, \{\xi \in \Xi_T | d_T(\xi, \xi^P) \leq \epsilon\}, e_T)$ is the function whose value at $\epsilon > 0$ is the minimum number of balls of radius $C\epsilon$ with respect to the d_T semimetric (i.e., d_T-balls of radius $C\epsilon$) needed to cover an e_T-ball of radius ϵ around the true parameter ξ^P. The logarithm of $N(C\epsilon, \{\xi \in \Xi_T | d_T(\xi, \xi^P) \leq \epsilon\}, e_T)$ is metric entropy.

The metric entropy is the relevant measure of the model's complexity, and hence the "size" of the sieve. In the following sections, we show our model satisfies the conditions required to apply (Ghosal and van der Vaart, 2007a, Theorem 1). This theorem provides general conditions for convergence of posterior distributions even if the data are not i.i.d.[9]

4.2 Contraction Rates

Given the existence of uniformly consistent tests, we show the remaining conditions for (Ghosal and van der Vaart, 2007a, Theorem 1) hold by proving Proposition 4 and deriving the marginal and transition densities as special cases of it.

Proposition 4 (Bounding the Posterior Divergence). *Let X_T satisfy Assumption 1 and Assumption 2. Let $p_T := \sum_k \Pi_{k,t}\phi(x_t \,|\, \mu_t, \Sigma_t)$ denote the true density. Let $\Xi_T \subset \Xi$ and $T \to \infty$. Let q_T be a mixture approximation with $K_T^i = \frac{\log(T)^i}{\eta_T^2}$ components for $i \in \{1, 2\}$. Assume the following condition holds with probability $1 - 2\delta$ for $\delta \in (0, 1/2)$, and constant C:*

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

Let $\epsilon_T := \sqrt{\frac{\log(T)}{T}}$. Then there exist constants C_2 and C_3 such that the following two conditions hold with probability $1 - 2\delta$:

$$\sup_{\epsilon_i \geq \epsilon_T} \log N\left(C_2\epsilon_i, \left\{\xi \in \Xi_T \,\middle|\, h_\infty(\xi, \xi^P) \leq \epsilon_i\right\}, h_\infty\right) \leq T\epsilon_T^2,$$

and

$$\mathcal{Q}_T\left(B_T\left(\xi^P, \epsilon_T, 2\right) \,\middle|\, X_T\right) \geq \exp\left(-C_3 T\epsilon_T^2\right).$$

where $B_T\left(\xi^P, \epsilon_T, 2\right)$ is a ϵ_T-ball with respect to the divergence measure as in Ghosal and van der Vaart (2007a).

By taking $i = 2$, we can apply Proposition 4 to the transition density whose representation is in Theorem 3. As a consequence, the following result holds for the transition density.

[9] We show that uniformly consistent tests exist with respect to the semimetric we use: h_∞ in Lemma 8 in the online appendix.

Theorem 5 (Contraction Rate of the Transition Density)**.** *Let X_T satisfy Assumption 1 and Assumption 2. Denote its density $p_T := \sum_k \Pi_{t,k}\phi(x_t \,|\, \mu_t, \Sigma_t)$. Let $T \to \infty$, then the following holds with $\epsilon_T := \sqrt{\frac{\log(T)}{T}}$ with probability $1-2\delta, \delta \in (0, 1/2)$ with respect to the prior: There exists a constant C independent of T such that the posterior over the transition densities constructed as in Theorem 3 and the true transition density satisfies*

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

The constant C in Theorem 5 implicitly depends on a given δ as shown in Theorem 3. Estimating the Markov transition density with respect to h_∞ is more difficult than estimating the marginal density. A similar argument shows that Proposition 4 implies the following result for the marginal density by taking $i = 1$.

5 Estimation Strategy

The previous section focused on theoretical results. This section develops a Bayesian Gibbs sampler for our model. Algorithm 1 summarizes the steps. Recall the definition of the approximating model for the transition density:

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

We must place a prior on each of the components δ_t in this model. We start by placing a Dirichlet process prior on $\Pi_{t,k} := \Pi(\delta_t = k \,|\, \delta_{t-1})$ and, hence, implicitly on K_T. We then construct priors for β_k and Σ_k. A substantial literature exists on efficiently estimating Dirichlet mixture models (Ishwaran and James, 2001; Papaspiliopoulos and Roberts, 2008; Griffin and Walker, 2011). We use the slice sampler of Walker (2007) to handle the potentially infinite number of mixtures and compute a valid upper bound for K_T. Conditional on K_T, we draw the $\{\delta_t\}_{t=1}^T$ from their marginal distributions. We update the transition matrix Π so it has the correct marginal distributions. Given $\delta_t = k$, we apply standard Bayesian regression methods to obtain posterior draws on β_k and Σ_k. In addition, we use a conditionally conjugate hierarchical prior and draw from the hyperparameters' posterior.

Algorithm 1 Gibbs Sampler

1. **Posterior of $\{\delta_t\}_{t=1}^{T}$**
   - (a) Use Walker (2007) to determine the number of mixtures (clusters) K_T.
   - (b) Update the marginal probabilities for $\{\delta_t\}_{t=1}^{T}$, and the transition matrix, Π.
   - (c) Given K_T and $\{x_t\}_{t=1}^{T}$, use multinomial sampling to draw δ_t with

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

2. **Posterior of Π**
   - (a) Obtain the posterior of Π conditional on $\{\delta_t\}_{t=1}^{T}$ where the (j,k) element of Π is

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

   Recall that $\mathcal{Q}_0$ denotes the Dirichlet prior distribution.

3. **Posterior of Component-Specific Parameters**
   - (a) Given each cluster k, conduct a Bayesian regression to draw $\{\beta_k, \Sigma_k\}$.
4. **Posterior of Hyperparameters**
   - (a) Draw the hyperparameters governing $\{\beta_k, \Sigma_k\}$ from their conjugate posteriors.
5. **Iterate**

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

5.1.1 Bounding K_T

Our problem takes the same form as estimating a mixture model does in an i.i.d. context except our mixture identities have time-varying dynamics. We adapt the algorithm developed by Walker (2007) to our context and obtain the marginal mixture probabilities π_k.[10]

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

where v_k are distributed

[10] More details on this algorithm is described in the online appendix.

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

for $k = 0, 1, \ldots$ and $\alpha > 0$.

5.1.2 Correcting Π to have the Correct Marginal Distribution

We must construct a transition matrix where the relationship between two clusters, k and k^*, remains the same as they did in the previous draw of the sampler, but the marginal distribution is updated appropriately. We know that Markov transition matrices and their associated marginal distributions have the following relationship for each cluster k:[11]

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

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

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

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

If the new distribution $\widetilde{\pi}$ has more clusters than the previous draw π did, we use the prior. From $\widetilde{\Pi}$, we compute $\Pi_{t,k}$ for each t by drawing the first cluster identity, δ_0, from its stationary distribution and then using the Markov property of δ_{t-1} for $t > 1$ to iterate forward. Then the posterior of δ_t satisfies

$$\Pr\left(\delta_t = k \mid x_t, \Pi_{t,k}, \beta_k, \Sigma_k\right) \propto \phi\left(x_t \mid \beta_k x_{t-1}, \Sigma_k\right)\Pi_{t,k}.$$

Hence, categorical variables δ_t can be sampled directly with known probabilities.

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

5.2 Posterior on the Transition Matrix

We place the Dirichlet process prior on these cluster identities in each period to allow for an arbitrary number of clusters. By stacking the Dirichlet processes over time, we obtain a Dirichlet process over the (δ_{t-1}, δ_t) product space. Intuitively, we are constructing the transition matrix, Π, as a Dirichlet-distributed infinite-dimensional square matrix as noted by Lin et al. (2010).

Given the cluster identities $\{\delta_t\}_{t=1}^{T}$ which we drew in Section 5.1, we obtain the posterior on the transition matrices by counting the proportion of realized transitions and combining it with the prior probability of each transition.

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

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

5.3 Posterior for the Coefficient Parameters

Definition 8 gives the mixture component-specific likelihood where $X_k \coloneqq \{x_{t-1} \,|\, \delta_{t-1} = k\}$, $Y_k \coloneqq \{x_t \,|\, \delta_t = k\}$, and T_k is the number of datapoints in cluster k.

Definition 8. Component-Specific Likelihood

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

We estimate these parameters using component-by-component Bayesian regression. Each mixture component has varying amounts of data. When the forecast generates a new mixture component, we cannot condition on the data in that component. There is none. Consequently, we specify a hierarchical model to pool information across components. Hence, our model is component-by-component Bayesian regression with a conjugate Gaussian Inverse-Wishart prior generalized to allow for a hierarchical structure over the regression parameters.

Definition 9. Component-Specific Parameters' Prior

$$\begin{aligned} \{\beta_k\}_{k=1}^K \,|\, \Sigma_k, \bar{\beta}, U &\sim \mathcal{MN}\left(\bar{\beta}, \Sigma_k, U\right) \\ \{\Sigma_k\}_{k=1}^K \,|\, \Omega &\sim \mathcal{W}^{-1}\left((\mu_1 - 2)\Omega, \mu_1 + D - 1\right) \end{aligned}$$

where $\mathcal{MN}$ stands for matrix normal distribution and $\mathcal{W}^{-1}$ for Inverse-Wishart distribution.

This prior is the conjugate prior for the likelihood in Definition 8, and so we can use the standard formulas to estimate component-specific parameters. Derivations on their posteriors are provided in the online appendix. We now specify the hyperparameters' prior. As we did above, we place a conjugate matrix-normal prior on the coefficient matrix and an Inverse-Wishart prior on the covariance matrix.

Definition 10. Coefficient Hyperparameters' Prior

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

We adapt the hierarchical prior for $\Omega := \mathbb{E}[\Sigma_k]$ from Huang and Wand (2013). We have two degree of freedom parameters, μ_1 and μ_2, and D scale parameters for Ω: $a_1, \ldots, a_D$. Given these prior specifications, we derive the posteriors in a fairly standard way in the online appendix.

Definition 11 (Prior for the Covariances)**.**

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

6 Simulation

6.1 Data

We analyze how our estimator works when we know what the true data generating process (DGP) is. The DGP we consider is a vector autoregressive model with the Student's t-distributed innovations.[12] The Student's t-distribution is an infinite mixture of normal distributions where

[12] We also conducted simulation exercises with other specifications. These results are available upon request.

the variance is inverse-gamma distributed. The degrees of freedom for t-distributed innovations, which govern the fat-tailedness, is set to 5.7 as in Brunnermeier et al. (2019). Our DGP of bivariate ($D = 2$) data x_t is as follows:

$$x_t = \Phi_0 + \Phi_1(x_{t-1} - \Phi_0) + \Sigma^{1/2}\epsilon_t$$

$$\Phi_0 = \begin{bmatrix} 0.2 \\ 0.1 \end{bmatrix}, \quad \Phi_1 = \begin{bmatrix} 0.6 & -0.1 \\ 0.0 & 0.9 \end{bmatrix}, \quad \Sigma^{1/2} = \begin{bmatrix} 0.3 & 0.0 \\ 0.2 & 0.3 \end{bmatrix}, \quad \epsilon_{it} \sim_{i.i.d.} t(5.7)$$

6.2 Prior

As stated in Table 1, the prior for the mixture component coefficients has a Kronecker structure where we specify beliefs over the relationship between regressands and regressors separately.

Table 1: Prior

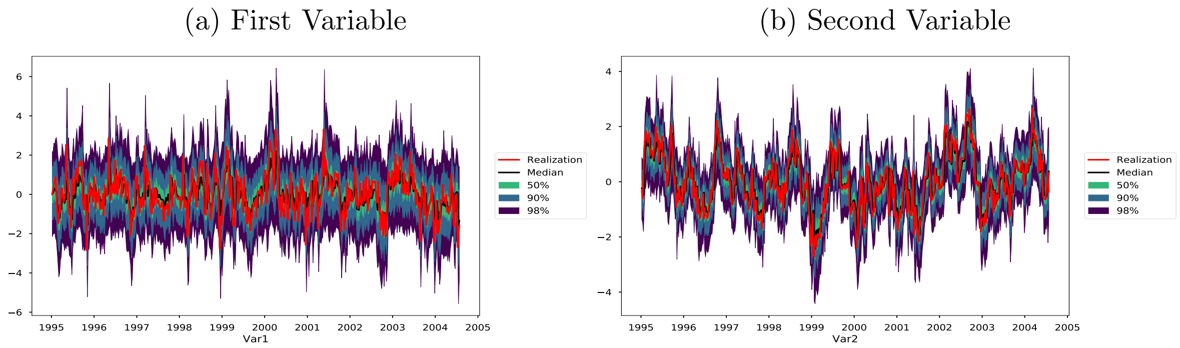
Degrees of freedom for the hierarchical prior	5
Expected number of mixture components	5
Component Coefficients	
Intercept	0
Expected diagonal autocorrelation	0.9
Expected off-diagonal autocorrelation	0
Component Covariances	
Mean	$.25^2\mathbb{I}_D$
μ_1	3
μ_2	3

The prior we use for the component parameters and base Dirichlet measure is rather flat, which means that we are not imposing a great deal of a priori structure. Lastly, although we do not have an explicit step in merging similar clusters in our sampler, our hierarchical prior will reduce separation between two similar clusters.

6.3 Simulation Results

We consider the data generating process of VAR(1) with the Student's t-distributed innovations. Figure 1 shows the in-sample predictive posterior density of x_t given x_{t-1}. The colored intervals show the credible sets based on posterior draws 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.

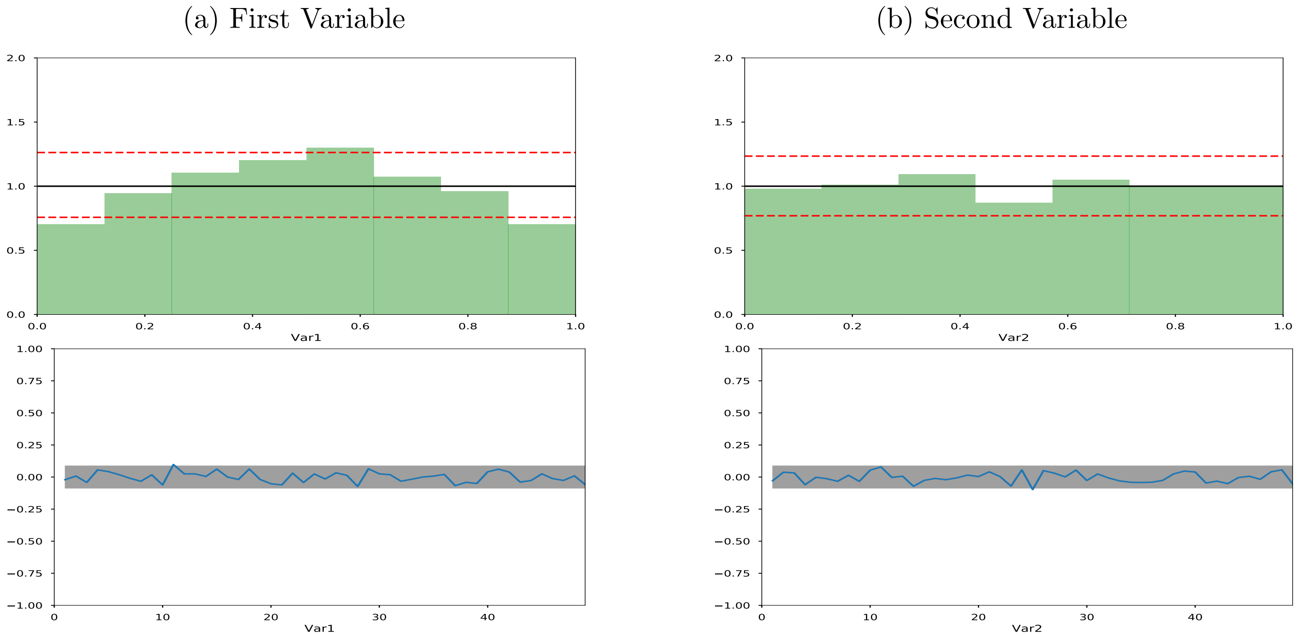
Figure 1: One-period Ahead Density Forecasts


(a) First Variable

(b) Second Variable

The first row of Figure 2 shows the probability integral transition (PIT) histograms. The PIT is the cumulative density of the random variable x_{T+1} evaluated at the true realization. The second row of Figure 2 shows the PIT autocorrelation functions (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 shaded area around the ACF is the credible set drawn using Barlett's formula. Based on Figure 2, our one-period ahead predictive density is correctly conditionally calibrated.

Figure 2: PIT Histogram and Autocorrelation Function (ACF)


(a) First Variable

(b) Second Variable

We can see from Figure 3 that we use more clusters as time progresses. Since the Student's

t-distribution has fatter tails than the normal distribution, we use at least three clusters in all of the periods. The rate at which the number of clusters increases is approximately logarithmic in the posterior, not just the prior, as predicted by our theory. In addition, when there arises a more complex dynamics compared to the past, our procedure is likely to add more clusters to approximate this dynamics. In Figure 3, we can see some spikes in the number of clusters over time. The blue solid line inside the green band stands for the median number of active clusters, which fluctuates between 5 and 12.

Figure 3: Number of Clusters Over Time

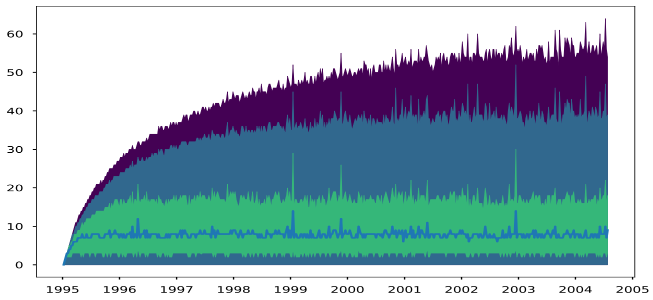
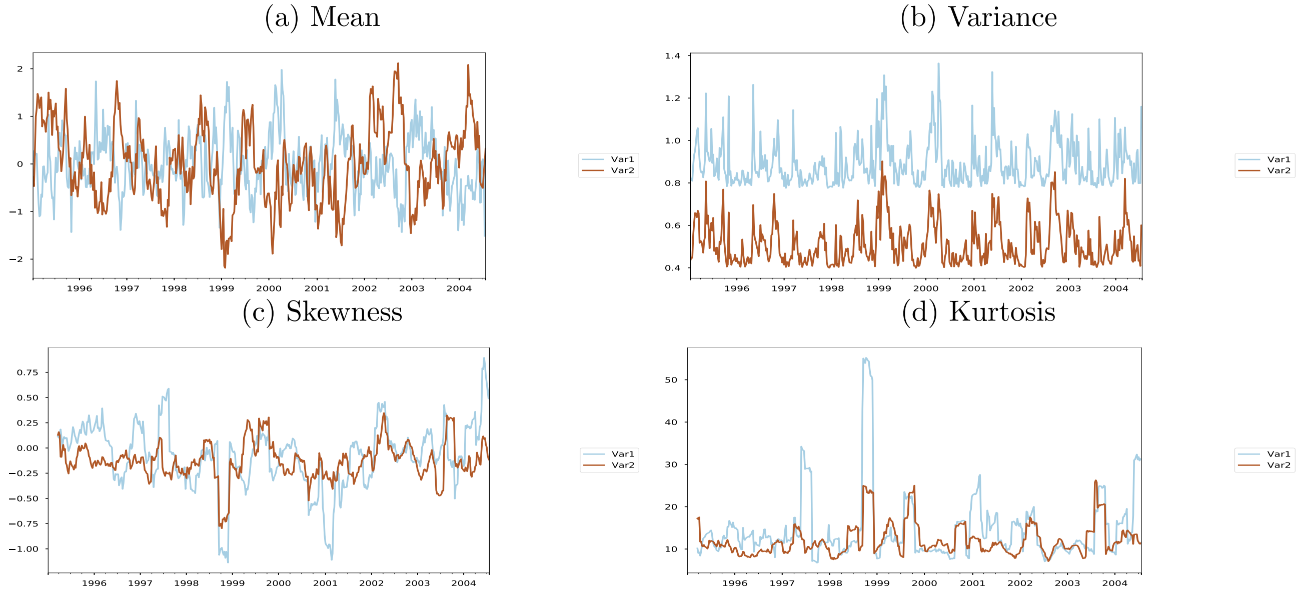


Figure 4: Time-varying Moments from One-period Ahead Density Forecasts



(a) Mean

(b) Variance

(c) Skewness

(d) Kurtosis

7 Empirical Analysis

7.1 Data and Prior

We downloaded monthly data on real consumption (DPCERAM1M225NBEA), personal consumption expenditure price index (PCEPI), industrial production (INDPRO), housing supply (MSACSR), unemployment rate (UNRATE), and 10-year government bond yields (IRLTLT01USM156N) from the Federal Reserve Bank of Saint Louis economic database (FRED). All of the data are seasonally-adjusted by FRED. We convert to approximate percent changes by log-differencing all of the data except for the consumption measure, which is already measured in percent changes, the unemployment rate, and the long-term interest rate. We then demean the data and rescale them so they have standard deviations equal to 1. This is useful because it puts all of the data on the same scale. The data are from January 1963 to December 2018. The time dimension is 671, and the cross-sectional dimension is 6.

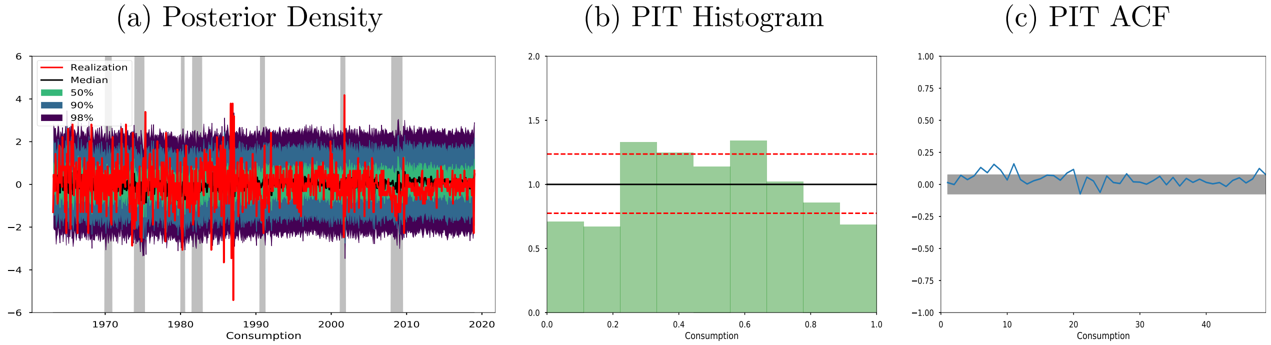
We use the prior as in Table 1, which is also used in our simulation. This prior specification does not impose too much structure a priori. Specifically, we do not impose how many clusters are necessary to approximate the evolution of densities. To the extent the simulation analysis and the empirical analysis require different numbers of clusters, this reflects different complexities in the datasets' dynamics.

7.2 Dynamics of Monthly Consumption Expenditure

To show that our algorithm works reasonably well in practice, we display the conditional density forecast for consumption in Figure 5. The online appendix provides predictive densities, PIT histograms, and ACFs for the other macroeconomic series. If the model works perfectly, the probability integral transform should be independent and distributed Uniform$[0, 1]$. As we can see, it is roughly independent and distributed approximately uniform.

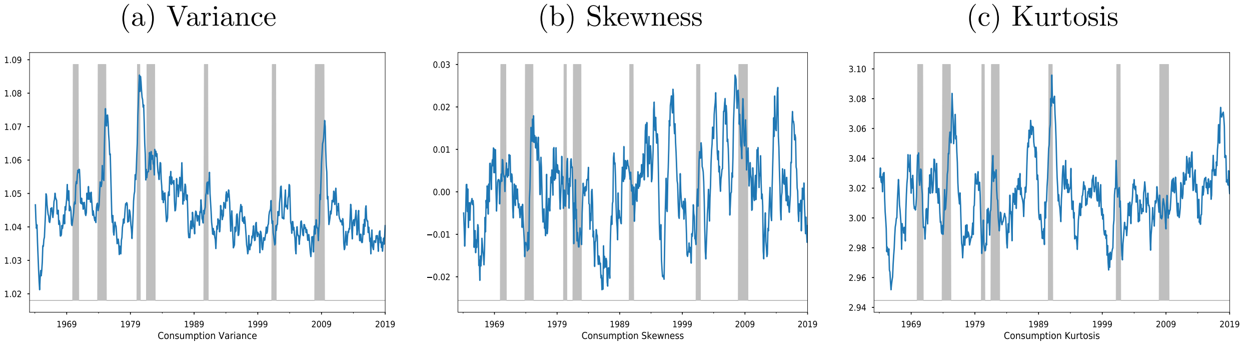
The dynamics of the data in Figure 5a are not obviously non-Gaussian or non-linear. One may question whether we are effectively just estimating a simple VAR. We show that this is not the case by Figure 6. Figure 6a illustrates that the conditional variance spikes a great deal in recessions when we compute the rolling averages over 1 year. Similar to Schorfheide

Figure 5: One-Period Ahead Conditional Forecasts: Consumption Expenditure


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

et al. (2018), we find stochastic volatility for consumption growth at business cycle frequencies using purely macroeconomic data. A VAR(1) could not capture this. We also find interesting results regarding higher moments of consumption. Skewness (Figure 6b) and kurtosis (Figure 6c) exhibit significant time-variation. Interestingly, skewness appears to decrease and kurtosis to increase during the shaded NBER recessions.

Figure 6: Consumption Variability


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

One may wonder how our model differs from a regime-switching model, which is quite popular in the literature. Our nonparametric approach uses an endogenously determined number of components to approximate the recession regime, instead of using just one as standard regime-switching models do. We use multiple clusters because our clusters serve two purposes. They let the mean change, as they do in regime-switching models, but they also model non-Gaussianity.

We find that the data are substantially less Gaussian during recessions, and this increase in the distributional complexity with time-varying higher moments holds for all the series considered. This finding aligns with the recent literature in macroeconomics and finance. For instance, Guvenen et al. (2014) point out that the left-skewness of income risk is counter-cyclical. That is, income shocks become more risky during recessions. Furthermore, the evolution of kurtosis shows that the consumption density becomes more fat-tailed in recessions. 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.

8 Conclusion

We construct a Bayesian nonparametric density estimator that, with high probability, converges fast for a large, fixed number of series. We devise a discrete random compression operator that induces a Dirichlet Gaussian mixture model to approximate a wide variety of densities. Based on this model, we provide a computationally efficient Gibbs sampler to estimate marginal and transition densities of multivariate processes.

We provide new theory that shows the posterior distributions of our density estimators converge more rapidly, with arbitrarily high probability with respect to random compression, than the literature has yet achieved. We show our estimators for the marginal and transition densities converge at a $\sqrt{\log(T)/T}$ rate with high probability.

We show that our estimators perform well in simulations and when applied to macroeconomic data. Our empirical analysis shows that macroeconomic data's dynamics are often far from Gaussian and change over the business cycle.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66:671–687.

Barro, R. J. and Jin, T. (2011). On the size distribution of macroeconomic disasters. *Econometrica*, 79(5):1567–1589.

Birgé, L. (2013). Robust tests for model selection. In Banerjee, M., Bunea, F., Huang, J., Koltchinksii, M., and Maathius, M. H., editors, *From Probability to Statistics and Back: High-Dimensional Models and Processes — A Festscrift in Honor of Jon A. Wellner*, volume 9 of *IMS Collections*, pages 47–68. Institute of Mathematical Statistics.

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

Brunnermeier, M., Palia, D., Sastry, K. A., and Sims, C. A. (2019). Feedbacks: Financial markets and economic activity.

de la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564.

Fox, E. (2009). Bayesian nonparametric learning of complex dynamical phenomena. *MIT Ph.D. Thesis*.

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

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

Ghosal, S. and van der Vaart, A. W. (2007a). Convergence rates of posterior distributions for non-i.i.d. observations. *The Annals of Statitics*, 35:192–223.

Ghosal, S. and van der Vaart, A. W. (2007b). Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statitics*, 35:697–723.

Ghosal, S. and van der Vaart, A. W. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Griffin, J. E. and Walker, S. G. (2011). Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259.

Guha, A., Ho, N., and Nguyen, X. (2020). On posterior contraction of parameters and interpretability in bayesian mixture modeling. *Bernoulli*.

Guhaniyogi, R. and Dunson, D. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110:1500–1514.

Guvenen, F., Ozkan, S., and Song, J. (2014). The nature of countercyclical income risk. *Journal of Political Economy*, 122:621–660.

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452.

Ichimura, H. and Todd, P. E. (2007). Implementing nonparametric and semiparametric estimators. volume 6, Part B of *Handbook of Econometrics*, chapter 74, pages 5369–5468. Elsevier.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

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

Klartag, B. and Mendelson, S. (2005). Empirical processes and random projections. *Journal of Functional Analysis*, 225(1):229–245.

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

Lin, D., Grimson, E., and Fisher, J. (2010). Construction of dependent dirichlet processes based on poisson processes. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23, pages 1396–1404. Curran Associates, Inc.

Nguyen, X. (2016). Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure. *Bernoulli*, 22(3):1535–1571.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, 38(3):1733–1766.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, 116:456–472.

Schorfheide, F., Song, D., and Yaron, A. (2018). Identifying long-run risks: A Bayesian mixed-frequency approach. *Econometrica*, 86(2):617–654.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.

Talagrand, M. (1996). Majorizing measures: The generic chaining. *The Annals of Probability*, 24(3):1049–1103.

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

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pages 90–110.

Tsai, J. and Wachter, J. A. (2016). Rare booms and disasters in a multisector endowment economy. *The Review of Financial Studies*, 29(5):1113–1169.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics – Simulation and Computation*, 36(1):45–54.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.

Feasible Multivariate Density Estimation Using Random Compression: Online Appendix

Online Appendix A Measure Concentration

A.1 Generic Chaining

We start by recalling a few definitions and fixing some notation. Recall the definition of a γ-functional:

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

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 $|\mathcal{X}_s| \leq 2^{2^s}$, $|\mathcal{X}_0| = 1$, and d is a metric. This $\gamma_2(\mathcal{X}, d)$ functional is useful because it controls the expected size of a Gaussian process by the majorizing measures theorem (Talagrand, 1996).

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

A.2 Properties of the Θ_T-operator

Lemma 2. *Let K_T be the number of columns of Θ_T as defined in Definition 3. Then its probability density function has the following form, where $b := \Pr(\zeta = 1)$.*

$$\Pr(K_T \leq \widetilde{K}) = \left(1 - (1-b)^{\widetilde{K}}\right)^T$$

Proof. Let θ_t denote a row of Θ_T. Then

$$\begin{aligned}\Pr(K \leq \widetilde{K}) &= \Pr(\theta_t \text{ includes 1 or -1 for all } t = 1, \ldots, T) = (\Pr(\theta_t \text{ includes 1 or -1}))^T \\ &= (1 - \Pr(\theta_t \text{ only includes 0's}))^T = \left(1 - (1-b)^{\widetilde{K}}\right)^T.\end{aligned}$$

□

Lemma 3. *Let K_T be the number of columns of Θ_T as defined in Definition 3, with $\Pr(\theta_{t,k} \neq 0) = b$. Then for any $\gamma \in (0, 1)$ there exist constants C_1 and C_2:*

$$\frac{C_1 \log(T)}{|\log(1-b)|} \leq K_T \leq \frac{C_2 \log(T)}{|\log(1-b)|}.$$

Proof. We set the cumulative distribution function equal to $1 - \gamma$, i.e. the survival function equal to γ:

$$(1-\gamma) = (1 - (1-b)^{K_T})^T \implies \log(1-\gamma)/T = \log(1 - (1-b)^{K_T}). \quad (6)$$

By Taylor's theorem there exist constant C_3, C_4 such that

$$-C_3(1-b)^{K_T} \leq \frac{\log(1-\gamma)}{T} \leq -C_4(1-b)^{K_T}.$$
$$C_4(1-b)^{K_T} \leq \frac{-\log(1-\gamma)}{T} \leq C_3(1-b)^{K_T}$$

Taking logs and multiplying through by -1:

$$-\log C_3 - K_T \log(1-b) \leq \log(T) - \log(-\log(1-\gamma)) \leq -\log C_4 - K_T \log(1-b)$$

$$-K_T \log(1-b) + \log(-\log(1-\gamma)/C_3) \leq \log(T) \leq -K_T \log(1-b) + \log(-\log(1-\gamma)/C_4) \quad (7)$$

Consider the lower bound in (7), if $\log(-\log(1-\gamma)/C_3) \geq 0$, we can just drop it. So assume without loss of generality that $\log(-\log(1-\gamma)/C_3) < 0$. Note (6) implies $-K_T \log(1-b) \geq -\log(\gamma)$. Hence, we get $\frac{\log(-\log(1-\gamma)/C_3)}{-K_T \log(1-b)} \geq \frac{\log(-\log(1-\gamma)/C_3)}{-\log(\gamma)}$. Then the left-hand side of (7) is bounded by $-K_T \log(1-b)\left(1 + \frac{\log(-\log(1-\gamma)/C_3)}{-\log(\gamma)}\right)$.

Consider the upper bound in (7), if $\log(-\log(1-\gamma)/C_4) < 0$, we can just drop it. Assume without loss of generality that it is positive. Then the right-hand side of (7) is bounded by $-K_T \log(1-b)\left(1 + \frac{\log(-\log(1-\gamma)/C_4)}{-\log(\gamma)}\right)$.

That is, there exist positive constants C_5, C_6 independent of b, K_T, and T such that

$$C_5 K_T |\log(1-b)| \leq \log(T) \leq C_6 K_T |\log(1-b)|.$$

Rearranging,

$$\frac{1}{C_6 |\log(1-b)|} \log(T) \leq K_T \leq \frac{1}{C_5 |\log(1-b)|} \log(T). \quad (8)$$

□

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 be an operator comprised of draws $\theta_{t,k}$ take from {-1, 0, 1} that the rows of Θ_T are i.i.d. and the columns of Θ_T form a martingale difference sequence. Let $b \in (0,1)$ denote $\Pr(\theta_{t,k} \neq 0)$. 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. First, we start by showing Item 1. Let Θ_k denote a column of the matrix. The root of the proof follows from realizing that Θ_T is a generalized selection matrix, and covariances are dominated by variances:

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

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

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

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

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

$$\begin{aligned}&\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{|\langle\Theta_k,x\rangle|^2}{C^2}\right)\right]-1\le 1\right\}\\ =&\inf\left\{C>0\,\middle|\,\mathbb{E}\left[\exp\left(\frac{\sum_{t=1}^{T}|\theta_{t,k}|x_t'x_t+2\sum_{t,\tau\neq t}\theta_{t,k}\theta_{\tau,k}x_t'x_\tau}{C^2}\right)\right]\le 2\right\}.\end{aligned}$$

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

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

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

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

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

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

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

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

where the last inequality follows because $b<1$. Hence, we have that the L_2-norm dominates the ψ_2-norm.

□

A.4 Norm Equivalence

In the section below we reproduce (Klartag and Mendelson, 2005, Prososition 2.2). The one change that we make is that we spell out one of the constants as a function of its arguments.

Proposition 6 (Klartag and Mendelson (2005) Proposition 2.2)**.** *Let $(\mathcal{X},d)$ be a metric space and let $\{Z_x\}_{x\in\mathcal{X}}$ be a stochastic process. Let $K>0,\Upsilon:[0,\infty)\to\mathbb{R}$ and set $W_x:=\Upsilon(|Z_x|)$ and $\epsilon:=\frac{\gamma_2(\mathcal{X},d)}{\sqrt{K}}$. Assume that for some $\eta>0$ and $\exp\left(-c_1(\eta)K\right)<\delta<\frac{1}{4}$, the following hold.*

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

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

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

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

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

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

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

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

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

Theorem 7 (Bernstein's Inequality for Martingales)**.** *Let $\{x_i,\mathcal{F}_i\}$ be a martingale difference sequence with $\mathbb{E}\left[x_i\,|\,\mathcal{F}_{i-1}\right]=0,\mathbb{E}\left[x_i^2\,|\,\mathcal{F}_{i-1}\right]=\sigma_i^2,v_k=\sum_{i=1}^k\sigma_i^2$. Furthermore, assume that $\mathbb{E}\left[|x_i|^n\,|\,\mathcal{F}_{i-1}\right]\leq\frac{n!}{2}\sigma_i^2M^{n-2}$ almost everywhere. Then, for all $x,y>0$,*

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

If we choose c small enough, this implies

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

A.5 Bounding the Norm Perturbation (Theorem 1)

Proof. We mimic the proof of (Klartag and Mendelson, 2005, Theorem 3.1), verifying the conditions of Proposition 6. 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 conclusion is weaker than theirs as $\gamma_2\left(\widetilde{\mathcal{X}}, \|\cdot\|_{L_2}\right) < C\sqrt{\log(T)}$. We can see this by combining the majorizing measure theorem (Talagrand, 2014, Theorem 2.4.1), and the minoration theorem (Talagrand, 2014, Lemma 2.4.2).

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

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

Consider $Z_x^K - Z_y^K$.

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

Let $Y_k := x^2(\theta_k) - y^2(\theta_k)$, then

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

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

The Y_k are a martingale difference sequence, and so we can apply Theorem 7. 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 7, we have the following if set $v := 4K\|\theta\|_{\psi_1}^2$ and $M = \|\theta\|_{\psi_1}$:

$$\Pr\left(\left\{\left|\frac{1}{K}\sum_{k=1}^{K}\theta_k\right| > u\right\}\right) \leq 2\exp\left(-cK\min\left\{\frac{u^2}{\|\theta\|_{\psi_1}^2}, \frac{u}{\|\theta\|_{\psi_1}}\right\}\right) \tag{9}$$

Then by applying (9) to $\Pr(\left|z_x^k - z_y^k\right| > u)$, we have the following.

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

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

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

By the triangle inequality,

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

Applying (9) for $u > 1$:

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

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

For any $x \in \mathcal{X}$ by (9),

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

We can bound the derivative of Υ:

$$\Upsilon'(0) = 1/2 > 0.$$

□

Online Appendix B Representation Theory

B.1 The Joint Density Setup

Lemma 5 (Bouding Ratio of Sums by Max Ratio)**.** *Let x_t, y_t be a sequence of positive numbers with a finite sum. Then the ratio of the sums is bounded by the supremum of the ratios, i.e.,*

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

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

$$\frac{x_1 + x_2}{y_1 + y_2} > \max\left\{\frac{x_1}{y_1}, \frac{x_2}{y_2}\right\} \implies x_1 + x_2 > \max\left\{x_1 + \frac{x_1 y_2}{y_1}, x_2 + \frac{x_2 y_1}{y_2}\right\}$$
$$\implies x_1 > \frac{x_2 y_1}{y_2} \text{ and } x_2 > \frac{x_1 y_2}{y_1} \implies x_1 > \frac{y_1}{y_2}\frac{x_1 y_2}{y_1} \implies x_1 > x_1.$$

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

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

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

Lemma 6. *Consider the ratio of the densities between p_T and q_T. Let δ_k^q be a clustering of x_t with respect to q_T. Let these clusters δ_k^q satisfy the following, where $\mu_k^q = \mathbb{E}_{P_T}[x_t \,|\, t \in \delta_k^q]$ and $\Sigma_k^q = \mathbb{C}\mathrm{ov}_{P_T}[x_t \,|\, x_t \in \delta_t^q]$:*

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

Then the log-divergence satisfies

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

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

$$\sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)'\Sigma_t^{-1}(x_t - \mu_t) - (x_t - \mu_k^q)' \left(\Sigma_k^q\right)^{-1} (x_t - \mu_k^q) \right| < C\epsilon. \tag{10}$$

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

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

By the definition of proportionality constant, we can write

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

where we can change measures to P_T. By Lemma 5, this is bounded by the supremum of the ratios, since we are integrating over the same space in both sums:

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

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

$$\left|\log\frac{\mu_1}{\mu_2}\right| < \frac{1}{2}C\epsilon \tag{11}$$

because the $k^{\cdot}(x)$ are Gaussian kernels, and we bounded the log-ratio in (10). The total deviation is the sum of the deviation in the constants and in the kernels. The result holds by combining (11) and (10).

□

Proposition 8 (Bounding the Supremum of the Rescaled Data)**.** *The data $X_T \in \mathbb{R}^{T\times D}$ is obtained by stacking the vector $x_t \in \mathbb{R}^D$ over $t = 1,\dots,T$. Let $p(x_t \,|\, \mathcal{F}_{t-1})$ satisfy Assumption 1. Let $\widetilde{X}_T$ denote the rescaled data as in equation (3). Let Θ_T be the random compression operator defined in Definition 3. Let δ_t^Q denote the mixture identity induced by Θ_T, and δ_t^P denote the true mixture identity. Let G_t^P and G_t^Q be the associated mixing measures. Given $\epsilon > 0$ and $\delta \in (0, 1/2)$, there exists a constant C such that*

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

with probability at least $1-2\delta$ with respect to Θ_T.

Proof. Let $\mathcal{K}$ be a coupling between the space of G^P and G^Q. Consider

$$\sup_t h^2\left(\int_{G_t}\phi\left(\widetilde{x}_t \,\middle|\, \delta_t^P\right) dG_t^P(\delta_t^P), \int_{G_t}\phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\right).$$

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

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

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

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

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

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

$$= \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} \sup_t \left| \exp\left(\frac{1}{2} \log\left(\frac{\phi\left(\widetilde{x}_t \,\middle|\, \delta_t^P\right)}{\phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right)} \right) \right) - 1 \right|^2 d\Phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right) d\mathcal{K}(G_t^P, G_t^Q).$$

By a first-order Taylor expansion of the exponential function and by Lemma 6 in the online appendix on the log divergence between the kernels:

$$\leq C_1 \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} \sup_t \left| (\widetilde{x}_t - \widetilde{\mu}_t^P)'(\widetilde{\Sigma}_t^P)^{-1}(\widetilde{x}_t - \widetilde{\mu}_t^P) - (\widetilde{x}_t - \widetilde{\mu}_t^Q)(\widetilde{\Sigma}_t^Q)^{-1}(\widetilde{x}_t - \widetilde{\mu}_t^Q) \right|^2 d\Phi\left(\widetilde{x}_t \,\middle|\, \delta_t^Q\right)$$
$$d\mathcal{K}(G_t^P, G_t^Q).$$

where $\widetilde{\mu}_t^P = \mathbb{E}[\widetilde{x}_t \,|\, \delta_t^P]$, $\widetilde{\Sigma}_t^P = \mathbb{C}\mathrm{ov}[\widetilde{x}_t \,|\, \delta_t^P]$, $\widetilde{\mu}_t^Q = \mathbb{E}[\widetilde{x}_t \,|\, \delta_t^Q]$, and $\widetilde{\Sigma}_t^Q = \mathbb{C}\mathrm{ov}[\widetilde{x}_t \,|\, \delta_t^Q]$. Note $\widetilde{Q}_T$ was obtained by applying Θ_T to $(\widetilde{\Sigma}_t^P)^{-1/2}(\widetilde{x}_t - \widetilde{\mu}_t^P)$. Also, the variance of the rescaled $\widetilde{X}_T$ is proportional to the variance of the non-rescaled $\widetilde{X}_T$ because they are both proportional to T. Hence, this norm perturbation is bounded by $C\epsilon^2$ with probability $1 - 2\delta$ with respect to Θ_T by Theorem 1:

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

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

B.2 Representing the Joint Density (Theorem 2)

Proof. Let G^P, G^Q be the associated mixing measures of the associated covariances. Let $\mathcal{K}$ be a coupling between the space of G^P and G^Q. 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. We can represent both densities succinctly as follows.

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

The squared supremum Hellinger distance h^2_∞ between the two densities is:

$$\begin{aligned} h^2_\infty\left(\widetilde{p}_T(\widetilde{\mathcal{X}}), \widetilde{q}_T(\widetilde{\mathcal{X}})\right) =& h^2_\infty\Bigg(\int_G \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta^P_t\right) dG^P_t(\delta^P_t)\, dG^P(dG^P_t), \\ & \int_G \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta^Q_t\right) dG^Q_t(\delta^Q_t)\, dG^Q(dG^Q_t)\Bigg). \end{aligned}$$

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:

$$= h^2_\infty\Bigg(\int_G \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta^P_t\right) dG^P_t(\delta^P_t)\, d\mathcal{K}(dG^P_t, dG^Q_t), \int_G \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta^Q_t\right) dG^Q_t(\delta^Q_t)\, d\mathcal{K}(dG^P_t, dG^Q_t)\Bigg).$$

Since supremum of squared Hellinger distance is convex, by Jensen's inequality:

$$\leq \int_{G\times G} \sup_t h^2 \left(\int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta^P_t\right) dG^P_t(\delta^P_t), \int_{G_t} \phi\left(\widetilde{x}_t \,\middle|\, \delta^Q_t\right) dG^Q_t(\delta^Q_t)\right) d\mathcal{K}(dG^P_t, dG^Q_t). \tag{12}$$

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 8 inside (12). Since we are integrating $C(1+\log(1/\delta))^2\epsilon^2$ over a joint density that is bounded above by 1, we have with probability $1-2\delta$ with respect to Θ_T:

$$h^2_\infty(\widetilde{p}_T(\widetilde{\mathcal{X}}), \widetilde{q}_T(\widetilde{\mathcal{X}})) < C\left(1+\log\left(\frac{1}{\delta}\right)\right)^2 \epsilon^2.$$

□

Lemma 7. *Let f, g be two densities of locally asymptotically mixed normal (LAMN) processes with respect to the sample size T. Squared Hellinger distance and Kullback-Leibler divergence are equivalent.*

Proof. Consider the following decomposition of the Hellinger distance:

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

Taking a Taylor expansion of the exponential function:

$$= \int \left(1 + \frac{1}{2}\log\left(\frac{f}{g}\right) + O\left(\log\left(\frac{f}{g}\right)^2\right) - 1\right) dG \tag{13}$$

$$= \int \frac{1}{2}\log\left(\frac{f}{g}\right) dG + O\left(\int \log\left(\frac{f}{g}\right)^2 dG\right). \tag{14}$$

Consider one-half the Kullback-Leibler divergence:

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

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

$$= \frac{1}{2}\int \log\left(\frac{f}{g}\right)\left(1+\log\left(\frac{f}{g}\right)\right)dG = \frac{1}{2}\int \log\left(\frac{f}{g}\right)dG + O\left(\int\left(\log\left(\frac{f}{g}\right)\right)^2 dG\right). \quad (15)$$

The first terms in (13) and (15) are the same. By the locally asymptotically mixed normal assumption $\log f(x) \propto (x-\mu_f)'\Sigma_f^{-1}(x-\mu_f) + o(T)$, where Σ is a random matrix. Choose $\epsilon \propto \frac{1}{T}$. Let z denote the deviation above. By the convexity of the square function and Jensen's inequality, it is sufficient to bound the value inside the integral:

$$\int \log(f/g)^2\,dG \leq \int |z|^2\,dG + O(\epsilon) \leq \int |z|\,dG + O(\epsilon) = \int \log(f/g)\,dG + O(\epsilon), \quad (16)$$

where the first inequality holds by the LAMN property, the second inequality holds since $|z| < 1$, and the third-inequality holds by the LAMN property. By (13) and (15), the last term in (16) is bounded by both the Hellinger and Kullback-Leibler divergences.

□

B.3 Representing the Transition Density (Theorem 3)

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

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

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

The first equality holds because the elements in each cluster have the same Gaussian distribution under q_T. The last equality holds because the elements of θ_{t-1} are in $\{-1, 0, 1\}$, we can absorb the θ_{t-1} into the $\beta_{k,k'}$ without increasing the number of clusters more than two-fold. This is because the vectors θ_{t-1} that contain at most one non-zero element form a convex hull, and we take the weighted averages over them in (17).

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

$$\widetilde{x}_t \,|\, \theta_{t-1} = k, \widetilde{x}_{t-1} \sim \sum_{k'} \phi(\beta_{k,k'}\widetilde{x}_{t-1}, \Sigma_{k,k'}) \Pr(\theta_t = k') \quad (17)$$

The last probability — $\Pr(\theta_t = k')$ — does not have any conditioning information because the rows of the Θ_T process are independent except for the stopping rule, which is not relevant here. Define a set of clusters in $(\widetilde{x}_t, \widetilde{x}_{t-1})$ space by grouping the ones whose associated $\{\beta, \Sigma\}$ are equal. In other words, take the Cartesian product of the clusters used in (17) and denote the cluster identities by δ_t's. Integrating out the cluster identities gives

$$\widetilde{x}_t \,|\, \widetilde{x}_{t-1}, \delta_{t-1} \sim \sum_j \phi(\beta_j \widetilde{x}_{t-1}, \Sigma_j) \Pr\left(\delta_t = j \,|\, \delta_{t-1}\right). \tag{18}$$

Clearly, there are $K_T^2 \propto \log(T)^2$ different clusters.

We make a similar argument to the one we made in the marginal density case. That is, we must show that the appropriate divergence between the transition densities is $1/T$ times the difference between the joint distributions. The goal is to show that the approximating transition distribution converges to the true transition distribution. From Proposition 8, we can bound the supremum Hellinger distance between the distributions of the rescaled data.

Consider the sup-squared-Hellinger distance considered in the proof of the joint density representation. Let $\mathcal{K}(G^P, G^Q)$ be any coupling between the two densities and integrate with respect to their joint density:

$$\sup_t h^2\left(\int_G \int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P)\, d\mathcal{K}(dG_t^P, dG_t^Q), \int_G \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q)\, d\mathcal{K}(dG_t^P, dG_t^Q)\right). \tag{19}$$

Taking the Schweppe decomposition of the joint distribution gives

$$\sup_t h^2\left(\prod_t \int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P \,|\, \mathcal{F}_{t-1}^P), \prod_t \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q \,|\, \mathcal{F}_{t-1}^Q)\right).$$

By Lemma 7, we can replace the squared Hellinger distance by Kullback-Leibler divergence

$$= C \sup_t \mathrm{D_{KL}}\left(\prod_t \int_{G_t} \phi\left(x_t \,|\, \delta_t^P\right) dG_t^P(\delta_t^P \,|\, \mathcal{F}_{t-1}^P) \,\middle\|\, \prod_t \int_{G_t} \phi\left(x_t \,\middle|\, \delta_t^Q\right) dG_t^Q(\delta_t^Q \,|\, \mathcal{F}_{t-1}^Q)\right).$$

Simplifying notation gives:

$$= C \sup_t \mathrm{D_{KL}}\left(\prod_t p_T\left(x_t \,|\, \mathcal{F}_{t-1}^P\right) \,\middle\|\, \prod_t q_T\left(x_t \,|\, \mathcal{F}_{t-1}^P\right)\right).$$

We can split apart the $\sup_t$ and write out the definition of Kullback-Leibler divergence:

$$C \sup_{F_{t-1}^P, \mathcal{F}_{t-1}^Q} \sup_{t \in \mathcal{F}_{t-1}^P \cap \mathcal{F}_{t-1}^Q} \int_{\mathbb{R}^{T \times D}} \log\left(\frac{\prod_t p_T\left(x_t \,|\, \mathcal{F}_{t-1}^P\right)}{\prod_t q_T\left(x_t \,|\, \mathcal{F}_{t-1}^P\right)}\right) \prod_t p_T(x_t \,|\, \mathcal{F}_{t-1}^P)\, dX_T.$$

Dropping the inner supremum cannot make the value larger:

$$\geq C \sup_{F_{t-1}^{P}, \mathcal{F}_{t-1}^{Q}} \int_{\mathbb{R}^{T \times D}} \log\left(\frac{\prod_t p_T\left(x_t \,|\, \mathcal{F}_{t-1}^{P}\right)}{\prod_t q_T\left(x_t \,|\, \mathcal{F}_{t-1}^{P}\right)}\right) \prod_t p_T(x_t \,|\, \mathcal{F}_{t-1}^{P})\, dX_T.$$

We can replace $\mathcal{F}_{t-1}^{P}$ and $\mathcal{F}_{t-1}^{Q}$ by the hidden Markov assumption.

$$= C \sup_{x_{t-1}, \delta_{t-1}^{P}, \delta_{t-1}^{Q}} \int_{\mathbb{R}^{T \times D}} \log\left(\frac{\prod_t p_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{P}\right)}{\prod_t q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{Q}\right)}\right) \prod_t p_T(x_t \,|\, x_{t-1}, \delta_{t-1}^{P})\, dX_T.$$

We can pull the supremum through the integral because it doesn't depend upon t; it only depends on the values of x_{t-1}, δ_{t-1}^{P}, and δ_{t-1}^{Q}:

$$= C \int_{\mathbb{R}^T} \sup_{x_{t-1}, \delta_{t-1}^{P}, \delta_{t-1}^{Q}} \int_{\mathbb{R}^D} \sum_t \log\left(\frac{p_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{P}\right)}{q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{Q}\right)}\right) \prod_t p_T(x_t \,|\, x_{t-1}, \delta_{t-1}^{P})\, dx_t\, d(\mathbb{R}^T).$$

We can pull the sum out:

$$= C \int_{\mathbb{R}^T} \sum_t \sup_{x_{t-1}, \delta_{t-1}^{P}, \delta_{t-1}^{Q}} \int_{\mathbb{R}^D} \log\left(\frac{p_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{P}\right)}{q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{Q}\right)}\right) \prod_t p_T(x_t \,|\, x_{t-1}, \delta_{t-1}^{P})\, dx_t\, d(\mathbb{R}^T).$$

The values inside the sum are all the same:

$$\geq CT \int_{\mathbb{R}^T} \sup_{x_{t-1}, \delta_{t-1}^{P}, \delta_{t-1}^{Q}} \int_{\mathbb{R}^D} \log\left(\frac{p_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{P}\right)}{q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{Q}\right)}\right) \prod_t p_T(x_t \,|\, x_{t-1}, \delta_{t-1}^{P})\, dx_t\, d(\mathbb{R}^T).$$

We can interchange the integral over $\mathbb{R}_T$ and the supremum because they are over different arguments of p_T and q_T; we also expand out the integral:

$$= CT \sup_{x_{t-1}, \delta_{t-1}^{P}, \delta_{t-1}^{Q}} \int_{\mathbb{R}^D} \cdots \int_{\mathbb{R}^D} \log\left(\frac{p_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{P}\right)}{q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{Q}\right)}\right) dP_T(x_1 \,|\, x_0, \delta_0^{P}) \cdots dP_T(x_T \,|\, x_{T-1}, \delta_{T-1}^{P}).$$

As in the marginal case, the only place that the densities inside the logarithm interact with the values is at t. We are taking the supremum over the conditioning argument so it cannot create any correlation. Where they do not interact we are simply integrating a constant over its entire domain.

$$= CT \sup_{x_{t-1}, \delta_{t-1}^{P}, \delta_{t-1}^{Q}} \int_{\mathbb{R}^D} \log\left(\frac{p_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{P}\right)}{q_T\left(x_t \,|\, x_{t-1}, \delta_{t-1}^{Q}\right)}\right) dP_T(x_t \,|\, x_{t-1}, \delta_{t-1}^{P}).$$

This is the sup-Kullback-Leibler divergence between the Markov transition densities:

$$= CT \sup_{\mathcal{F}^P_{t-1}, \mathcal{F}^Q_{t-1}} \mathrm{D}_{\mathrm{KL}} \left(p_T \left(x_t \,|\, \mathcal{F}^P_{t-1} \right) \middle|\middle| \, q_T \left(x_t \,|\, \mathcal{F}^Q_{t-1} \right) \right). \tag{20}$$

Equation (19) equals the distance between the joint distributions. Hence, by Theorem 2, we can bound it by $T(1 + \log(1/\delta))^2 \epsilon^2)$. The T term comes because we are no longer using rescaled data. By Lemma 7, we can replace the Kullback-Leibler divergence in (20) by squared Hellinger.

This gives

$$T \sup_{\mathcal{F}^P_{t-1}, \mathcal{F}^Q_{t-1}} h^2 \left(p_T \left(x_t \,|\, \mathcal{F}^P_{t-1} \right), q_T \left(x_t \,|\, \mathcal{F}^Q_{t-1} \right) \right) \leq CT(1 + \log(1/\delta))^2 \epsilon^2.$$

Canceling the T terms finishes the proof.

$$\sup_{\mathcal{F}^P_{t-1}, \mathcal{F}^Q_{t-1}} h^2 \left(p_T \left(x_t \,|\, \mathcal{F}^P_{t-1} \right), q_T \left(x_t \,|\, \mathcal{F}^Q_{t-1} \right) \right) \leq C(1 + \log(1/\delta))^2 \epsilon^2.$$

□

B.4 Replacing Θ_T with a Dirichlet Process (Lemma 1)

Proof. We can represent a Dirichlet process as $\Pr(x) = \sum_{i=1}^{\infty} \beta_i \delta_{x_i}(x)$, where δ_{x_i} is a indicator function with $\delta_{x_i}(x_i) = 1$, and the β_i satisfy a stick-breaking process. In other words, $\beta_i = \beta_i' \prod_{j=1}^{i-1}(1 - \beta_j')$ with $\beta_j' \sim \mathrm{Beta}(1, \alpha)$ for some positive scalar α. Consider the probability mass function of a row of Θ_T, θ_t. Then $\Pr(|i| = 1) = b \prod_{j=1}^{j-1}(1 - b)$. Since draws from the beta distribution lie in $(0, 1)$ with probability 1, these two stick-breaking processes are clearly mutually absolutely continuous.

Because these two processes are mutually absolutely continuous, a Radon-Nikodym derivative exists because both measures are σ-finite. Since the rows are independent, and Dirichlet processes are normalized random measures (Lin et al., 2010), we can extend this to the entire Θ_T process. In other words, we can choose the base measure of the Dirichlet process so that it puts positive probability on any atom that Θ_T does. Consequently, any process that is representable as an integral with respect to Θ_T can be represented as an integral with respect to to a Dirichlet process.

□

Online Appendix C Contraction Rates

C.1 Exponentially Consistent Tests with Respect to h_∞

Lemma 8 (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 $\epsilon > 0$, each $\xi_1 \in \Xi$, and true parameter ξ^P with $h_\infty(\xi_1, \xi^P)$:*

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

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

Proof. We can represent the joint density as a product density conditionally on a sequence of latent mixing measures G_t:

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

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

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

Then by (Birgé, 2013, Corollary 2), there exists a test ϕ_T that satisfies the following:[13]

$$\begin{aligned} &\Pr_T \left(\phi_T(X) \,\middle|\, \left\{G_t^f, G_t^g\right\}\right) \\ &\quad \leq \exp\left(-\frac{1}{3} T h_{\text{avg}}^2 \left(\int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi\left(x_t \,|\, \delta_t^g\right) dG_t^g(\delta_t^g)\right)\right) \end{aligned} \tag{21}$$

and

$$\begin{aligned} &\Pr_T \left(1 - \phi_T(X) \,\middle|\, \left\{G_t^f, G_t^g\right\}\right) \\ &\quad \leq \exp\left(-\frac{1}{3} T h_{\text{avg}}^2 \left(\int_{G_t^f} \phi\left(x_t \,\middle|\, \delta_t^f\right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi\left(x_t \,|\, \delta_t^g\right) dG_t^g(\delta_t^g)\right)\right). \end{aligned} \tag{22}$$

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

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

[13] To map his notation into ours, take his $z = 0$, and take his measure R equal to P. Equation (21) is obvious then, and (22) follows by taking the exponential of both sides in the inequality inside the probability and rearranging.

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

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

The terms inside the sum are all the same:

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

Taking the supremum over G_t^f and G_t^g is equivalent to taking supremum over $\mathcal{F}_{t-1}^f$ and $\mathcal{F}_{t-1}^g$ because the G_t^f and G_t^g are measurable functions of the later, and we are taking the supremum outside of the integral. They both span the same information sets. Since we can bound the error probabilities in both directions, using exponentially consistent tests, we have shown both items in Lemma 8 hold. □

C.2 Bounding the Posterior Divergence (Proposition 4)

Proof. We are looking at locally asymptotically mixed normal models, as discussed in Lemma 7, and we bind the Hellinger distance and Kullback-Leibler divergence in terms of $(x_t - \mu_t)'\Sigma_t^{-1}(x_t - \mu_t)$. In addition, the supremum of the deviations is clearly greater than the average of the deviations, and so the h_∞-norm forms smaller balls than both $\mathrm{D_{KL}}\left(f\,||\,g\right)$ and $V_{k,0}$. Consequently, we can replace $B_T(\xi_0, \epsilon_T, 2)$ with $\{\xi \in \Xi \,|\, h_\infty^2(\xi, \xi_0) < \epsilon_T^2\}$. We use 2 as the last argument of B because we are using $V_{2,0}$, i.e., effectively the 2nd-moment of the Kullback-Leibler divergence.

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

$$\sup_{\epsilon_i > \epsilon_T} \log N\left(C_2\epsilon_i, \{\xi \in \Xi_T \,|\, h_\infty(\xi, \xi_0) \leq \epsilon_i\}, h_\infty\right) \leq T\epsilon_T^2 \tag{23}$$

and

$$\mathcal{Q}_T\left(\left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi, \xi_0) < \epsilon_T^2\right\} |X_T\right) \geq \exp\left(-C_3 T\epsilon_T^2\right). \tag{24}$$

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

By assumption, there exists a covering with $K_T^i = \frac{\log(T)^i}{\eta_T^2}$ components such that the following holds:

$$\sup_t h\left(q_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^Q\right), p_T\left(x_t \,\middle|\, \mathcal{F}_{t-1}^P\right)\right) < C\eta_T. \tag{25}$$

Since ϵ_T^2 asymptotically dominates T, the right-hand-side of (24) is clearly less than $1-\delta$ for large T. The $\mathcal{Q}_T$ puts probability at least $1-\delta$ on the $(C\eta)-$ball surrounding ξ_0 by (25). So (24) is clearly satisfied if $C\eta_T \leq \epsilon_T$. Setting $C\eta_T = \epsilon_T$ gives $\eta_T = \frac{1}{C}\sqrt{\frac{\log(T)}{T}}$.

Solving for K_T gives $K_T = \frac{\log(T)^i}{\eta_T^2} = C^2 T \log(T)^{i-1}$. This K_T is proportional to the number of terms we are using, and the bracketing number is proportional to the covering number. In other words, for come constant C_1,

$$\log N\left(C_2\epsilon_i, \left\{\xi \in \Xi \,\middle|\, h_\infty^2(\xi,\xi_0) \leq \epsilon_i\right\}, h_\infty^2\right) = \log C_1 K_T$$
$$= \log\left(T\log(T)^{i-1}\right) + \log(C_1) = \log(T) + \log\left((i-1)\log(T)\right) + \log(C_1)$$

Because $\log(T)$ dominates the other terms, for some constant $C_4 > 1$:

$$\leq C_4 \log(T) = C_4 T\frac{\log(T)}{T} = C_4 T\epsilon_T^2.$$

This completes the proof because we can allow for a constant C_4 multiplied on the right-hand side of (23)[14].

□

C.3 Contraction Rate of the Transition Density (Theorem 5)

Proof. The proof of this is essentially identical to the marginal density case, mutatis mutandis. Lemma 8 implies the that h_∞ has the required exponentially consistent tests. We verify the conditions in Proposition 4. If we take $i = 2$ in the condition in Proposition 4, Theorem 3 implies the necessary bound on the sieve complexity exists.

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

□

Online Appendix D Estimation Strategy and Posterior Derivations

D.1 Bounding K_T with Walker (2007)

We draw the cluster identities by adapting 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.

[14] The proof of Ghosal and van der Vaart (2007a) goes through with this additional constant unchanged. $\sqrt{C_4}\epsilon_T$ characterizes the convergence rate for the distance h_∞. Since multiplying a norm by a constant clearly does not change the convergence rate, ϵ_T characterizes the convergence rate under $\frac{1}{\sqrt{C_4}}h_\infty$ (equivalently h_∞) as well. We have

$$\log N\left(C_2\sqrt{C_4}\epsilon_i, \left\{\xi \in \Xi \,\middle|\, \frac{1}{\sqrt{C_4}}h_\infty^2(\xi,\xi_0) \leq \epsilon_i\right\}, h_\infty^2\right) \leq T\epsilon_T^2.$$

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

$$q_T(x_t) = \sum_{k=1}^{\infty} \Pi_{t,k} \phi\left(x_t \,|\, \beta_k x_{t-1}, \Sigma_k\right). \tag{26}$$

As mentioned above, we introduce a latent variable $u_t \sim U(0, \Pi_{t,k})$ so we can rewrite (26) as

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

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

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

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

$$\Pr\left(\{v_k\}_{k=1}^{K} \,\middle|\, \{\delta_t\}_{t=1}^{T}\right) = \mathcal{Q}_0\left(\{v_k\}_{k=1}^{K}\right) \prod_{t=1}^{T} \mathbf{1}\left(v_{k=\delta_t} \prod_{\kappa<\delta_t} (1 - v_\kappa) > u_{k=\delta_t}\right), \tag{27}$$

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

The dependence between the u_t does not affect (27) because the v_k do not depend upon t. Hence, the v_k are conditionally independent given $\{u_t\}_{t=1}^{T}$. Exploiting this independence and the stick-breaking representation of the prior, we can draw v_k from (27); 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 (27) to draw v_k. As shown by Papaspiliopoulos and Roberts (2008), this implies v_k are distributed:

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

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

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

D.2 Component Coefficients Posterior

Let X_k be the $T_k \times N$ vector and Y_k be the $T_k \times D$ vector of data in component K. This implies that Σ_k is a $D \times D$ matrix and β_k is an $N \times D$ matrix.[15] Meanwhile, V is a $D \times D$ matrix and U is a $N \times N$ matrix.

The joint density is

$$\begin{aligned}\Pr\left(Y_k, \beta_k, \Sigma_k \mid X_k\right) = \exp\left(-\frac{1}{2}\operatorname{tr}\Big\{V_k^{-1}\left(\beta_k - \bar{\beta}\right)' U^{-1}\left(\beta_k - \bar{\beta}\right)\Big\}\right) \exp\left(-\frac{1}{2}\operatorname{tr}\big\{\left(Y_k - X_k\beta_k\right)\Sigma_k^{-1}\left(Y_k - X_k\beta_k\right)'\big\}\right) \\ \frac{|\Sigma_k|^{-T_k/2}}{(2\pi)^{T_k/2}} \frac{1}{\sqrt{(2\pi)^{ND}|V|^N|U|^D}} \frac{|(\mu_1 - 2)\Omega|^{\nu/2}}{\sqrt{2^{\nu D}}\Gamma_D(\frac{\nu}{2})} |\Sigma_k|^{-\frac{\nu+D+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\big\{(\mu_1 - 2)\Omega\Sigma_k^{-1}\big\}\right)\end{aligned} \tag{28}$$

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

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

Combining the two kernels of β_k and expanding gives

$$\begin{aligned}&\propto |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\Big\{V^{-1}\left(\left(\beta_k - \bar{\beta}\right)' U^{-1}\left(\beta_k - \bar{\beta}\right)\right) + \left(\left(Y_k - X_k\beta_k\right)'\left(Y_k - X_k\beta_k\right) + (\mu_1 - 2)\Omega\right)\Sigma_k^{-1}\Big\}\right) \\ &= |\Sigma_k|^{-\frac{\nu+D+T+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\big\{V_k^{-1}\left(\beta_k' U^{-1}\beta_k - 2\beta_k' U^{-1}\bar{\beta} + \bar{\beta}' U^{-1}\bar{\beta}\right) + \Sigma_k^{-1}\left(Y_k' Y_k - 2\beta_k' X_k' Y_k + \beta_k' X_k' X_k \beta_k + (\mu_1 - 2)\Omega\right)\big\}\right).\end{aligned}$$

Isolating the terms that have a β_k in them:

$$\begin{aligned}&= \exp\left(-\frac{1}{2}\operatorname{tr}\big\{V_k^{-1}\left(-2\beta_k' U^{-1}\bar{\beta} + \beta_k' U^{-1}\beta_k\right) + \Sigma_k^{-1}\left(-2\beta_k' X_k' Y_k + \beta_k' X_k' X_k \beta_k\right) + V_k^{-1}\bar{\beta}' U^{-1}\bar{\beta} + \Sigma_k^{-1}\left(Y_k' Y_k + (\mu_1 - 2)\Omega\right)\big\}\right). \\ &\quad |\Sigma_k|^{-\frac{\nu+D+T+1}{2}}\end{aligned}$$

[15] The likelihood in (28) is correct because the trace is the sum of the diagonal elements.

Rewriting the traces in terms of the vectorization operator:

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

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

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

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

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

$$= \exp\left(\mathrm{vec}\{\beta_k\}'\,\mathrm{vec}\big\{\left(U^{-1}\bar{\beta} + X_k' Y_k\right)\Sigma_k^{-1}\big\} - \frac{1}{2}\,\mathrm{vec}\{\beta_k\}'\left(\Sigma_k^{-1}\otimes\left(U^{-1} + X_k' X_k\right)\right)\mathrm{vec}\{\beta_k\}\right)$$
$$\exp\left(-\frac{1}{2}\,\mathrm{tr}\big\{\Sigma_k^{-1}\left(\bar{\beta}' U^{-1}\bar{\beta} + Y_k' Y_k + (\mu_1 - 2)\Omega\right)\big\}\right)|\Sigma_k|^{-\frac{\nu+D+T+1}{2}}. \tag{29}$$

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

and $W_k \coloneqq (U^{-1} + X_k' X_k)$. We can rewrite (29) as

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

I now eliminate all of the Kronecker products:

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

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

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

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

D.3 Hierarchical Mean Posterior with Heteroskedastic Data

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

$$\Pr\left(\{\beta\}_{k=1}^{K}, \bar{\beta}, \{\Sigma\}_{k=1}^{K}\right) = \exp\left(-\frac{1}{2}\,\mathrm{tr}\Big\{V^{-1}\left(\bar{\beta}-\beta^{\dagger}\right)' U^{-1}\left(\bar{\beta}-\beta^{\dagger}\right)\Big\}\right) \exp\left(\sum_{k=1}^{K} -\frac{1}{2}\,\mathrm{tr}\Big\{\Sigma_k^{-1}\left(\beta_k-\bar{\beta}\right)' U^{-1}\left(\beta_k-\bar{\beta}\right)\Big\}\right)$$
$$\sqrt{(2\pi)^{ND}|U|^{D}}\,|U|^{-\frac{\nu_U+N+1}{2}} \exp\left(-\frac{1}{2}\,\mathrm{tr}\big\{\Psi_U U^{-1}\big\}\right) \prod_{k=1}^{K} \frac{1}{\sqrt{(2\pi)^{ND}|\Sigma_k|^{N}|U|^{D}}}$$

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

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

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

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

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

$$\propto \exp\left(-\frac{1}{2}\,\mathrm{vec}\big\{\beta^{\dagger}\big\}'\left(V^{-1}\otimes W^{-1}\right)B + \mathrm{vec}\Big\{W^{-1}\beta^{\dagger\prime}V^{-1}\Big\}'B - \frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}\big\{(\Sigma_k^{-1}\otimes U^{-1})BB'\big\} + \mathrm{vec}\Bigg\{\sum_{k=1}^{K} U^{-1}\beta_k\Sigma_k^{-1}\Bigg\}' B\right)$$
$$|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}} \exp\left(-\frac{1}{2}\,\mathrm{tr}\Bigg\{V^{-1}\beta^{\dagger\prime}U^{-1}\beta^{\dagger} + \sum_{k=1}^{K}\Sigma_k^{-1}\beta_k' U^{-1}\beta_k + \Psi_U U^{-1}\Bigg\}\right).$$

We can simplify using the circular commutativity of the trace:

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

Collecting terms:

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

$$\begin{aligned}\propto \exp&\left(-\frac{1}{2}\operatorname{tr}\left\{\left(\sum_{k=1}^{K}\Sigma_k^{-1}+V^{-1}\right)\bar{\beta}'U^{-1}\bar{\beta}+\left(\beta^{\dagger}V^{-1}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\right)'U^{-1}\bar{\beta}\right\}\right)\\ &|U|^{-\frac{\nu_U+N+(K+1)D+1}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\left\{\left(\beta^{\dagger}V^{-1}\beta^{\dagger\prime}+\sum_{k=1}^{K}\beta_k\Sigma_k^{-1}\beta_k'+\Psi_U\right)U^{-1}\right\}\right).\end{aligned}\tag{30}$$

We now vectorize the first line of (30) after using the circular commutativity of the trace to simplify the square term. We drop the second line for now to simplify the exposition. We will bring it back in later. This gives

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

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

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

$$\exp\left(\frac{1}{2}\mathrm{vec}\{U^{-1}Z\}'\left(Z \otimes U^{-1}\right)^{-1}\mathrm{vec}\{U^{-1}Z\}\right)$$

We can simplify the vectorization.

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

We can replace the vectorizations with traces.

$$= \exp\left(-\frac{1}{2}\mathrm{tr}\{U^{-1}\left(\bar{\beta} - ZW^{-1}\right)W\left(\bar{\beta} - ZW^{-1}\right)\}\right)\exp\left(\frac{1}{2}\mathrm{tr}\{U^{-1}ZW^{-1}Z'\}\right) \tag{31}$$

Equation (31) is the kernel of a matrix normal distribution given the covariance matrices. We substitute the definitions of W and Z back in. The row matrix covariance is U, the column posterior covariance is $(\sum_{k=1}^K \Sigma_k^{-1} + V^{-1})$, and the mean is $(\beta^\dagger V^{-1} + \sum_{k=1}^K \beta_k \Sigma_k^{-1})(\sum_{k=1}^K \Sigma_k^{-1} + V^{-1})^{-1}$ Note, there is no reason here that β_k cannot itself be a matrix.

To compute the distribution of U, we combine the last lines of (30) and (31). This gives

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

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

D.4 Innovation Covariances' Mean Posterior

The product of the relevant likelihood and prior is

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

Since matrix multiplication distributes over matrix addition:

$$\begin{aligned}
=&|\Omega|^{\frac{K(\mu_1+D-1)}{2}} \exp\left(-\frac{\mu_1-2}{2}\sum_{k=1}^{K}\mathrm{tr}\left\{\Omega\Sigma_k^{-1}\right\}\right) \cdot |\Omega|^{\frac{\mu_2-2}{2}} \exp\left(-\frac{1}{2}\,\mathrm{tr}\left\{\mathrm{diag}(a_1,\ldots,a_D)^{-1}\Omega\right\}\right)\\
=&|\Omega|^{\frac{K(\mu_1+D-1)+\mu_2-2}{2}} \exp\left(-\frac{1}{2}\,\mathrm{tr}\left\{\left(\mathrm{diag}(a_1,\ldots,a_D)^{-1} + (\mu_1-2)\sum_{k=1}^{K}\Sigma_k^{-1}\right)\Omega\right\}\right).
\end{aligned}$$

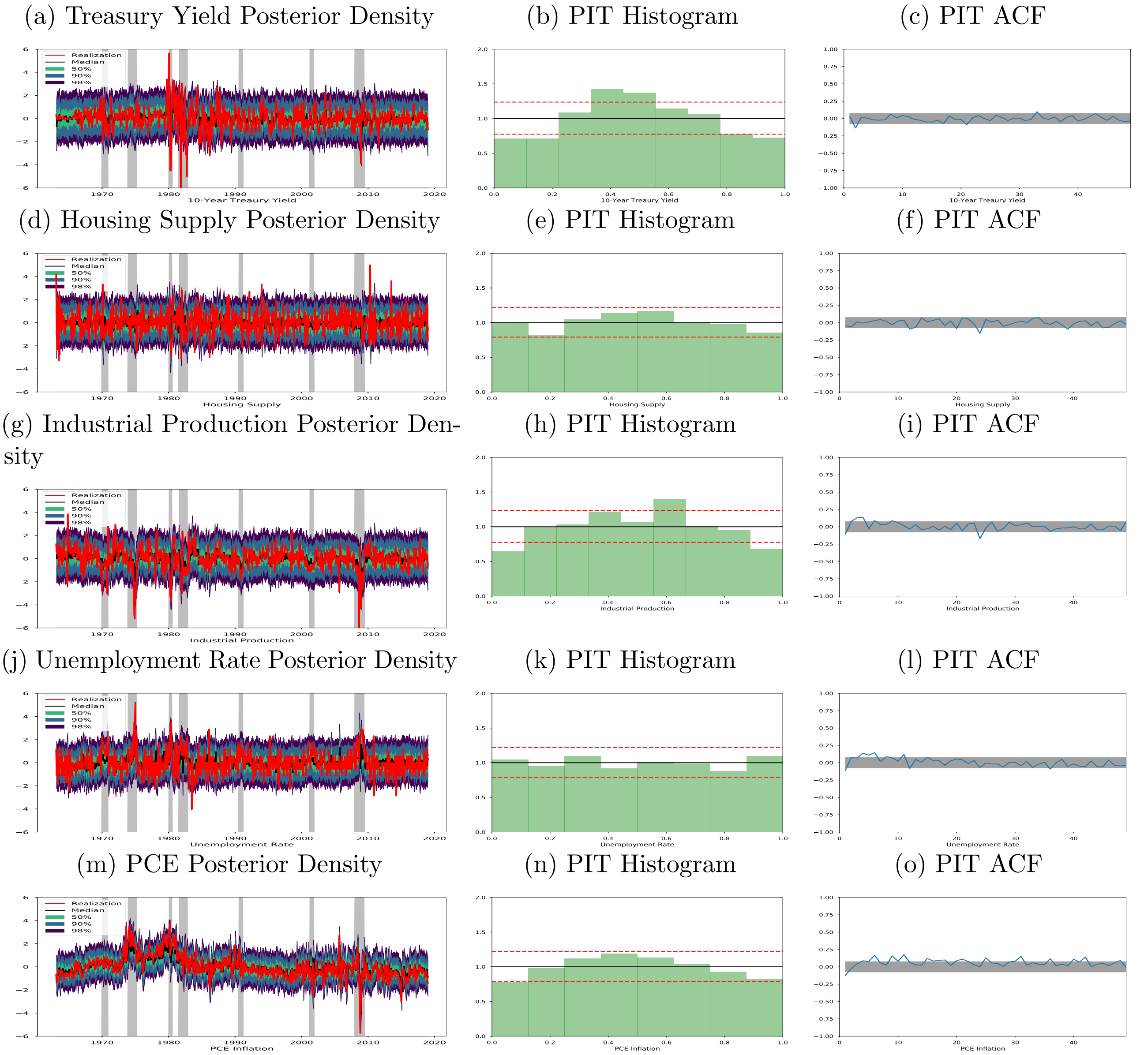
This is the kernel of a Wishart distribution. That is

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

Online Appendix E Empirical Analysis

E.1 One-Period Ahead Conditional Forecasts: Macroeconomic Variables

Figure 7: One-Period Ahead Conditional Forecasts



(a) Treasury Yield Posterior Density

(b) PIT Histogram

(c) PIT ACF

(d) Housing Supply Posterior Density

(e) PIT Histogram

(f) PIT ACF

(g) Industrial Production Posterior Density

(h) PIT Histogram

(i) PIT ACF

(j) Unemployment Rate Posterior Density

(k) PIT Histogram

(l) PIT ACF

(m) PCE Posterior Density

(n) PIT Histogram

(o) PIT ACF