Bypassing the Curse of Dimensionality: Feasible Multivariate Density Estimation*

Minsu Chang † Paul Sangrey ‡

Current Version
This Version: May 16, 2020

Abstract

Given vector-valued data — \( \{x_t\}_{t=1}^T \) — the curse of dimensionality makes non-parametrically estimating the data’s density infeasible when the number of series, \( D \), is large. We bypass the curse of dimensionality by adapting random compression to represent the density as a parsimonious mixture. For a number of periods, \( T \), the number of mixture components required to approximate a density to a given tolerance is a random variable. We construct a bound for this variable as a function of \( T \) that holds with high probability. We construct a nonparametric Bayesian estimator using Dirichlet processes. With high probability, our estimator’s convergence rates — \( \sqrt{\log(T)/\sqrt{T}} \) in the unconditional case and \( \log(T)/\sqrt{T} \) in the conditional case — depend on \( D \) only through the constant term. Because estimators that always converge rapidly do not exist, we construct estimators that converge rapidly most of the time. Our procedure produces a well-calibrated joint predictive density for a macroeconomic panel.

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

JEL Codes: C11, C14, C32, C55

*We are indebted to our advisors, Francis X. Diebold, Jesús Fernández-Villaverde, and Frank Schorfheide. We have also benefited greatly from conversations with Karun Adusumilli, Ben Connaught, Frank DiTraglia, Domenico Giannone, Laura Liu, Andriy Norets, Allan Timmermann, our colleagues at Amazon and Georgetown University, and seminar participants at 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 — $\{x_t\}_{t=1}^T$ — nonparametrically estimating the data’s density is infeasible when the number of series, $D$, is large. This phenomenon is called the curse of dimensionality.

This paper constructs a nonparametric, easy-to-use multivariate density estimator that scales well with the number of series. First, we construct an alternative representation of the density in the form of a parsimonious mixture of Gaussians by utilizing ideas from the literature on random compression. Our theory contribution lies in extending this idea of randomly compressing the data to the space of densities, which enables the computationally feasible high-dimensional density estimation. Second, we develop a Bayesian density estimator using a Gibbs sampler based on Dirichlet mixture models.

Unlike this paper’s approach based on random compression, the original curse of dimensionality papers, such as Stone (1980), examine how to approximate functions indexed by some smoothness class. They show that requiring the estimators to be consistent causes the estimator and the approximation, which is deterministic, to use the same number of terms asymptotically. To understand this, consider creating a multidimensional histogram. Dividing a $D$-dimensional hypercube into small hypercubes with width $1/T$ requires $T^D$ terms. The various deterministic approximations since the minimax estimation by Stone (1980) essentially form these high-dimensional histograms asymptotically (Yang and Barron (1999); Ichimura and Todd (2007)). In general, solving this deterministic problem requires $T^{g(D)}$ terms for some $g$ that depends upon the set of functions under consideration.

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

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

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

It is impossible to create a nonparametric density estimator that never requires exponentially many terms. 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 aforementioned data-agnostic procedure that determines the number of mixture components. We convert these bounds on the mixture’s complexity into convergence rates for the estimators. With high probability, our estimators’ convergences rates \(- \sqrt{\log(T)}/\sqrt{T} \) in the unconditional case and \( \log(T)/\sqrt{T} \) in the conditional case — depend on \( D \) only through the constant term, instead of decaying exponentially fast in \( D \) as minimax rates do.

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

We organize the paper as follows. Section 2 provides the intuition underlying
the results in the random compression literature, and hence, our results. Section 3
describes the data generating process. Section 4 constructs the sieve and provides
conditions under which it approximates the true density well. Section 5 proves our
estimators converge at the rates given above with high probability. Section 6 provides
a computationally efficient Gibbs sampling algorithm to estimate our sieve. Section 7
analyzes the model’s performance in a simulation with Student’s t-distributed shocks.
Section 8 empirically analyzes a monthly macroeconomic panel showing our method
works well in practice. Section 9 concludes. The appendices contain the proofs.

2 Intuition

The convergence rates discussed above likely seem surprising, so we now explain why
they are reasonable. We do this by discussing the intuition that drives the results
in the random compression literature. As discussed above, the standard convergence
rates are consequences of the number of bins of width $1/T$ required to fill a $D$-
dimensional hypercube equaling $T^D$. The random compression algorithms use fewer
terms than the deterministic approximations do by exploiting two facts.

First, random data tend to cluster in balls. For example, given a coverage level
(size) alpha, the lowest volume Gaussian confidence regions are ellipsoids. If the data
are i.i.d., they are $D$-dimensional balls. Why is this? A $D$-dimensional draw is in the
corner of a $D$-dimensional hypercube when all $D$ components of this draw are in their
tails. Obviously, if $D$ is large, this is incredibly unlikely. This intuition still holds in
correlated, non-Gaussian cases as long as the tails decay sufficiently rapidly.

Second, the volume of a $D$-dimensional ball grows exponentially slower with $D$
than the hypercube does as shown in Figure 1. As $D$ gets large, more and more
of the volume of the hypercube lies in the corners. If $D$ equals 1, the ball and the
hypercube coincide. They are both intervals. If $D = 2$, the ratio of volumes equals
\[(\pi r^2)/(4r^2) = \pi/4.\] If \(D = 3\), the volume ratio equals \((4/3\pi r^3)/(8r^3)) = \pi/6.\]

We exploit this behavior by constructing a sieve for the \(D\)-dimensional ball instead of constructing a sieve for the \(D\)-dimensional hypercube. A key insight is that constructing a sieve for the ball is much easier to deal with in high dimensions compared to a sieve for the hypercube as in Stone (1980). Since the volume of the ball grows more slowly, our sieve requires far fewer terms especially when \(D\) is large. To be clear, we consider asymptotic experiments when \(D\) is medium to large, but fixed and \(T\) grows.

Figure 1: Volume of a Ball Relative to a Hypercube

This paper exploits the simplicity this behavior implies about high-dimensional probability distributions to bound the number of terms required to estimate a density, instead of just compressing the data. Previous methods have shown how to compress the data while only slightly perturbing the data’s first two sample moments. We construct a sparse discretization operator (i.e., we bin the data) that does not significantly perturb the data’s first two sample moments. To convert this distance between the sample moments into a distance between densities, we use the fact that if a process is locally asymptotically mixed normal, its first two component-wise moments asymptotically form a component-wise sufficient statistic for the density. Consequently, densities are close when the first two component-wise moments are close.

We build a Dirichlet mixture process and adopt the standard Bayesian mixture framework. Below, we will show how the random compression operator described above implicitly creates a prior. The number of mixture components determines the complexity of a Gaussian mixture and the estimator’s convergence rate. Hence, we have a series of distributions indexed by \(T\). The distances between the estimator and

\[\frac{\pi}{2^D \Gamma(\frac{D}{2}+1)}\]

where \(\Gamma\) denotes the Gamma function.
the truth form a random variable whose distribution is indexed by $T$. The critical difference between our results and the previous ones in the literature is that we only require the convergence rate to hold in a $1 - 2\delta$ probability region with respect to the prior. In other words, we want our estimator to converge rapidly “most of the time” where “most” means with probability at least $1 - 2\delta$ and this probability is only taken with respect to the prior. We still require the convergence rate to be uniform with respect to the likelihood.

Because the previous literature requires the convergence rate to be uniform with respect to randomness in the prior, they cannot exploit the smoothness that the prior induces in deriving their convergence rates. At a technical level, for any fixed $T$ our sieve is not a measurable function of the data and so the Stone (1980) bounds do not apply.

### 3 Data Generating Process

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

**Definition 1 (Data Generating Process).** The data $X'_T$'s conditional densities given filtration $\mathcal{F}_{t-1}$ for each time period are

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

where $\Pi_{t-1,k}$ is the mixture probability of the $k^{th}$ component and $\phi \left( x_{k,t} \mid x_{t-1} \beta_{k,t}, \Sigma_{k,t} \right)$ stands for the probability density function of normal distribution having the mean $x_{t-1} \beta_{k,t}$ and the covariance $\Sigma_{k,t}$.

In words, $X'_T$'s conditional densities — $p_T(x_t \mid \mathcal{F}_{t-1})$ — have an infinite Gaussian mixture representations for each time period. Each mixture component has an associated mixture probability, $\Pi_{t-1,k}$ 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. The approximating model is a Gaussian mixture with $K_T$ components. The number of components — $K_T$ — governs the complexity of the model and so grows with $T$. 
Definition 2 (Approximating Model).

\[ q_T(x_t | F_{t-1}) := \sum_{k=1}^{K_T} \Pi_{t-1,k} \phi(x_t | x_{t-1}\beta_k, \Sigma_k). \]  

(2)

In this paper, we use the terms mixture and cluster interchangeably. Each cluster’s (mixture’s) components, \((\beta_k, \Sigma_k)\), no longer have time \(t\) subscripts. The idea is that we can reuse the latent variables \((\beta_{k,t}, \Sigma_{k,t})\) across time without loss of generality. If two separate periods have sufficiently similar dynamics, we group them into one component with the same parameters. Since the clusters are defined differently in the Definition 1 and Definition 2, no simple relationship between the parameters exists in general. This is also the case with the mixture probabilities.

Throughout, we use \(\mu_T\) to refer to the \(T \times D\) mean vector. We also consider the rescaled data:

\[ \tilde{X}_T := \frac{X_T - \mu_T}{\sqrt{\|X_T - \mu_T\|_{L^2}}} \in S^{TD-1} = \{ x \in \mathbb{R}^{TD} \mid \|X\|_{L^2} = 1 \}, \]  

(3)

where \(\|\cdot\|_{L^2}\) is the \(L^2\)-norm. Since we are on the unit hypersphere, we are in a compact space for any fixed \(T\). Since \(X_T - \mu_T\) is a zero-mean conditional Gaussian process, its \(TD \times TD\) covariance matrix completely determines its component-wise distributions. We define the densities of \(\tilde{X}_T\) as we did for \(X_T\) above and denote them \(\tilde{p}_T\) and \(\tilde{q}_T\).

In practice, we are making the following assumptions in this paper.

Assumption 1. Let \(X_T := \{x_t\}_{t=1}^{T}\) be a \(D\)-dimensional series where the conditional densities \(p(x_t | F_{t-1})\) given filtration \(F_{t-1}\) can be represented as infinite Gaussian mixtures for all \(t\). Further assume that the \(x_t\) have uniformly bounded means \(\mu_t\) and covariances \(\Sigma_t\) where the \(\Sigma_t\) are positive-definite.

Assumption 1 is a very general assumption. Tokdar (2006) shows that if there exists an \(\eta > 0\) such that the true distribution \(p_0\) satisfies \(\int |x|^\eta \, dP_0(x) < \infty\), then the first part of Assumption 1 on represent-ability as infinite Gaussian mixtures is satisfied. Also, Assumption 1 does not impose any structure on the relationship between the \(p(x_t | F_{t-1})\) over different time periods. The positive-definite assumption rules out perfect correlation between the various components in the vector \(x_t\).

We do need to restrict this relationship. In particular, we assume that \(X_T\) is a first-order hidden Markov process.
Assumption 2. Assume that there exists a latent state $z_t$ such that $(x_t', z_t')'$ form a uniformly ergodic first-order Markov sequence.

Note, if the $x_t$ form a Markov sequence, then this holds automatically; we can take $z_t$ to be a constant. In this paper, we will sometimes focus on the independent case where the $X_T$ are independent across $t$. This serves as a special case of a Markov sequence. Note that the independent case does not assume that the data are identically-distributed, just that they are independent. In other words, the density of $X_T$ is the product of the densities of each of the $x_t$.

4 Sieve Construction

4.1 Setting up the Problem

This section constructs a sieve that approximates a wide variety of DGPs while remaining simple. By simple, we mean that the metric entropy of these approximating models grows slowly with the number of datapoints. This property is useful because metric entropy controls the rate at which noteworthy posteriors converge (Ghosal, Ghosh, and van der Vaart (2000); Shen and Wasserman (2001)), and the minimax rate at which they can converge (Wong and Shen, 1995; Yang and Barron, 1999).

We approximate both a marginal density in the space of densities over $\mathbb{R}^D$ and a transition density that lies in the associated product space. These approximation problems are not well-posed because multiple equivalent representations exist for each density given $X_T$ that satisfy a given bound on the distance to $p_T$. We can exploit this multiplicity by choosing a representation that is particularly amenable to estimation for each $T$. We want a very parsimonious representation.

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

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

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

A standard discretization operator with $K$ bins is a $T \times K$ matrix where each row $\theta_t$ contains exactly one 1 and the rest of the elements equal zero. A variable $x_t$ is in bin $k$ if and only if $\theta_{t,k} = 1$, i.e., $\Theta_T$ has a 1 in row $t$ column $k$. We cannot use a standard discretization operator for two reasons. First, since all of the elements are weakly positive $E[\theta_{t,k}] \neq 0$. Second, once we see a 1, the rest of the columns in the row must be identically zero. This property makes the columns too dependent for our results to hold.

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

There are two important features of this sieve that differ from many of the sieves in the literature such as a kernel smoother. First, this approximation is global. The clustering is latent. We place different datapoints in the same bin if their densities are locally similar, not if the $x_t$ themselves are similar. Second, since each row of $\Theta$ 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. The densities in each of the components may not, by themselves, approximate the densities of any of the data. Local models such as kernel density estimators only use one density for each datapoint.

To make the discussion in the previous few paragraphs more formal, we define the random operator $\Theta_T$. We use a stick-breaking process to construct $\Theta_T$, adapting the process commonly used to construct Dirichlet processes (Sethuraman, 1994).\(^2\)

\(^2\)In Section 4.7, we show that a Dirichlet process can replace $\Theta_T$ without affecting our results.
**Definition 3 (Θₜ Operator).** Let b be a Bernoulli random variable with Pr(b = 1) ∈ (0, 1). Draw another random variable χ ∈ {−1, 1} with probability 1/2 each. Let T ∈ N be given. Draw T variables ϑ := χ · b independently of all of the previous values, and form them into a column-vector — Θ₁. Form another column vector Θ₂ the same way and append it to the right of Θ₁. Continue this until all of the rows of Θₜ contain at least one nonzero element.

We form the Θₜ operator this way so that E[ϑ] = 0 and Var(ϑ) = E[|ϑ|] = Pr(b = 1). Furthermore, its rows are independent and its columns form a martingale-difference sequence. The only dependence between the columns of Θₜ arises through the stopping rule, and stopped martingales are still martingales. In addition, Θₜ is independent of X₁. Since Θₜ is discrete, Θₜ implicitly clusters X₁. Consider some row θₜ of Θₜ. For each column of θₜ, define a bin as |θₜₖ| × sign(θₜₖ). Clearly, if Θₜ has Kₜ columns, there are 2Kₜ possible total bins.

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

\[
Pr(Kₜ \leq \bar{K}) \propto (1 - (1 - Pr(b = 1))^{\bar{K}})^T.
\]

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

We claimed above that Θₜ constructs an approximate sufficient statistic by binning X₁. In other words, we compress the data. Equation (4) quantifies the amount by which we compress the data. Instead of considering each of the T values of xₜ separately, and use a single parametric distribution within each bin. Since Kₜ ∝ log(T) ≪ T this substantially reduces the complexity.

We also must show that Θₜ preserves the xₜ’s densities. It is not a sufficient statistic if we lose any necessary information. We do this by adapting a well-known result — (Klartag and Mendelson, 2005, Theorem 3.1) — from the random compression literature.

The intuition behind this is that we can construct both of them using similar stick-breaking processes. Consequently, they are mutually absolutely continuous, a so a density exists that converts integrals with respect to one distribution into integrals with respect to the other.
**Theorem 1** (Bounding the Norm Perturbation). Let $\Theta_T$ be constructed as in Definition 3 with the number of columns denoted by $K_T$. Let $\epsilon > 0$ be given. Let $0 < \delta < 1/2$ be given such that $0 < \log(\frac{1}{\delta}) < c_1 \epsilon^2 K_T$ for some constant $c_1$. Let $\tilde{X}_T$ be in the unit hypersphere in $\mathbb{R}^{TD-1}$. Then with probability greater than $1 - 2\delta$ with respect to $\Theta_T$, there exists a constant $c_2$ such that for any $\epsilon > \sqrt{\frac{\log T}{K_T}}$,

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

Theorem 1 implies that when the number $K_T$ of $\Theta_T$’s columns satisfies $K_T \propto \log(T)$ applying $\Theta_T$ perturbs the norms of $\tilde{x}_t$ by at most $\epsilon$. This result holds with probability at least $1 - 2\delta$ with respect to the distribution over $\Theta_T$. Since $\tilde{X}_T \in S^{TD-1}$, we can map $S^{TD-1}$ onto a smaller space $S^{K_TD-1}$, with $K_T \ll T$, without perturbing the individual elements’ norms significantly.

The basic idea is that we are pre-multiplying the data by a martingale, i.e., a process whose expectation equals to one. This does not affect the mean or the variance. This increased randomness “smoothes” the data. To gain intuition, one can think about the average value. Koop, Korobilis, and Pettenuzzo (2019) do precisely this, focusing on Bayesian model averaging. This gives us tight bounds on the tails of the distribution with high probability. Since we have not changed the first two population moments and can tightly bound the tails of the distribution, we can place strong bounds on how much the clustering moves the sample moments. This is precisely what Theorem 1 does.

### 4.3 Distances on the Space of Densities

In the previous section, we showed that $\Theta_T$ does not affect $\tilde{x}_t$’s norms significantly. This is useful because these norms form a sufficient statistic for the components of Gaussian mixture distributions. To show the densities are close, we must convert the distances between the norms into distances on the space of densities.

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

Since $\Theta_T$ is almost surely discrete, this approximating distribution is a mixture, as
in Definition 2. We represent it as an integral with respect to a latent mixing measure — $G^Q_t$ — for each $t$. The parameters in each component are means and covariances, and so the $G^Q_t$ measure is over the space of means and covariances. Because $\Theta_T$ can have multiple nonzero elements $G^Q_t$ is a mixture distribution in each period, even conditional on $\Theta_T$.

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

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

$$\tilde{q}_T(\tilde{X}) = \int_G \int_G \phi(\tilde{x}_t | \delta^Q_t) dG^Q_t(\delta^Q_t) dG^Q(dG^Q_t). \tag{5}$$

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

$$\tilde{p}_T(\tilde{X}) = \int_G \int_G \phi(\tilde{x}_t | \delta^P_t) dG^P_t(\delta^P_t) dG^P(dG^P_t), \tag{6}$$

with its associated latent mixing measures and mixture identities. Note, the approximating cluster identities, $\{\delta^Q_t\}_{t=1}^T$, are different than the true cluster identities, $\{\delta^P_t\}_{t=1}^T$, because $\Theta_T$ induces $Q$’s clustering. It is not induced by the underlying true clustering.$^3$

We construct this bound in the space of densities by converting the bounds in $\tilde{x}_t$-space into bounds in $\Sigma_t$-space, which we then convert into bounds in the density-space. The norms of $\tilde{x}_t$ and $\tilde{x}_{t^*}$ being close does not imply that the associated matrix norms for $\Sigma_t$ and $\Sigma_{t^*}$ are close. Consequently, we cluster the rescaled data $\Sigma_t^{-1/2}\tilde{x}_t$ directly.

The error bound Theorem 1 provides does not depend on $X_T$ and so it does not depend on $\Sigma_t$. In other words, for times $t, t^*$ such that the associated $\tilde{x}_t$ and $\tilde{x}_{t^*}$ are

$^3$(6) 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.
contained in the same cluster, $\delta_k^Q$, the following holds:

$$\sup_{t,t^* \in \delta_k^Q} \left| \bar{x}_t \Sigma_t^{-1} \bar{x}_t - \bar{x}_{t^*} \Sigma_{t^*}^{-1} \bar{x}_{t^*} \right| < \epsilon. \quad (7)$$

Here $\epsilon$ is independent of $t$, $t^*$, and the cluster identity. The right-hand side of (7) is a “distance” on the space of covariance matrices. That is, we introduce the following semimetric on the space of covariance matrices.

**Definition 4 (Weighted-L$_2$ Semimetric).**

$$\delta_wl_2(\Sigma_k, \Omega_k) := \sup_{t,t^* \in \delta_k^Q} \left| \bar{x}_t \Sigma_k^{-1} \bar{x}_t - \bar{x}_{t^*} \Omega_k^{-1} \bar{x}_{t^*} \right|. \quad (8)$$

The space of covariances matrices equipped with $\delta_wl_2$ generates a Polish space. In particular, $\delta_wl_2$ constructs a set of equivalence classes over the space of covariance matrices, where two sample covariances are equivalent if the implied second-moment behavior of the $\{\bar{x}_t \in \delta_k^Q\}$ is indistinguishable.

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

**Definition 5.** Hellinger Distance

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

The Hellinger distance is useful because it is a valid norm on the space of densities. Since the covariance matrix is a sufficient statistic for a centered Gaussian, we can convert bounds between the covariances into bounds in this distance. Instead

---

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

5 It is a semimetric because we can have $\Sigma \neq \Omega$ but $\delta_wl_2(\Sigma, \Omega) = 0$. The two matrices may differ that cannot be identified by the set $x \in \delta_k$.

6 This norm is compatible with and weaker than the max-norm. If $x$, $y$ in $x \Sigma^{-1} y$ are (possibly) different unit selection vectors we can pick out the maximum absolute deviation between elements in the two matrices. This difference is clearly at least as big as the $\delta_wl_2$ because that semimetric requires $x$, $y$ to be the same. The max-norm is equivalent to the $L_\infty$-norm up to a scale transformation, and the relevant scale is a constant since we only consider full-rank matrices. This implies the space of covariance matrices is isomorphic to $\mathbb{R}^{D \times D}$. We are choosing an open subset of that space.
of applying this directly to the joint distribution, we take the supremum over the conditional distributions.

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

\[
    h^2_\infty(p, q) := \sup_{F \in \mathcal{P}_{T-1}, \mathcal{Q}_{T-1}, 1 \leq t \leq T} h^2\left(p\left(\cdot \mid \mathcal{F}_{t-1}^P\right), q\left(\cdot \mid \mathcal{F}_{t-1}^Q\right)\right) \quad (10)
\]

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

### 4.4 Representing the Joint Density

We now show that the approximating distribution of \( \tilde{X}_T \) induced by \( \Theta_T \) is close to the true distribution \( \tilde{P}_T \) using \( h_\infty \). We can do this whenever the rescaled trace is a component-wise sufficient statistic for the density. Hence, we can use bounds on divergences in the space of \( \tilde{X}_t - \tilde{X} \) to bound divergences in the space of densities.

**Theorem 2** (Representing the Joint Density). Let \( \tilde{X}_T := \frac{X_T - \mu_T}{\sqrt{\|X_T - \mu_T\|_2}} \) where \( X_T \) satisfies Assumption 1. Let \( \Theta_T \) be the generalized selection matrix constructed in Definition 3. Let \( \tilde{P}_T \) denote the distribution of \( \tilde{X}_T \). Then given \( \epsilon > 0 \) and \( \delta \in (0, \frac{1}{2}) \), the approximating distribution, \( Q_T \), which is the mixture distribution over \( \tilde{X} \) that \( \Theta_T \) induces, satisfies the following with probability at least \( 1 - 2\delta \) with respect to \( \Theta_T \) for some constant \( C \):

\[
    h_\infty\left(\tilde{P}_T(\tilde{X}), Q_T(\tilde{X})\right) < C \left(1 + \log\left(\frac{1}{\delta}\right)\right) \epsilon.
\]

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

Unlike the previous compression operators in the literature, \( \Theta_T \) is discrete, and so it clusters \( \tilde{x}_t \). This property implies that the density of \( \tilde{x}_t \) is a process with respect to a
discrete measure. That is, \( Q_T \) is a mixture distribution. Also, we show in Section 4.7, that we can assume that this latent measure is Dirichlet without loss of generality. In other words, our method represents the \( \tilde{X}_T \) process as an integral with respect to a Dirichlet process. Consequently, since \( \tilde{X}_T \) is a Gaussian mixture process, and hence locally mixed Gaussian, we can represent \( \tilde{X}_T \) using a Gaussian mixture process whose latent mixing measure is a Dirichlet process.

The leading issue that remains is that Theorem 2 bounds the rescaled data, \( \tilde{X}_T \), not \( X_T \). As one might expect, estimating the true joint density of \( X_T \) is impossible. Since \( \|X_T\|_2^2 \propto T \), the bound we have is of the order \( \sqrt{T} \epsilon \), which is useless. Instead, we consider quantities such as \( X_T \)'s marginal density (Section 4.5) and transition density (Section 4.6). We show that sample means of the marginal and transition densities converge to those implied by \( Q_T \), and hence those implied by \( P_T \). This convergence occurs because sample means converge to population means.

## 4.5 Representing the Marginal Density

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

**Theorem 3** (Representing the Marginal Density). *Let \( X_T \) satisfy Assumption 1 and assume that the \( X_T \) are independent across \( t \). Let \( \Theta_T \) be constructed as in Definition 3. Let \( \epsilon > 0, \delta \in (0, 1/2) \) be given. Construct \( Q_T \) as the mixture model in Definition 2 where \( \Theta_T \) groups the data into components. Then, with probability \( 1 - 2\delta \) with respect to \( \Theta_T \), there exists a constant \( C \) such that the following holds uniformly over \( T \)

\[
\epsilon \left( \int_{G_t} \phi \left( x_t \mid \delta_{t}^{P} \right) dG_t(\delta_{t}^{P}), \int_{G_t} \phi \left( x_t \mid \delta_{t}^{Q} \right) dG_t(\delta_{t}^{Q}) \right) < C \left( 1 + \log \left( \frac{1}{\delta} \right) \right) \epsilon.
\]

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

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

### 4.6 Representing the Transition Density

We now show our model approximates transition densities well. Since the data are Markov, we construct the sample transition density as an average of the transitions in the data. Component by component, we solve for the correct conditional distributions in the approximating model. We relate the error in the transition densities to the error for the joint densities. The space of transitions form a product space: $X_T \otimes X_T$. We construct the transitions’ marginal density in the space. As in Section 4.5, the approximate product form gives us a $1/T$ term in the convergence rate. Proposition 11 gives us a $T\epsilon^2$ term. The $T$ terms cancel, and so $\epsilon^2$ bounds the distance between the densities.

**Theorem 4** (Transition Density Representation). *Let $X_T$ satisfy Assumption 1 and Assumption 2. Let $p_T$ denote the true density. Let $\epsilon > 0, \delta \in (0, 1/2)$ be given. Let $\Theta_T$ be constructed as in Definition 3. Let $K := C(\text{number of columns of } (\Theta_T))^2$ for some constant $C$. Let $\delta_t$ be the cluster identity at time $t$. Then there exists a mixture density $q_T$ with $K$ clusters with the following form:*

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

*Construct $q_T(x_t | F^Q_{t-1})$ from $q_T(x_t | x_{t-1}, \delta_{t-1})$ by integrating out $\delta_{t-1}$ using $\Pr(\delta_{t-1} | X_T)$. Then with probability $1 - 2\delta$ with respect to the prior

$$h_\infty\left(p_T(x_t | F^P_{t-1}), q_T(x_t | F^Q_{t-1})\right) < C \left(1 + \log \left(\frac{1}{\delta}\right)\right) \epsilon.$$
4.7 Replacing $\Theta_T$ with a Dirichlet Process

The previous subsections use $\Theta_T$ 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 $\Theta_T$ can be chosen to be a Dirichlet process without loss of generality.

Consider the $\Theta_T$ process as in Definition 3 except we no longer stop when we no longer need columns. Then we can replace $\Theta_T$ 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, Ghosh, and van der Vaart, 2000; Walker, 2007).

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

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

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

5 Bayesian Nonparametrics and Convergence Rates

5.1 Problem Setup

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

We assume the data \( \{x_t\}_{t=1}^{T} \) are drawn from some distribution \( P_T \) which is parameterized \( P_T(\cdot | \xi) \), for \( \xi \in \Xi \). This parameter set is equipped with the Borel \( \sigma \)-algebra \( \mathcal{B} \) with associated prior distribution \( Q_0(\xi) \). We assume there exists a regular version of the conditional distribution of \( \xi \) given \( X_T \), which is called the posterior:

\[
Q_T(B | X_T) := \Pr(\{\xi \in B\} | X_T), \; B \in \mathcal{B}.
\]

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

**Definition 7.** Contraction Rate A sequence \( \epsilon_T \) is a posterior contraction rate at parameter \( \xi^P \) with respect to the semimetric \( d \) if \( Q_T \left( \{\xi | d(\xi^P, \xi) \geq M_T \epsilon_T\} | X_T \right) \to 0 \) in \( P_T \left( X_T \right| \xi^P \right) \)-probability for every \( M_T \to \infty \).

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

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

**Definition 8.** Metric Entropy \( N(C\epsilon, d_T(\xi, \xi^P), e_T) \) is the function whose value for \( \epsilon > 0 \) is the minimum number of balls of radius \( C\epsilon \) with respect to the \( d_T \) semimetric (i.e., \( d_T \)-balls of radius \( C\epsilon \)) needed to cover an \( e_T \)-ball of radius \( \epsilon \) around the true parameter \( \xi^P \).

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

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

$$V_{k,0}(f, g) := \int |\log(f/g) - D_{KL}(f \parallel g)|^k dF. \quad (11)$$

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

$$B_T(\xi^P, \epsilon, k) := \left\{ \xi \in \Xi \left| D_{KL} \left( f \left( X_T \mid \xi^P \right) \bigg\| f \left( X_T \mid \xi \right) \right) \leq T\epsilon^2, \right. \right. \left. \left. V_{k,0} \left( f \left( X_T \mid \xi^P \right), f \left( X_T \mid \xi \right) \right) \leq T\epsilon^2 \right\}. \quad (12)$$

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

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

1. There exist exponentially consistent tests $Y_T$ as in Lemma 2 with respect to $d_T$.

2. \[ \sup_{\epsilon_T > \epsilon} \log N \left( \frac{C_2}{2} \epsilon, \left\{ \xi \in \Xi_T \mid d_T(\xi, \xi^P) \leq \epsilon \right\}, e_T \right) \leq T\epsilon^2. \quad (13) \]

3. \[ \frac{Q_T \left( \left\{ \xi \in \Xi_T \mid n\epsilon_T < d_T(\xi, \xi^P) \leq 2n\epsilon_T \right\} \mid X \right)}{Q_T \left( B_T(\xi^P, \epsilon_T, C_1) \mid X \right)} \leq \exp \left( \frac{C_2T\epsilon^2n^2}{2} \right) \quad (14) \]

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

$$P_T \left( Q_T \left( \left\{ \xi \in \Xi_T \mid d_T(\xi, \xi^P) \geq M_T\epsilon_T \right\} \mid X \right) \mid \xi^P \right) \to 0. \quad (15)$$
5.2 Contraction Rates

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

Note, $h_{\infty}^2$ should be interpreted as a distance on the joint distributions because we can always factor a joint distribution as

$$f(X_T) = f(x_T | F_{T-1}) \cdot f(x_{T-1} | F_{T-2}) \cdots f(x_2 | F_1) \cdot f(x_1 | F_0),$$

where $F_0$ denotes information that is always known, as is standard.

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

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

1. $P_T (\mathcal{Y}_T | \xi^P) \leq \exp(-C_2 T \epsilon^2)$

2. $\sup_{\xi \in \Xi, \epsilon_\alpha(\xi, \xi) < \epsilon} P_T (1 - \mathcal{Y}_T | \xi^P) \leq \exp(-C_2 T \epsilon^2)$

Having shown the appropriate tests exist, we now show (13) and (14). As noted in (Ghosal and van der Vaart, 2007, 197), the numerator is trivially bounded by 1, as long as $T \epsilon_T \to \infty$ which it does in this case. We do this by proving a proposition that covers both the marginal and transition density cases. We can deduce the main theorems as results of it.

**Proposition 6** (Bounding the Posterior Divergence). Let $X_T$ satisfy Assumption 1 and Assumption 2. Let $p_T := \sum_k \Pi_k \phi(x_t | \mu_t, \Sigma_t)$ denote the true density. Let $\Xi_T \subset \Xi$ and $T \to \infty$. Let $Q_T$ be a mixture approximation with $K_T$ components. Assume the following condition holds with probability $1 - 2\delta$ for $\delta \in (0, 1/2)$ and constants $C$ and $i \in \mathbb{N}$:

$$\sup_{t} h \left( q_T \left( x_t \mid F^p_{t-1} \right), p_T \left( x_t \mid F^p_{t-1} \right) \right) < C \eta_T.$$
Let $\epsilon_{i,T} := \frac{\log(T)^{\sqrt{T}}}{\sqrt{T}}$. Then the following two conditions hold with probability $1 - 2\delta$ with respect to the prior

\[
\sup_{\epsilon_i \geq \epsilon_{T,i}} \log N \left( \epsilon_i, \{ \xi \in \Xi_T \mid h_{\infty}(\xi, \xi^p) \leq \epsilon_i \} , h_{\infty} \right) \leq T \epsilon_{T,i}^2, \tag{20}
\]

and

\[
Q_T \left( B_T (\xi^P, \epsilon_{T,i}, 2) \mid X_T \right) \geq C \exp \left( -C_0 T \epsilon_{T,i}^2 \right). \tag{21}
\]

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

**Theorem 7** (Contraction Rate of the Transition Density). Let $X_T$ satisfy Assumption 1 and Assumption 2. Denote its density $p_T := \sum_k \Pi_t, k \phi(x_t \mid \mu_t, \Sigma_t)$. Let $T \to \infty$, then the following holds with $\epsilon_T := \sqrt{\frac{\log(T)^2}{T}}$ 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 above and the true transition density satisfies

\[
P_T \left( Q_T \left( \sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} h \left( p_T (x_t \mid \mathcal{F}_{t-1}^P), q_T (x_t \mid \mathcal{F}_{t-1}^Q) \right) \geq C \epsilon_T \mid X_T \right) \right) \to 0.
\]

We also bound the convergence rate of the marginal density. Estimating the Markov transition density with respect to $h_{\infty}$ is strictly harder than estimating the marginal distribution. You can integrate out the marginal distribution using the stationary distribution. (In this context, the stationary and marginal distributions are the same.) A similar argument shows that Proposition 6 implies the following theorem.

**Theorem 8** (Contraction Rate of the Marginal Density). Let $X_T$ satisfy Assumption 1 and assume that the $X_T$ are independent across $t$. Denote its density $p_T := \sum_k \Pi_t, k \phi(x_t \mid \mu_t, \Sigma_t)$. Let $T \to \infty$, then the following holds with $\epsilon_T := \sqrt{\frac{\log(T)^2}{T}}$ and 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 above and the true transition density satisfies

\[
P_T \left( Q_T \left( h (p_T (x_t), q_T (x_t)) \geq C \epsilon_T \mid X_T \right) \right) \to 0.
\]
6 Estimation Strategy

Thus far, the discussion has been rather abstract, focusing on theoretical results. We estimate our model using Bayesian methods by constructing a Gibbs Sampler, which we summarize in Algorithm 1. Recall the definition of the approximating model for the transition density:

\[ q_T(x_t | F_{t-1}) = \sum_{k=1}^{K_T} \Pi(k = \delta_t | \delta_{t-1}) \phi(x_t | \beta_k, x_{t-1}, \Sigma_k). \]  

(22)

We must place a prior on each of the components — \( \delta_t \) — 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 \( \beta_k \) and \( \Sigma_k \).

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 clusters without truncation and compute a valid upper bound for \( K_T \). Conditional on \( K_T \) we draw the \( \delta_t \)s from their marginal distribution. This is straightforward because (22) is a standard Gaussian mixture model conditional on \( \delta_{t-1} \). We update the transition matrix \( \Pi \) so it has the correct marginal distributions and the correct relative transition weights. We then draw the \( \{\delta_t\}_{t=1}^T \). Given \( \delta_t = k \) and the hyperparameters, we apply standard Bayesian regression methods to obtain \( \beta_k \) and \( \Sigma_k \). We use a conditionally conjugate hierarchical prior and draw from the hyperparameters’ posterior. We present the procedure in Algorithm 1.

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

6.1.1 Bounding \( K_T \)

In each period, the approximating model and implied marginal density are Dirichlet mixtures. We draw the cluster identities by adapting existing algorithms. Our problem has the same form as estimating a mixture model in an i.i.d. context except we have a time-varying prior distribution.

Sampling Dirichlet mixtures is difficult for two reasons. First, the prior allows...
Algorithm 1 Gibbs Sampler

1. Posterior of $\{\delta_t\}_{t=1}^T$
   (a) Use Walker (2007) to determine the number of clusters $K_T$.
   (b) Draw the new marginal probabilities, $\pi$, and update the transition matrix, $\Pi$.
   (c) Given $K_T$ and $\{x_t\}_{t=1}^T$, use multinomial sampling to draw $\delta_t$ with
   \[
   \Pr(\delta_t = k) \propto \phi(x_t | \beta_k x_{t-1}, \Sigma_k) \Pi_{t,k}.
   \]

2. Posterior of $\pi$
   (a) Estimate the posterior of $\Pi$ conditional on $\{\delta_t\}_{t=1}^T$:
   \[
   \Pi_{k,j} = \frac{Q_0(\delta_{t-1} = k)Q_0(\delta_t = j) + \sum_{t=2}^T 1(\delta_{t-1} = k)1(\delta_t = j)}{Q_0(\delta_{t-1} = k) + \sum_{t=2}^T 1(\delta_{t-1} = k)}.
   \]

3. Posterior of Component-Specific Parameters
   (a) Given each cluster $k$, use Bayesian regression to draw $\{\beta_k, \Sigma_k\}$.

4. Posterior of Hyperparameters
   (a) Draw the hyperparameters governing $\{\beta_k, \Sigma_k\}$ from their conjugate posteriors.

5. Iterate

for infinitely many clusters, and so we cannot sum the probabilities to compute the resulting marginal cluster probabilities. This inability arises because we cannot numerically solve the probability of cluster $k$: $\Pr(k) = 1 - \sum_{k^{*} \neq k} \Pr(k^{*})$. All Dirichlet mixture models share this property and so several authors have developed ingenious ways to deal with this issue. We adopt the algorithm developed by Walker (2007) because this algorithm is exact (we do not need to truncate the distribution) and computationally efficient. He does this by introducing a random variable — $u_t$ — so that, conditional on $u_t$, the distributions are available in closed form.
Given the cluster parameters, we can write the distribution of $x_t$ as

$$q_T(x_t) = \sum_{k=1}^{\infty} \Pi_{t,k} \phi(x_t | \beta_k x_{t-1}, \Sigma_k). \quad (23)$$

As mentioned above, we introduce a latent variable $u_t \sim U(0, \Pi_{t,k})$ so we can rewrite (23) as

$$q_T(x_t) = \sum_{k=1}^{\infty} 1(u_t < \Pi_{t,k}) \phi(x_t | \beta_k x_{t-1}, \Sigma_k) = \sum_{k=1}^{\infty} \Pi_{t,k} U(u_t | 0, \Pi_{t,k}) \phi(x_t | \beta_k x_{t-1}, \Sigma_k). \quad (24)$$

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

$$\Pr(u_t | \{\Pi_{t,k}\}_{k=1}^{K}) = \sum_{k=1}^{\infty} \Pi_{t,k} U(u_t | 0, \Pi_{t,k}) = \sum_{k=1}^{\infty} 1(u_t < \Pi_{t,k}). \quad (25)$$

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

$$\Pr(\{v_k\}_{k=1}^{K} | \{\delta_t\}_{t=1}^{T}) = Q_0(\{v_k\}_{k=1}^{K}) \prod_{t=1}^{T} 1(v_k = \delta_t) \prod_{\kappa < \delta_t} (1 - v_{\kappa}) > u_{k=\delta_t}), \quad (26)$$

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

The dependence between the $u_t$ does not affect (26) 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 (26); 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 (26) to draw $v_k$. As shown by Papaspiliopoulos and Roberts (2008), this implies $v_k$ are distributed:

$$v_k \sim \text{Beta}\left(1 + \sum_{t=1}^{T} 1(\delta_t = k), T - \sum_{\kappa=1}^{k} \sum_{t=1}^{T} 1(\delta_t = \kappa) + \alpha\right). \quad (27)$$

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

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

(28)

6.1.2 Correcting $\Pi$ to have the Correct Marginal Distribution

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

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

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

(29)

Let $\tilde{\pi}$ be a new marginal distribution that is equivalent (in the measure-theoretic sense) to $\pi$. Define a transition matrix $\tilde{\Pi}$ whose elements satisfy $\tilde{\Pi}_{j,k} = \Pi_{j,k} \frac{\pi_j}{\pi_k} \frac{\pi_k}{\pi_j}$. We now show that $\tilde{\pi}$ is the marginal distribution associated with $\tilde{\Pi}$ by showing it satisfies (29):$^9$

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

(30)

We constructed a matrix $\tilde{\Pi}$ that induces the correct marginal distributions. In doing this, we only changed the marginal distribution. The relative probabilities between different states has not been affected.

To run a Gibbs sampler, we view the operation in (30) as a draw from a conditional posterior. We condition on all but the first left eigenvector (the one associated with the eigenvector 1) of the transition matrix, $\Pi$ and replace it with the one associated

$^8$This condition holding for all $k$ is the standard condition that a stationary distribution is a left-eigenvector of the transition matrix.

$^9$The multiplication and division in (30) is the scalar version.
with $\tilde{\Pi}$. Equivalently, we condition on the stationary distribution of the Markov chain, but not the relative transition probabilities. We then calculate the resulting transition matrix. Transition matrices associated with irreducible Markov chains have exactly one stationary distribution, and that stationary distribution is the first left eigenvector. So this algorithm computes the unique new transition matrix associated with the previous relative transition probabilities and the new marginal distribution.

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

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

From $\tilde{\Pi}$ we can compute $\Pi_{t,k}$ for each $t$ by drawing the first cluster identity, $\delta_0$ from the stationary distribution, and then using the Markov property of $\delta_{t-1}$ for $t > 1$, and iterating forward. We can now compute $\{k \mid \Pi_{t,k} > u_t\}$ for each $t$. Then the posterior of $\delta_t$ is

$$
\Pr(\delta_t = k \mid \Pi_{t,k}, u_t, x_t, \beta_k, \Sigma_k) \propto 1 \left( k \in \{k \mid \Pi_{t,k} > u_t\} \right) \phi(x_t \mid \beta_k x_{t-1}, \Sigma_k). \tag{31}
$$

This is a finite set with known probabilities, and the $\delta_t$ are categorical variables. These can be sampled directly.

### 6.2 Posterior on the Transition Matrix

We place the Dirichlet process prior over these cluster identities in each period to allow for an arbitrary number of clusters. By stacking the Dirichlet processes over time, we obtain a Dirichlet process over the $(\delta_{t-1}, \delta_t)$ product space. Intuitively, we are constructing the transition matrix, $\Pi$, as a Dirichlet-distributed infinite-dimensional square matrix as noted by Lin, Grimson, and Fisher (2010).

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

$$
\Pi_{k,j} = \frac{Q_0(\delta_{t-1} = k)Q_0(\delta_t = j) + \sum_{t=2}^T 1(\delta_{t-1} = k)1(\delta_t = j)}{Q_0(\delta_{t-1} = k) + \sum_{t=2}^T 1(\delta_{t-1} = k)}.
$$

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

### 6.3 Identification Strategy and Cluster Labeling Problem

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

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

In order to enforce this identification restriction, we re-order the cluster identities right before returning the next posterior draw so that they always arrive in time order. This reordering does not solve the identification problem; the data do not identify the cluster labels. It does, however, reduce the amount of multi-modality in the posterior.

### 6.4 Posterior for the Coefficient Parameters

Definition 9 gives the component-specific likelihood where $X_k := \{x_t \mid t - 1 \in \delta_k\}$, $Y_t := \{x_t \mid t \in \delta_k\}$, and $T_k$ is the number of datapoints in cluster $k$. We are factoring the likelihood into the component-specific terms.
\textbf{Definition 9.} Component-Specific Likelihood

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

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

The first level is the standard Gaussian Inverse-Wishart prior that is conjugate to the prior specified in Definition 9.\textsuperscript{10} The only difference is that we parameterize the innovation covariance distribution in terms of its mean: \( \Omega \).\textsuperscript{11} If we need to add a new component during the course of the algorithm we draw from the distribution of \( \beta_k, \Sigma_k \) conditional on the \( \bar{\beta}, U, \Omega, \mu_1 \). We cannot condition on the data in the new component because none exists.

\textbf{Definition 10.} Component-Specific Parameters’ Prior

\[ \left\{ \beta_k \right\}_{k=1}^{K} \mid \Sigma_k, \bar{\beta}, U \sim \mathcal{MN} \left( \bar{\beta}, \Sigma_k, U \right) \] \hspace{1cm} (32)

\[ \left\{ \Sigma_k \right\}_{k=1}^{K} \mid \Omega \sim \mathcal{W}^{-1} \left( (\mu_1 - 2)\Omega, \mu_1 + D - 1 \right) \] \hspace{1cm} (33)

This prior is the conjugate prior for the likelihood in Definition 9, and so we can use the standard formulas to estimate it. This gives the following marginal posterior

\textsuperscript{10} Throughout, we use the parametric formulas given in the Wikipedia pages for the distribution. For example, the Matrix-Normal distribution is parameterized as it is at \url{https://en.wikipedia.org/wiki/Matrix_normal_distribution}.

\textsuperscript{11} The scale parameter and the degrees of freedom parameter are chosen in the appropriate way to make \( \Omega \) the mean matrix: \( \mathbb{E}[\Sigma_k] = \text{Scale}/(\text{Degrees of freedom} - D - 1) = (\mu_1 - 2)\Omega/(\mu_1 + D - 1 - D - 1) = \Omega \).
for the $\Sigma_k$:

$$
\Sigma_k \mid X_k, Y_k \sim \mathcal{W}^{-1}\left(\bar{\beta}'U^{-1}\bar{\beta} + (\mu_1 - 2)\Omega + Y_k'Y_k \right)$$

$$
- \left(U^{-1}\bar{\beta} + X_k'Y_k\right)'\left(U^{-1} + X_k'X_k\right)^{-1}\left(U^{-1}\bar{\beta} + X_k'Y_k\right), \mu_1 + D - 1 + T_k
$$

(34)

We can also compute the following conditional posterior for $\beta_k$ given $\Sigma_k$:

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

(35)

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

**Definition 11. Coefficient Hyperparameters’ Prior**

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

The product of the priors for $\beta_k$’s given in (32) now behaves as the likelihood. Since we have Gaussian priors and likelihoods, this is a fairly standard posterior calculation. The only complication is that the $\{\beta_k\}_{k=1}^K$ are heteroskedastic.\(^{12}\) Consequently, we provide the derivation in Appendix D.2:

$$
U \mid \{\Sigma_k, \beta_k\}_{k=1}^K \sim \mathcal{W}^{-1}\left(\beta^t\beta' + \Psi_U + \sum_{k=1}^K \beta_k\Sigma_k^{-1}\beta_k' - \left(\beta^t + \sum_{k=1}^K \beta_k\Sigma_k^{-1}\right)\left(\sum_{k=1}^K \Sigma_k^{-1} + \mathbb{I}_D\right)^{-1}\left(\beta^t + \sum_{k=1}^K \beta_k\Sigma_k^{-1}\right), \nu_U + (K + 1)D\right),
$$

(36)

\(^{12}\)They must be in order for the prior in (32) to be conjugate with its likelihood because the likelihood is heteroskedastic itself.
and

$$\beta | U, \{\Sigma_k, \beta_k\}_{k=1}^K \sim \mathcal{MN}
$$

$$\left(\left(\beta^2 + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1}\right) \left(\sum_{k=1}^{K} \Sigma_k^{-1} + \mathbb{I}_D\right)^{-1}, \left(\sum_{k=1}^{K} \Sigma_k^{-1} + \mathbb{I}_D\right), U\right).$$

(37)

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

In particular, we parameterize the hierarchy for the $\Sigma_k$ as follows. We have two degree of freedom parameters, $\mu_1$ and $\mu_2$, a mean matrix, $\Omega = \mathbb{E}[\Sigma_k]$, and $D$ scale parameters for $\Omega$: $a_1, \ldots, a_D$.

**Definition 12** (Prior for the Covariances).

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

If we send $\mu_2 \to \infty$, the implied prior for the prior for $\Omega$ becomes fully dogmatic. If $\nu_2 = 1/2$ and $D = 1$, the root diagonal elements — $\sqrt{(\Sigma_k)_{dd}}$ — have half-$t$ distributions. In general, the $(\Sigma_k)_{dd}$ have appropriately scaled $F$-distributions.\(^{13}\) If the off-diagonal elements of $\Omega$ almost surely equal to 0, the diagonal elements satisfy $(\Sigma_k)_{dd} \sim \Gamma^{-1}\left(\mu_1/2, (\mu_1/2 - 1)\Omega_{dd}\right)$. This is why we let the number of degrees of freedom in (33) depend upon $D$. In general, the mean of these elements is the same, but the distribution is different since the off-diagonal elements of $\Omega$ affect the distribution of $(\Sigma_k)_{dd}$.

Obviously, conditional on $\Omega$, everything is independent. The posterior distribution

\(^{13}\) $\sigma^2 \sim F(1, \mu_1 + D - 1) \implies \sigma \sim \text{half-}t(\mu_1 + D - 1)$. In the one dimensional case, $\mu_1 + D - 1 = 1/2$ implies that $\sigma^2 \sim F(1, \mu_1 + D - 1)$. This result is not feasible in the multivariate case while maintaining a density with respect to Lebesgue measure. If we let $\mu_1 \to 2$, we recover this expression. However, $\Omega$ is not well-defined in this case.
of $\Sigma_k$ given $\Omega$, $\{x_t|\delta_t = k\}$ is

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

(38)

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

7 Simulation

7.1 Data

Having characterized our estimators’ theoretical properties, we now consider their behavior in practice. We analyze the performance in simulations to better understand how the approximation works when we know what the true DGP is. The data generating process (DGP) we consider is a vector autoregressive model with the Student’s t-distributed innovations.\footnote{We also conducted simulation experiments with other specifications. These results are available upon request.} The Student’s t-distribution is an infinite mixture of normal distributions where the variance is inverse-gamma distributed. The degrees of freedom for $t$-distributed innovations, which governs the fat-tails of a distribution, are set to be 5.7 as in Brunnermeier, Palia, Sastry, and Sims (2019). Our data generating process of a bivariate data $x_t$ is as follows:
\[ x_t = \Phi_0 + \Phi_1(x_{t-1} - \Phi_0) + \Sigma^{1/2}\epsilon_t \]  
(39)

\[
\Phi_0 = \begin{bmatrix} 0.2 \\ 0.1 \end{bmatrix}, \quad \Phi_1 = \begin{bmatrix} 0.6 & -0.1 \\ 0.0 & 0.9 \end{bmatrix}, \quad \Sigma^{1/2} = \begin{bmatrix} 0.3 & 0.0 \\ 0.2 & 0.3 \end{bmatrix}, \quad \epsilon_{it} \sim_{i.i.d.} t(5.7) \]  
(40)

### 7.2 Prior

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

<table>
<thead>
<tr>
<th>Table 1: Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degrees of freedom for the hierarchical prior</td>
</tr>
<tr>
<td>Expected Number of Components</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Expected Diagonal Autocorrelation</td>
</tr>
<tr>
<td>Expected Off-Diagonal Autocorrelation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component Covariances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ( .25^2 \Pi_d )</td>
</tr>
<tr>
<td>( \mu_1 )</td>
</tr>
<tr>
<td>( \mu_2 )</td>
</tr>
</tbody>
</table>

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. In addition, the theory tells us it will not matter asymptotically. 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.

### 7.3 Simulation Results

We consider the data generating process of VAR(1) with the Student’s t-distributed innovations. The Student’s t-distribution can be considered as an infinite mixture of normal distributions with a common mean but with different precisions.
Figure 2 shows the in-sample predictive posterior density of $x_t$ given $x_{t-1}$. The colored intervals shows the credible set 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 2: One-period Ahead Density Forecasts\textsuperscript{15}

The first row of Figure 3 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 3 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 gray shaded area around the ACF is the credible set drawn using Barlett’s formula. Based on Figure 3, we see that our one-period ahead predictive density is correctly conditionally calibrated.

We can see from Figure 4 that we use more clusters as time progresses. Since the Student’ $t$-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 4, 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 to 12.

\textsuperscript{15}2 std. shock. 5, 50, 95th percentiles.
\textsuperscript{16}2 std. shock. 5, 50, 95th percentiles.
Figure 3: PIT Histogram and Autocorrelation Function (ACF)\textsuperscript{16}

(a) Variable 1
(b) Variable 2

We are not interested in identifying and estimating structural parameters $\Phi_0$, $\Phi_1$ and $\Sigma$. Our goal is to approximate the density closely with a flexible mixture of Gaussians.\textsuperscript{17} Based on our density forecasts, we could obtain the dynamics of higher moments. We obtain the evolution of mean, standard deviation, skewness, and kurtosis based on the rolling average over 12 periods. There arises a fair amount of

\textsuperscript{17}Since the Student’s t-distribution is an infinite mixture of Gaussians with the same mean but with different precisions, the posterior mean over each mixture’s VAR coefficients turns out to be quite close to the truth given that the conditional mean dynamics is well-approximated by our estimated densities.

Figure 4: Number of Clusters Over Time
fluctuation in skewness centered around zero since the underlying t-distribution is symmetric. The kurtosis moves a lot over time capturing the innovations having thick tails.\footnote{A t-distributed random variable with 5.7 degrees of freedom is expected to have a kurtosis equal to 6.5. The time series of rolling average kurtosis over 12 periods (one year) seems to show larger values due to some outliers. When we compute the rolling median kurtosis for the first variable, its value is close to 9. The reason why we just report the values for the first variable is that the second variable is based on the linear combination of two independent t-distributed innovations, which is not t-distributed in general.}

Figure 5: Time-varying Moments from One-period Ahead Density Forecasts

8 Empirical Analysis

8.1 Empirical Data

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). We chose these data series because they are several of the fundamental economic series underlying the macroeconomy, and they span much of the interesting variation.
All of the data are seasonally-adjusted by FRED. We convert to approximate percent changes by log-differencing all of the data except for the consumption measure, which is already measured in percent changes, the 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 covers the January 1963 to December 2018. The time dimension is 671, and the cross-sectional dimension is 6. Figure 6 shows the standardized monthly macroeconomic data used in this subsection. The gray bars denote NBER recessions.

Figure 6: Monthly Macroeconomic Data

8.2 Prior

We use the prior as in Table 1, which is the same used in the simulation section. 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 and empirical results are different, this is reflection of the dynamics of the datasets.

8.3 Monthly Macroeconomic Series

To show that our algorithm works reasonably well in practice, we display the conditional density forecast for consumption in Figure 7. Predictive densities and PIT’s for the other series are provided in Figure 10. If the model works perfectly, the probability integral transform (PIT) should be independent and distributed $U[0,1]$. As we can see, it is roughly independent and distributed approximately uniformly.

The dynamics of the data in Figure 7a are not obviously non-Gaussian or non-linear. Are we effectively just estimating a VAR? No. Figure 8a shows that the
conditional variance spikes a great deal in recessions when we compute the rolling averages over 1 year. Similar to Schorfheide, Song, and Yaron (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 on higher-moments of consumption at business cycle frequencies. Skewness (Figure 8b) and kurtosis (Figure 8c) exhibit significant time-variation. Interestingly, skewness appears to decrease and kurtosis to increase during NBER recessions.

What drives the time-variation in these moments? We can divide the conditional variance in each period into two components using the law of total variance:

$$\text{Var} (x_t | \mathcal{F}_{t-1}) = \text{Var} \left( \mathbb{E}[x_t | \delta_t] | \mathcal{F}_{t-1} \right) + \mathbb{E} \left[ \text{Var} (x_t | \delta_t) | \mathcal{F}_{t-1} \right].$$  \hspace{1cm} (41)

Since the model is linear conditional on the cluster identity $\delta_t$, the first term comes from variation in $\beta_k x_{t-1}$, while the second arises from variation in the innovations. Figure 9d shows the volatility associated with autoregressive coefficients, whereas...
Figure 9b shows the volatility associated with innovations. The total variance, which Figure 8a reports for consumption, is the sum of the two. Comparing these two volatilities, we observe bigger changes in dynamics for the coefficient volatility. This implies that the stochastic volatility in macroeconomic data studied in papers such as Fernández-Villaverde and Rubio-Ramírez (2010) and Fernández-Villaverde, Guerrón-Quintana, Kuester, and Rubio-Ramírez (2015) can be more parsimoniously modeled using variation in the conditional mean than by using stochastic volatility.

Figure 9c shows the mixture probability of the first cluster in each period. From the empirical results, we see that 5 clusters become active in our sample but the mixture probability of the first cluster is very high. Hence, our model is very parsimonious, which we did not impose. We can also see that the mixture probabilities fluctuate at the monthly frequency. Specifically, the transition probability matrix of cluster identities is

\[
\begin{bmatrix}
0.994 & 0.003 & 0.002 & 0.000 & 0.000 \\
0.438 & 0.323 & 0.156 & 0.065 & 0.027 \\
0.324 & 0.167 & 0.399 & 0.074 & 0.035 \\
0.258 & 0.244 & 0.163 & 0.278 & 0.058 \\
0.348 & 0.180 & 0.271 & 0.109 & 0.093 \\
\end{bmatrix},
\]

and the unconditional cumulative probability distribution of cluster identities is

\[
\begin{bmatrix}
0.991 & 0.995 & 0.998 & 0.000 & 0.000 \\
\end{bmatrix},
\]

The first component has extremely high unconditional probability in (43). As asked above, are we essentially just estimating a VAR(1). No. Arguably the closest related models in the literature are regime-switching models. Consider lumping the clusters into two “regimes”, the first cluster into a “normal-times” regime, and the remaining clusters into a “recession” regime. If we’re not in the first cluster, the probability of entering the first cluster next period falls below 50% at the monthly frequency. For example, conditional on being in the second cluster, the probability of transitioning from the third cluster to the first cluster even lower at 32.4% and from the fourth cluster lower still 25.8%. Since these data are at the monthly frequency, if we view the first cluster as a normal-times regime, then the probability of entering the other bins (the recession
regime) is quite low. This should be expected because forecasting recessions is quite difficult. However, conditional on being in a recession, staying there is actually quite likely.

Figure 9: Empirical Results with Monthly Macroeconomic Series

(a) Mean

(b) Innovation Volatility

(c) Mixture Probability of the First Cluster

(d) Coefficient Volatility

The key difference between our model and a regime-switching model is that our model is more flexible. Our nonparametric approach uses an endogenously-determined number of components to fit the recession regime, instead of just one as a standard regime-switching model would do. The is a consequence of our clusters serving two purposes. They let the mean change, as you would in a regime-switching model, but they also model non-Gaussianity. Recall that a Student’s $t$-distribution is an infinite mixture of Gaussians. For example Figure 9d shows that the volatility of the coefficients spikes dramatically during recessions, especially 2008. Consequently, our model implies that the data become less Gaussian during recessions, and so the model uses more components to fit this non-Gaussian distribution.

The claim that some of the data become less Gaussian during recessions, should not be too surprising. However, our nonparametric, multivariate approach finds non-Gaussianity and increased volatility in all of our variables, including consumption. In other words, we find that the data are substantially less Gaussian during reces-
sions, and this increase in the distributional complexity holds for all the series considered. Since recessions are rare and the data are close to Gaussian during non-recession periods, this means that the data are unconditionally close to Gaussian.

This aligns with the recent literature and has important consequences for the types of models macroeconomics and finance researchers should use. Recently, Guvenen, Ozkan, and Song (2014) point out that the left-skewness of income risk is strongly counter-cyclical. That is, income shocks become more risky during recessions. Our empirical analysis shows that the change in consumption is also left-skewed in worse economic times. Furthermore, the joint evolution of kurtosis and coefficient volatility shows that the consumption density becomes more fat-tailed in recessions, even conditioning on the time-varying innovation volatility. 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. Lastly, the volatility and the kurtosis of consumption fluctuated more in magnitude in the 70s and 80s than it did in the period after 2000.

9 Conclusion

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

We provide new theory that shows, under some general assumptions, the posterior distribution of our estimators converges more rapidly than the previous literature has been able to achieve with arbitrarily high probability. In particular, we exploit the tail behavior of probability distributions in high dimensions to show that our estimator for the marginal densities converges at a $\sqrt{\log(T)/T}$ rate and our estimator for the transition densities converge at a $\log(T)/\sqrt{T}$ rate with high probability. These rates are noteworthy because they are the parametric rate up to a logarithmic term. They are remarkable because they only depend on the number of series through the constant term.
We show that this estimation strategy performs well in simulations and when applied to macroeconomic data. In the empirical applications, we show that macroeconomic data’s dynamics are often far from Gaussian and the dynamic structure moves across the business cycle. We further find that our proposed representation requires more than one mixture component, but only a few, to handle the data’s dynamics well.

**References**


Appendix A  Conditional Forecasts: The Other Variables

Figure 10: One-Period Ahead Conditional Forecasts

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

(d) Housing Supply Posterior Density

(g) Industrial Production Posterior Density

(j) Unemployment Rate Posterior Density

(m) PCE Posterior Density

(n) PIT Histogram  (o) PIT ACF
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 $\gamma$-functional:

$$
\gamma_\alpha(X, d) = \inf_{x \in X} \sup_{s=0}^{\infty} 2^{s/\alpha} d(s, X_s), \tag{44}
$$

where the infimum is taken with respect to all subsets $X_s \subset X \subset \mathbb{R}^{T \times D}$ such that the cardinality $|X_s| \leq 2^{2^s}$, $|X_0| = 1$, and $d$ is a metric. This $\gamma_2(X, d)$ 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 $n$: $\psi_n := \inf_{C > 0} \mathbb{E} \left[ \exp \left( \frac{|X|^n}{C^n} - 1 \right) \right]$. This is useful because a standard argument shows that if $X$ has a bounded $\psi_n$ norm then the tail of $X$ decays faster than $2 \exp \left( -\frac{s^n}{\|x\|_{\psi_n}} \right)$. Hence, if $x$ has a finite $\psi_2$-norm, it is subgaussian.

A.2  Definition and Properties of the $\Theta_T$-operator

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

$$
\Pr(K \leq \tilde{K}) = \left( 1 - (1 - \mu)^{\tilde{K}} \right)^T \tag{45}
$$

**Proof.** Let $\theta_t$ denote a row of $\Theta_T$. Then

$$
\Pr(K \leq \tilde{K}) = \Pr(1 \in \theta_t \text{ for all } t) = \Pr(1 \in \theta_t)^T = (1 - \Pr(1 \not\in \theta_t))^T = \left( 1 - (1 - \mu)^{\tilde{K}} \right)^T. \tag{46}
$$

**Lemma 4.** Let $K$ be the number of columns of $\Theta_T$ as defined in Definition 3. There exists a constant $\gamma \in (0, 1)$ and constants $c_1, c_2$, such that with probability at least $\gamma$, the following holds.

$$
c_1 \log(T) \leq K \leq c_2 \log(T) \tag{47}
$$
Proof. Let $B := \exp(\tilde{K})$ We set the cumulative distribution function equal to $1 - \gamma$, i.e. the survival function equal to $\gamma$:

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

(48)

Note, for positive $a$ and $b$, $a^{\log(b)} = b^{\log(a)}$.

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

(49)

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

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

(50)

This implies

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

(51)

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

\[\Box\]

A.3 Relationship between the Orlicz and $L_2$ norms.

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

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

1. The squared $L_2$-norm of $x$ is equivalent to $\mathbb{E}[(\Theta_k, x)^2]$.

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 $\Theta_k$ denote a column of the matrix. The root of the proof follows from realizing that $\Theta_T$ is a generalized selection matrix, and covariances are dominated by variances:

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

(52)
where the last line follows by the independence of the rows of $\Theta_k$.

Consider $E_\Theta [X'\Theta'X]$. Since the columns of $\Theta_T$ are a martingale difference sequence, variances of sums are sums of variances:

$$E_\Theta [X'\Theta'X] = \sum_{k=1}^{K} E_{\Theta_k} [X'\Theta_k'\Theta_k X] = \sum_{t=1}^{T} x_t x'_t. \quad (53)$$

Now that we have shown Item 1, we must show that $L_2$-norm dominates the $\psi_2$-norm. This is useful because it implies that if we can control the variance of the distribution, we automatically control the tails as well:

$$\inf \left\{ C > 0 \mid \mathbb{E} \left[ \exp \left( \frac{|\langle \Theta_k, x \rangle|^2}{C^2} \right) \right] - 1 \leq 1 \right\} \quad (54)$$

$$= \inf \left\{ C > 0 \mid \mathbb{E} \left[ \exp \left( \frac{\sum_{t=1}^{T} |\theta_{t,k}| x'_t x_t + 2 \sum_{t,\tau \neq t} \theta_{t,k} \theta_{\tau,k} x'_t x'_\tau}{C^2} \right) \right] \leq 2 \right\}. \quad (55)$$

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

$$\inf \left\{ C > 0 \mid \mathbb{E} \left[ \exp \left( \frac{2 \sum_{t=1}^{T} |\theta_{t,k}| x'_t x_t}{C^2} \right) \right] \leq 2 \right\}. \quad (56)$$

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

$$\inf \left\{ C > 0 \mid \mathbb{E} \left[ \sum_{h=0}^{\infty} \frac{2^h \left( \sum_{t=1}^{T} |\theta_{t,k}| x'_t x_t \right)^h}{C^{2h} h!} \right] \leq 2 \right\} \quad (57)$$

$$= \inf \left\{ C > 0 \mid \mathbb{E} \left[ \sum_{h=0}^{\infty} \frac{2^h \prod_{h} |\sum_{k_1=0}^{h} (k_1, k_2, \ldots, k_T) \prod_{t=1}^{T} \theta_{t,k} | (\sum_{t=1}^{T} |\theta_{t,k}| x'_t x_t)^{k_t}}{C^{2h} h!} \right] \leq 2 \right\}. \quad (58)$$

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

$$\inf \left\{ C > 0 \mid \sum_{h=0}^{\infty} \frac{2^h \prod_{h} |\sum_{k_1=0}^{h} (k_1, k_2, \ldots, k_T) \prod_{t=1}^{T} \frac{1}{K} (x'_t x_t)^{k_t}}{C^{2h} h!} \leq 2 \right\}. \quad (59)$$
Then we can use the multinomial theorem and the formula for the exponential function in the reverse direction, implying this equals
\[
\inf \left\{ C > 0 \left| \frac{1}{K} \exp \left( \frac{2\|x\|^2_{L^2}}{C^2} \right) \leq 2 \right. \right\} = \inf \left\{ C > 0 \left| \frac{2\|x\|^2_{L^2}}{C^2} = \log (2K) \right. \right\} \leq \frac{\sqrt{2}\|x\|_{L^2}}{\log (2)},
\]
where the last inequality follows because \( K \geq 1 \). Hence, we have that the \( L^2 \)-norm dominates the \( \psi_2 \)-norm.

\[\square\]

### A.4 Norm Equivalence

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

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

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

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

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

\[
\Pr (|W_x - W_y| > ud(x, y)) < \exp (-\eta Ku^2)
\]

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

4. \( \gamma \) is increasing, differentiable at zero and \( \gamma'(0) > 0 \).

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

\[
\sup_{x \in \mathcal{X}} |Z_x| < C(\gamma, \delta, \eta)\epsilon.
\]
Here we quote a version of Bernstein’s inequality for martingales due to (de la Peña, 1999, Theorem 1.2A), which we use later.

**Theorem 10** (Bernstein’s Inequality for Martingales). Let \(\{x_i, F_i\}\) be a martingale difference sequence with \(\mathbb{E}[x_i | F_{i-1}] = 0, \mathbb{E}[x_i^2 | F_{i-1}] = \sigma_i^2, v_k = \sum_{i=1}^k \sigma_i^2\). Furthermore, assume that \(\mathbb{E}[|x_i|^n | F_{i-1}] \leq \frac{n!}{2} \sigma_i^2 M^{n-2}\) almost everywhere. Then, for all \(x, y > 0\),

\[
\Pr \left( \left\{ \left| \sum_{i=1}^k x_i \right| \geq u, v_k \leq y \text{ for some } k \right\} \right) \geq 2 \exp \left( -\frac{u^2}{2(y + uM)} \right). \tag{61}
\]

If we choose \(c\) small enough, this implies

\[
\Pr \left( \left\{ \left| \frac{1}{k} \sum_{i=1}^k x_i \right| \geq u, v_k \leq y \text{ for some } k \right\} \right) \geq 2 \exp \left( -c \min \left\{ \frac{u^2k^2}{v}, \frac{uk}{M} \right\} \right). \tag{62}
\]

**Theorem 1** (Bounding the Norm Perturbation). Let \(\Theta_T\) be constructed as in Definition 3 with the number of columns denoted by \(K_T\). Let \(\epsilon > 0\) be given. Let \(0 < \delta < 1/2\) be given such that \(0 < \log(\frac{1}{\delta}) < c_1 \epsilon^2 K_T\) for some constant \(c_1\). Let \(\bar{X}_T\) be in the unit hypersphere in \(\mathbb{R}^{TD-1}\). Then with probability greater than \(1 - 2\delta\) with respect to \(\Theta_T\), there exists a constant \(c_2\) such that for any \(\epsilon > \sqrt{\frac{\log T}{K_T}}\),

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

**Proof.** We mimic the proof of (Klartag and Mendelson, 2005, Theorem 3.1), verifying the conditions of Proposition 9. Similar to them we use \(\Upsilon(t) := \sqrt{1 - t}\). Our conclusion is stated in terms of the logarithm of the sample size \(- T\). This conclusion is weaker than theirs as \(\gamma_2(\bar{X}, \|\cdot\|_{L_2}) < C \sqrt{\log(T)}\). 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 \(x, y \in X\). We use the functional notation \(x(\theta_k)\) to refer \(\sum_{d=1}^D \theta_k^d x_d\).

\[
Z_x^K := \frac{1}{K} \sum_{k=1}^K x^2(\theta_k) - \left\| x \right\|_{L_2}^2 \tag{63}
\]

Consider \(Z_x^K - Z_y^K\).
\[ Z^K_x - Z^K_y = \frac{1}{K} \sum_{k=1}^{K} x^2(\theta_k) - y^2(\theta_k) = \frac{1}{K} \sum_{k=1}^{K} (x-y)(\theta_k)(x+y)(\theta_k) \]  \hspace{1cm} (64)

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

\[
\Pr(|Y_k| > 4\|x - y\|_{\psi_2}\|x + y\|_{\psi_2}) \leq \Pr(|x(\theta_k) - y(\theta_k)| > 2\sqrt{u}\|x - y\|_{\psi_2}) + \Pr(|x(\theta_k) + y(\theta_k)| > 2\sqrt{u}\|x + y\|_{\psi_2}) \leq 2\exp(-u),
\]

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 \( Y_k \) used by Klartag and Mendelson because the entries in our \( \theta \) operator are uniformly bounded by 1 in absolute value.

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

\[
\Pr \left( \left\{ \frac{1}{K} \sum_{k=1}^{K} \theta_k > u \right\} \right) \leq 2 \exp \left( -cK \min \left\{ \frac{u^2}{\|\theta\|_{\psi_1}^2}, \frac{u}{\|\theta\|_{\psi_1}} \right\} \right) \hspace{1cm} (66)
\]

Then by applying (66) to \( \Pr(|z^K_x - z^K_y| > u) \), we have the following.

\[
\Pr \left( |Z^K_x - Z^K_y| > u \right) \leq 2 \exp \left( -c \min \left\{ \frac{u^2}{\|x - y\|_{L_2}^2}, \frac{u}{\|x - y\|_{L_2}} \right\} \right) \hspace{1cm} (67)
\]

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

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

\[
\Pr \left( |Z^K_x - Z^K_y| > 2\|x - y\|_{L_2} \right) \leq 2 \exp \left( \eta K \min \{u, u^2\} \right) < \exp \left( -\eta K \frac{u^2}{\delta_0} \right). \hspace{1cm} (68)
\]
By the triangle inequality,

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

(69)

Applying (66) for $u > 1$:

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

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

For any $x \in \mathcal{X}$ by (66),

$$\text{Pr}(|Z_x| > \epsilon) < \exp(-\eta K \epsilon^2) < \delta. \tag{71}$$

We can bound the derivative of $\Upsilon$:

$$\Upsilon'(0) = 1/2 > 0. \tag{72}$$

□

Online Appendix B  Representation Theory

B.1 The Joint Density

Lemma 6 (Bounding 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.,

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

Proof. Clearly, if $\#t = 1$, the result holds. Assume $\#t = 2$. Assume the claim is false. Then
\[
\frac{x_1 + x_2}{y_1 + y_2} > \max \left\{ \frac{x_1}{y_1}, \frac{x_2}{y_2} \right\} \quad \Rightarrow \quad x_1 + x_2 > \max \left\{ \frac{x_1 y_2}{y_1}, x_2 + \frac{x_2 y_1}{y_2} \right\} \quad (73)
\]

\[
\Rightarrow x_1 > \frac{x_2 y_1}{y_2} \quad \text{and} \quad x_2 > \frac{x_1 y_2}{y_1} \quad \Rightarrow \quad x_1 > \frac{y_1 x_1 y_2}{y_2} \quad \Rightarrow \quad x_1 > x_1.
\]

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

\[
\sum_{t} x_t \leq \max \left\{ \frac{\sum_{t \neq T} x_t}{\sum_{t \neq T} y_t} \right\} \leq \ldots \leq \max \left\{ \frac{x_t}{y_t} \right\}, \quad (74)
\]

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

**Lemma 7.** Consider the ratio of the densities between \(p_T\) and \(q_T\). Let \(\delta_k^q\) be a clustering of \(x_t\) with respect to \(q_T\). Let these clusters \(\delta_k^q\) satisfy the following, where \(\mu_k^q = E_{P_T} [x_t \mid t \in \delta_k^q]\) and \(\Sigma_k^q = \text{Cov}_{P_T} [x_t \mid x_t \in \delta_k^q]\):

\[
\sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - (x_t' - \mu_t')' \Sigma_k^{-1} (x_t' - \mu_t') \right| < C \epsilon. \quad (75)
\]

Then the log-divergence satisfies

\[
\sup_{x_t, x_t'} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - (x_t' - \mu_t')' \Sigma_k^{-1} (x_t' - \mu_t') \right| < C \epsilon \quad \Rightarrow \quad \sup_{x_t, x_t'} \left| \log \left( \frac{p_T(x_t)}{p_T(x_t')} \right) \right| < C \epsilon. \quad (76)
\]

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

\[
\sup_{\delta_k^q} \sup_{x_t \in \delta_k^q} \left| (x_t - \mu_t)' \Sigma_t^{-1} (x_t - \mu_t) - (x_t' - \mu_k^q)' \left( \Sigma_k^q \right)^{-1} (x_t' - \mu_k^q) \right| < C \epsilon. \quad (77)
\]

Consider the ratio of the proportionality constants \(\chi^p\) and \(\chi^q\) associated with the kernels \(k^p, k^q\) above:

\[
\chi^p = \int_{\mathcal{X}} k^p(x) \, dx, \quad \chi^q = \int_{\mathcal{X}} k^q(x) \, dx. \quad (78)
\]
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 6, this is bounded by the supremum of the ratios, since we are integrating over the same space in both sums:
\[
\leq \sup_x \log \left( \frac{k^q(x)/p_T(x)}{k^p(x)/p_T(x)} \right) \leq \sup_x \log \left( \frac{k^q(x)}{k^p(x)} \right),
\]
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
\[
\left| \log \frac{\mu_1}{\mu_2} \right| < \frac{1}{2} C \epsilon
\]
because the \(k'(x)\) are Gaussian kernels, and we bounded the log-ratio in (77). The total deviation is the sum of the deviation in the constants and in the kernels. The result holds by combining (81) and (77).

\(\blacksquare\)

**Proposition 11** (Bounding the Supremum of the Rescaled Data). Let \(\tilde{X} := \tilde{x}_1, \ldots, \tilde{x}_T\) be a D-dimensional mixed Gaussian process with finite stochastic means \(\mu_t\) and co-variances \(\Sigma_t\), where \(\Sigma_t\) is positive-definite for all \(t\). Let \(\Theta_T\) be the generalized selection matrix defined in Definition 3. Let \(\tilde{P}_T\) denote the distribution of \(\tilde{X}\). Then given \(\epsilon > 0\) and for some \(\delta \in (0, 1/2)\), the approximating distribution \(\tilde{Q}_T\), which is the mixture distribution over \(\{\tilde{\Sigma}_t^{-1/2}\tilde{x}_1\}_{t=1}^T\) defined by the clustering induced by \(\Theta_T\) satisfies the following with probability at least \(1 - 2\delta\) with respect to \(\Theta_T\).
\[
\sup_t h^2 \left( \int_{G_t} \phi (\tilde{x}_t \mid \delta^P_t) \, dG^P_t (\delta^P_t), \int_{G_t} \phi (\tilde{x}_t \mid \delta^Q_t) \, dG^Q_t (\delta^Q_t) \right) < c \left( 1 + \log \frac{1}{\delta} \right)^2 \epsilon^2
\]

**Proof.** In this proof, we drop the tilde’s over the \(x_t\) because all of the terms have them. Let \(G^P\) and \(G^Q\) be the associated mixing measures of the covariances. Let \(K\) be a coupling from between the space of \(G^P\) and \(G^Q\). Consider the supremum of the squared Hellinger distance — \(h^2\) — between \(P_T\) and \(Q_T\):
\[ \sup_t h^2 \left( \int_{G_t} \phi \left( x_t \mid \delta^P_t \right) dG_t^P(\delta^P_t), \int_{G_t} \phi \left( x_t \mid \delta^Q_t \right) dG_t^Q(\delta^Q_t) \right). \]  

(83)

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

\[ \leq \int_{G_t^P \times G_t^Q} \sup_t h^2 \left( \phi \left( x_t \mid \delta^P_t \right), \phi \left( x_t \mid \delta^Q_t \right) \right) dK(G_t^P, G_t^Q). \]  

(84)

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

\[ \leq \int_{G_t^P \times G_t^Q} \sup_t \int_{\mathbb{R}^D} \left| \left( \frac{\phi \left( x_t \mid \delta^P_t \right)}{\phi \left( x_t \mid \delta^Q_t \right)} \right)^{1/2} - 1 \right|^2 d\Phi \left( x_t \mid \delta^Q_t \right) dK(G_t^P, G_t^Q). \]  

(85)

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

\[ = \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} \sup_t \left| \left( \frac{\phi \left( x_t \mid \delta^P_t \right)}{\phi \left( x_t \mid \delta^Q_t \right)} \right)^{1/2} - 1 \right|^2 d\Phi \left( x_t \mid \delta^Q_t \right) dK(G_t^P, G_t^Q). \]  

(86)

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

\[ \leq C_1 \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} \sup_t \left| (x_t - \mu^P_t)\Sigma^P_t(x_t - \mu^P_t) - (x_t - \mu^Q_t)\Sigma^Q_t(x_t - \mu^Q_t) \right| d\Phi \left( x_t \mid \delta^Q_t \right) dK(G_t^P, G_t^Q). \]  

(87)

Since \(Q_T\) was defined through applying \(\Theta_T\) to \((\Sigma^P_t)^{-1/2}(x_t - \mu^P_t)\), by Theorem 1 this norm perturbation is bounded by \(\epsilon^2\); we just have to square the constant:

\[ \leq C \left( 1 + \log \frac{1}{\delta} \right)^2 \int_{G_t^P \times G_t^Q} \int_{\mathbb{R}^D} |\epsilon|^2 d\Phi \left( x_t \mid \delta^Q_t \right) dK(G_t^P, G_t^Q) = C \left( 1 + \log \frac{1}{\delta} \right)^2 \epsilon^2, \]  

(88)

where the last equality holds because all of the integrals integrate to 1.
Theorem 2 (Representing the Joint Density). Let $\tilde{X}_T := \frac{X_T - \mu_T}{\sqrt{\|X_T - \mu_T\|_2}}$ where $X_T$ satisfies Assumption 1. Let $\Theta_T$ be the generalized selection matrix constructed in Definition 3. Let $P_T$ denote the distribution of $\tilde{X}_T$. Then given $\epsilon > 0$ and $\delta \in (0, \frac{1}{2})$, the approximating distribution, $Q_T$, which is the mixture distribution over $\tilde{X}$ that $\Theta_T$ induces, satisfies the following with probability at least $1 - 2\delta$ with respect to $\Theta_T$ for some constant $C$:

$$h_\infty \left( P_T(\tilde{X}), Q_T(\tilde{X}) \right) < C \left( 1 + \log \left( \frac{1}{\delta} \right) \right) \epsilon.$$

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

We can represent both densities succinctly as follows. Importantly, we do not require that the $G^P_t$ are independent:

$$p_T(\tilde{X}) = \int_G \int_{G_t} \phi \left( x_t \mid \delta^P_t \right) dG^P_t \left( \delta^P_t \right) dG^P \left( dG^P_t \right). \quad (89)$$

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

$$q_T(\tilde{X}) = \int_G \int_{G_t} \phi \left( x_t \mid \delta^Q_t \right) dG^Q_t \left( \delta^Q_t \right) dG^Q \left( dG^Q_t \right). \quad (90)$$

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

$$h^2_\infty \left( p_T(\tilde{X}), q_T(\tilde{X}) \right) \quad (91)$$

$$= h^2_\infty \left( \int \int \phi \left( x_t \mid \delta^P_t \right) dG^P_t \left( \delta^P_t \right) dG^P \left( dG^P_t \right), \int \int \phi \left( x_t \mid \delta^Q_t \right) dG^Q_t \left( \delta^Q_t \right) dG^Q \left( dG^Q_t \right) \right).$$
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_\infty^2 \left( \int_G \int_{G_t} \phi \left( x_t \mid \delta_t^P \right) dG^P_t(dG^P_t, dG^Q_t), \right.

\left. \int_G \int_{G_t} \phi \left( x_t \mid \delta_t^Q \right) dG^Q_t(dG^P_t, dG^Q_t) \right).
$$

(92)

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

$$
\leq \int_{G \times G} \sup_t h^2 \left( \int_{G_t} \phi \left( x_t \mid \delta_t^P \right) dG_t^P(d\delta^P_t), \int_{G_t} \phi \left( x_t \mid \delta_t^Q \right) dG_t^Q(d\delta^Q_t) \right) d\mathcal{K}(dG^P_t, dG^Q_t).
$$

(93)

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

We can place the bound obtained in Proposition 11 inside (93). Since we are integrating $C\epsilon^2$ over a joint density, the density is bounded above by 1, and we are done.

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

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

(94)

Lemma 8. Let $f, g$ be two densities of locally asymptotically mixed normal (LAMN) processes with respect to the sample size $T$.\textsuperscript{19} Squared Hellinger distance and Kullback-Leibler divergence are equivalent.

Proof. Consider the following decomposition of the Hellinger distance:

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

(95)

\textsuperscript{19}This trivially covers all Gaussian processes with finite-means and variances.
Taking a Taylor expansion of the exponential function:

\[
\begin{align*}
&= \int \left( 1 + \frac{1}{2} \log \left( \frac{f}{g} \right) + O \left( \log \left( \frac{f}{g} \right)^2 \right) - 1 \right) dG \\
&= \int \frac{1}{2} \log \left( \frac{f}{g} \right) dG + O \left( \int \log \left( \frac{f}{g} \right)^2 dG \right). 
\end{align*}
\]

(96)

(97)

Consider one-half the Kullback-Leibler divergence:

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

(98)

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

\[
\begin{align*}
&= \frac{1}{2} \int \log \left( \frac{f}{g} \right) \left( 1 + \log \left( \frac{f}{g} \right) \right) dG = \frac{1}{2} \int \log \left( \frac{f}{g} \right) dG + O \left( \int \log \left( \frac{f}{g} \right) \log \left( \frac{f}{g} \right) dG \right). 
\end{align*}
\]

(99)

The first terms in (96) and (99) 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 \( \Sigma \) 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:

\[
\int \log(f/g)^2 dG \leq \int |z|^2 dG + O(\epsilon) \leq \int |z| dG + O(\epsilon) = \int \log(f/g) dG + O(\epsilon), 
\]

(100)

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 (96) and (99), the last term in (100) is bounded by both the Hellinger and Kullback-Leibler divergences.

\[ \square \]

**B.2 Representing the Marginal Density**

**Theorem 3** (Representing the Marginal Density). Let \( X_T \) satisfy Assumption 1 and assume that the \( X_T \) are independent across \( t \). Let \( \Theta_T \) be constructed as in Definition 3. Let \( \epsilon > 0, \delta \in (0, 1/2) \) be given. Construct \( Q_T \) as the mixture model in Definition 2
where $\Theta_T$ groups the data into components. Then, with probability $1 - 2\delta$ with respect to $\Theta_T$, there exists a constant $C$ such that the following holds uniformly over $T$

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

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

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

$$\prod_{t=1}^{T} \int_{G_t} \phi \left( x_t \mid \delta_t \right) dG_t(\delta_t) = \prod_{t=1}^{T} \int_{G_m} \phi \left( x_t \mid \delta_t \right) dG_m(\delta_t). \quad (101)$$

All (101) is saying is that the joint $T$ independent draws from the marginal are the same as $T$ independent draws from a sequence $G_1, \ldots, G_T$, which is drawn from $G$. By assumption $G$ has a product form. The Kullback-Leibler divergence between the two joint distributions is

$$D_{KL} (q_T \mid \mid p_T) = \int_{R^T \times D} \log \left( \frac{q_T}{p_T} \right) dP_T = \int_{R^T \times D} \log \left( \frac{\prod_{t=1}^{T} \int_{G_t} \phi \left( x_t \mid \delta_t^Q \right) dG_t(\delta_t^Q)}{\prod_{t=1}^{T} \int_{G_t} \phi \left( x_t \mid \delta_t^P \right) dG_t(\delta_t^P)} \right) dP_T. \quad (102)$$

Ratios of products are products of ratios, and logs of products are sums of logs, and we can substitute in the definition of the marginal distribution, (101), giving

$$= \int_{R^T \times D} \sum_{t=1}^{T} \log \left( \frac{\int_{G_m} \phi \left( x_t \mid \delta_t^Q \right) dG_m(\delta_t^Q)}{\int_{G_m} \phi \left( x_t \mid \delta_t^P \right) dG_m(\delta_t^P)} \right) dP_T. \quad (103)$$
We can rewrite \( P_T \) in terms of its mixture representation:

\[
\int_{G_t} \int_{\mathbb{R}^D} \sum_{t=1}^{T} \log \left( \frac{\int_{G_m} \phi \left( x_t \mid \delta^Q_t \right) dG^Q_t(\delta^Q_t)}{\int_{G_m} \phi \left( x_t \mid \delta^P_t \right) dG^P_t(\delta^P_t)} \right) \prod_{t=1}^{T} \phi \left( x_t \mid \delta_t \right) dx \ dG_m(\delta_t). \tag{104}
\]

The only interactions between the two terms are the \( x_t \):

\[
= \sum_{t=1}^{T} \left( \int_{G_t} \int_{\mathbb{R}^D} \log \left( \frac{\int_{G_m} \phi \left( x_t \mid \delta^Q_t \right) dG^Q_t(\delta^Q_t)}{\int_{G_m} \phi \left( x_t \mid \delta^P_t \right) dG^P_t(\delta^P_t)} \right) \phi \left( x_t \mid \delta_t \right) dx \ dG_m(\delta_t) \right) \tag{105}
\]

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

\[
= \sum_{t=1}^{T} \left( \int_{G_t} \int_{\mathbb{R}^D} \log \left( \frac{\int_{G_m} \phi \left( x_t \mid \delta^P_t \right) dG^P_t(\delta^P_t)}{\int_{G_m} \phi \left( x_t \mid \delta^Q_t \right) dG^Q_t(\delta^Q_t)} \right) \phi \left( x_t \mid \delta_t \right) dx \ dG_m(\delta_t) \right). \tag{106}
\]

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

\[
= \sum_{t=1}^{T} D_{\text{KL}} \left( \int_{G_m} \phi \left( x_t \mid \delta^Q_t \right) dG^Q_m(\delta^Q_t) \bigg\| \int_{G_m} \phi \left( x_t \mid \delta^P_t \right) dG^P_m(\delta^P_t) \right) \tag{107}
\]

\[
= TD_{\text{KL}} \left( \int_{G_m} \phi \left( x_t \mid \delta^Q_t \right) dG^Q_m(\delta^Q_t) \bigg\| \int_{G_m} \phi \left( x_t \mid \delta^P_t \right) dG^P_m(\delta^P_t) \right). \tag{108}
\]

In other words, the distance between the joint densities is at least \( T \) times the distance between the distance marginal densities. Also, by Lemma 8 this is proportional to squared Hellinger distance. In other words, the difference between the joint densities is at least \( T \) times the distance between the distance between the marginal densities. We know by Theorem 2 that this is bounded above by \( CT^2 \varepsilon^2 \). The \( T \) arises because we are no longer using the rescaled data, and \( \|X\|^2 \propto T \). This gives

\[
h^2 \left( \int_{G_m} \phi \left( x_t \mid \delta^Q_t \right) dG^Q_m(\delta^Q_t), \int_{G_m} \phi \left( x_t \mid \delta^P_t \right) dG^P_m(\delta^P_t) \right) \leq \frac{1}{T} h^2(q_T, p_T) \leq C \frac{T}{T} \varepsilon^2 = C \varepsilon^2. \tag{109}
\]

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

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

\[ \square \]

B.3 Representing the Transition Density

Theorem 4 (Transition Density Representation). Let \( X_T \) satisfy Assumption 1 and Assumption 2. Let \( p_T \) denote the true density. Let \( \epsilon > 0, \delta \in (0, 1/2) \) be given. Let \( \Theta_T \) be constructed as in Definition 3. Let \( K = C( \text{number of columns of } (\Theta_T)^2 ) \) for some constant \( C \). Let \( \delta_t \) be the cluster identity at time \( t \). Then there exists a mixture density \( q_T \) with \( K \) clusters with the following form:

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

Construct \( q_T(x_t | F_{t-1}^Q) \) from \( q_T(x_t | x_{t-1}, \delta_{t-1}) \) by integrating out \( \delta_{t-1} \) using \( \Pr(\delta_{t-1} | X_T) \).

Then with probability \( 1 - 2\delta \) with respect to the prior

\[
h_\infty(p_T(x_t | F_{t-1}^P), q_T(x_t | F_{t-1}^Q)) < C \left( 1 + \log \left( \frac{1}{\delta} \right) \right) \epsilon.
\]

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

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

\[
\theta_t \tilde{x}_t | \tilde{x}_{t-1}, \theta_t, \theta_{t-1} \overset{\mathcal{L}}{=} \theta_t \tilde{x}_t | \theta_{t-1} \tilde{x}_{t-1}, \theta_t, \theta_{t-1} \overset{\mathcal{L}}{=} \phi(\beta_{k,k'} \theta_{t-1} \tilde{x}_{t-1}, \Sigma_{k,k'}) \overset{\mathcal{L}}{=} \phi(\beta_{k,k'} \tilde{x}_{t-1}, \Sigma_{k,k'}). \tag{110}
\]
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 \( \theta_{t-1} \) are in \( \{-1, 0, 1\} \), we can absorb the \( \theta_{t-1} \) into the \( \beta_{k,k'} \) without increasing the number of clusters more than two-fold. This is because the vectors \( \theta_{t-1} \) that contain at most one non-zero element form a convex hull, and we take the weighted averages over them in (111).

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

\[
\tilde{x}_t \mid \theta_{t-1} = k, \tilde{x}_{t-1} \sim \sum_{k'} \phi(\beta_{k,k'}; \tilde{x}_{t-1}, \Sigma_{k,k'}) \Pr(\theta_t = k') \quad (111)
\]

The last probability — \( \Pr(\theta_t = k') \) — does not have any conditioning information because the rows of the \( \Theta_T \) process are independent except for the stopping rule, which is not relevant here. Define a set of clusters in \( (\tilde{x}_t, \tilde{x}_{t-1}) \) space by grouping the ones whose associated \( \{\beta, \Sigma\} \) are equal. In other words, take the Cartesian product of the clusters used in (111) and denote the cluster identities by \( \delta_t \)'s. Integrating out the cluster identities gives

\[
\tilde{x}_t \mid \tilde{x}_{t-1}, \delta_{t-1} \sim \sum_j \phi(\beta_j; \tilde{x}_{t-1}, \Sigma_j) \Pr(\delta_t = j \mid \delta_{t-1}) . \quad (112)
\]

Clearly, there are \( K_T^2 \propto \log(T)^2 \) different clusters.\(^{20}\)

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 11, 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

\(^{20}\)The number of clusters used here is of the same asymptotic order as in the prior. This bound may no longer be tight.
and integrate with respect to their joint density:

\[
\sup_t h^2 \left( \int_{G_t} \int_{G_t} \phi(x_t \mid \delta^P_t) \ dG^P_t(\delta^P_t) \ dK(dG^P_t, dG^Q_t), \right. \\
\left. \quad \int_{G_t} \int_{G_t} \phi(x_t \mid \delta^Q_t) \ dG^Q_t(\delta^Q_t) \ dK(dG^P_t, dG^Q_t) \right). 
\]  

Taking the Schweppe decomposition of the joint distribution gives

\[
\sup_t h^2 \left( \prod_t \int_{G_t} \phi(x_t \mid \delta^P_t) \ dG^P_t(\delta^P_t \mid \mathcal{F}^P_{t-1}), \prod_t \int_{G_t} \phi(x_t \mid \delta^Q_t) \ dG^Q_t(\delta^Q_t \mid \mathcal{F}^Q_{t-1}) \right). 
\]

By Lemma 8, we can replace the squared Hellinger distance by Kullback-Leibler divergence

\[
= C \sup_t D_{KL} \left( \prod_t \int_{G_t} \phi(x_t \mid \delta^P_t) \ dG^P_t(\delta^P_t \mid \mathcal{F}^P_{t-1}), \prod_t \int_{G_t} \phi(x_t \mid \delta^Q_t) \ dG^Q_t(\delta^Q_t \mid \mathcal{F}^Q_{t-1}) \right). 
\]

Simplifying notation gives:

\[
= C \sup_t D_{KL} \left( \prod_t p(x_t \mid \mathcal{F}^P_{t-1}), \prod_t q(x_t \mid \mathcal{F}^P_{t-1}) \right). 
\]

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

\[
C \sup_{F^P_{t-1}, F^Q_{t-1}} \sup_{t \in F^P_{t-1} \cap F^Q_{t-1}} \int_{\mathbb{R}^T \times D} \log \left( \frac{\prod_t p(x_t \mid \mathcal{F}^P_{t-1})}{\prod_t q(x_t \mid \mathcal{F}^P_{t-1})} \right) \prod_t p(x_t \mid \mathcal{F}^P_{t-1}) \ dX_T. 
\]

Dropping the inner supremum cannot make the value larger:

\[
\geq C \sup_{F^P_{t-1}, F^Q_{t-1}} \int_{\mathbb{R}^T \times D} \log \left( \frac{\prod_t p(x_t \mid \mathcal{F}^P_{t-1})}{\prod_t q(x_t \mid \mathcal{F}^P_{t-1})} \right) \prod_t p(x_t \mid \mathcal{F}^P_{t-1}) \ dX_T. 
\]
We can interchange \( \mathcal{F}^P_{t-1} \) and \( \mathcal{F}^Q_{t-1} \) by the hidden Markov assumption.

\[
= C \sup_{x_{t-1}, \delta_{t-1}^P, \delta_{t-1}^Q} \int_{\mathbb{R}^T \times D} \log \left( \frac{\prod_t p_M(x_t \mid x_{t-1}, \delta_{t-1}^P)}{\prod_t q_M(x_t \mid x_{t-1}, \delta_{t-1}^Q)} \right) \prod_t p_M(x_t \mid x_{t-1}, \delta_{t-1}^P) \, dX_T.
\]  

(119)

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

\[
= C \int_{\mathbb{R}^T} \sup_{x_{t-1}, \delta_{t-1}^P, \delta_{t-1}^Q} \int_{\mathbb{R}^D} \sum_t \log \left( \frac{p_M(x_t \mid x_{t-1}, \delta_{t-1}^P)}{q_M(x_t \mid x_{t-1}, \delta_{t-1}^Q)} \right) \prod_t p_M(x_t \mid x_{t-1}, \delta_{t-1}^P) \, dx_t \, d(\mathbb{R}^T).
\]  

(120)

We can pull the sum out:

\[
= C \int_{\mathbb{R}^T} \sum_t \sup_{x_{t-1}, \delta_{t-1}^P, \delta_{t-1}^Q} \int_{\mathbb{R}^D} \log \left( \frac{p_M(x_t \mid x_{t-1}, \delta_{t-1}^P)}{q_M(x_t \mid x_{t-1}, \delta_{t-1}^Q)} \right) \prod_t p_M(x_t \mid x_{t-1}, \delta_{t-1}^P) \, dx_t \, d(\mathbb{R}^T).
\]  

(121)

The values inside the sum are all the same:

\[
\geq CT \int_{\mathbb{R}^T} \sup_{x_{t-1}, \delta_{t-1}^P, \delta_{t-1}^Q} \int_{\mathbb{R}^D} \log \left( \frac{p_M(x_t \mid x_{t-1}, \delta_{t-1}^P)}{q_M(x_t \mid x_{t-1}, \delta_{t-1}^Q)} \right) \prod_t p_M(x_t \mid x_{t-1}, \delta_{t-1}^P) \, dx_t \, d(\mathbb{R}^T).
\]  

(122)

We can interchange the integral over \( \mathbb{R}^T \) and the supremum because they are over different arguments of \( p_M \) and \( q_M \); we also expand out the integral:

\[
= CT \sup_{x_{t-1}, \delta_{t-1}^P, \delta_{t-1}^Q} \int_{\mathbb{R}^D} \cdots \int_{\mathbb{R}^D} \log \left( \frac{p_M(x_t \mid x_{t-1}, \delta_{t-1}^P)}{q_M(x_t \mid x_{t-1}, \delta_{t-1}^Q)} \right) \prod_t dP_M(x_1 \mid x_0, \delta_0^P) \cdots dP_M(x_T \mid x_{T-1}, \delta_T^P).
\]  

(123)

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.

\[ = CT \sup_{x_{t-1}, \delta^P_{t-1}, \delta^Q_{t-1}} \int_{\mathbb{R}^D} \log \left( \frac{p_M(x_t \mid x_{t-1}, \delta^P_{t-1})}{q_M(x_t \mid x_{t-1}, \delta^Q_{t-1})} \right) dP_M(x_t \mid x_{t-1}, \delta^P_{t-1}). \]  

(124)

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

\[ = CT \sup_{x^P_{t-1}, x^Q_{t-1}} D_{KL} \left( p_M(x_t \mid \mathcal{F}^P_{t-1}) \parallel q_M(x_t \mid \mathcal{F}^Q_{t-1}) \right). \]  

(125)

Equation (113) equals the distance between the joint distributions — It is (92) from the joint density proof. 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 8, we can can replace the Kullback-Leibler divergence in (125) by squared Hellinger.

This gives

\[ T \sup_{x^P_{t-1}, x^Q_{t-1}} h^2 \left( p_M(x_t \mid \mathcal{F}^P_{t-1}) , q_M(x_t \mid \mathcal{F}^Q_{t-1}) \right) \leq CT(1 + \log(1/\delta))^2 \epsilon^2. \]  

(126)

Canceling the \( T \) terms and taking square roots finishes the proof.

\[ \sup_{x^P_{t-1}, x^Q_{t-1}} h \left( p_M(x_t \mid \mathcal{F}^P_{t-1}) , q_M(x_t \mid \mathcal{F}^Q_{t-1}) \right) \leq C(1 + \log(1/\delta))\epsilon. \]  

(127)

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

Proof. We can represent a Dirichlet process as \( \Pr(x) = \sum_{i=1}^{\infty} \beta_i \delta_{x_i}(x) \), where \( \delta_{x_i} \) is a indicator function with \( \delta_{x_i}(x_i) = 1 \), and the \( \beta_i \) satisfy a stick-breaking process. In other words, \( \beta_i = \beta'_i \prod_{j=1}^{i-1} (1 - \beta'_j) \) with \( \beta'_j \sim \text{Beta}(1, \alpha) \) for some positive scalar \( \alpha \). Consider the probability mass function of a row of \( \Theta_T \), \( \theta_t \). Then \( \Pr(|i| = 1) = b \prod_{j=1}^{j-1} (1 - b) \). Since draws from the beta distribution lie in \((0,1)\) with probability 1, these two stick-breaking processes are clearly mutually absolutely continuous. If we take \( x \in \{-1,1\} \) with the probability 1/2 each as the Dirichlet base measure, the
process are mutually absolutely continuous after possibly extending the space so that
the Beta random variables are well-defined.

Because these two processes are mutually absolutely continuous, a Radon-Nikodym
derivative exists because both measures are $\sigma$-finite. Since the rows are independent,
and Dirichlet processes are normalized random measures (Lin, Grimson, and Fisher,
2010), we can extend this to the entire $\Theta_T$ process. Consequently, any process that
is representable as an integral with respect to $\Theta_T$ can be represented as an integral
with respect to to a Dirichlet process. $\square$

Online Appendix C Contraction Rates

C.1 Constructing Exponentially Consistent Tests with Respect to $h_\infty$

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

1. $P_T (\Upsilon_T | \xi^P) \leq \exp(-C_2 T \epsilon^2)$ (17)

2. $\sup_{\xi \in \Xi, e_a(\xi, \xi^P) < \epsilon C_3} P_T (1 - \Upsilon_T | \xi^P) \leq \exp(-C_2 T \epsilon^2)$ (18)

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

$$f (X_T | G_1, \ldots G_T) = \prod_{t=1}^T \int_{G_t^f} \phi \left( x_t \big| \delta^f_t \right) dG_t^f(\delta^f_t).$$ (128)

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:

$$h^2_{\text{avg}} (f (X | \{G_i^f\}), g (X | \{G_i^g\}))$$

$$:= \frac{1}{T} \sum_{t=1}^T h^2 \left( \int_{G_t^f} \phi \left( x_t \big| \delta^f_t \right) dG_t^f(\delta^f_t), \int_{G_t^g} \phi \left( x_t \big| \delta^g_t \right) dG_t^g(\delta^g_t) \right).$$ (129)
Then by (Birgé, 2013, Corollary 2), there exists a test \( \phi_T \) that satisfies the following:\(^{21}\)

\[
\Pr_T \left( \phi_T(X) \bigg| \{ G_f^t, G_g^t \} \right) \\
\leq \exp \left( -\frac{1}{3} T \overline{h}_{avg}^2 \left( \int_{G_f^t} \phi \left( x_t \bigg| \delta_f^t \right) dG_f^t(\delta_f^t), \int_{G_g^t} \phi \left( x_t \bigg| \delta_g^t \right) dG_g^t(\delta_g^t) \right) \right) \\
\text{and}
\]

\[
\Pr_T \left( 1 - \phi_T(X) \bigg| \{ G_f^t, G_g^t \} \right) \\
\leq \exp \left( -\frac{1}{3} T \overline{h}_{avg}^2 \left( \int_{G_f^t} \phi \left( x_t \bigg| \delta_f^t \right) dG_f^t(\delta_f^t), \int_{G_g^t} \phi \left( x_t \bigg| \delta_g^t \right) dG_g^t(\delta_g^t) \right) \right).
\]

The issue with these equations is that they are not in terms of \( h_\infty \) and only hold conditionally. The reason that we can get around this is because they hold for all \( G_f^t \) and for all \( G_g^t \). Consequently, we can take the infimum of both sides, and bound the right-hand side of both equations by

\[
\frac{T}{3} \sup_{\{ (G_f^t, G_g^t) \}} \overline{h}_{avg}^2 \left( \int_{G_f^t} \phi \left( x_t \bigg| \delta_f^t \right) dG_f^t(\delta_f^t), \int_{G_g^t} \phi \left( x_t \bigg| \delta_g^t \right) dG_g^t(\delta_g^t) \right) \\
\text{for any length } T \text{ sequence. This equals the least favorable } G_f^t \text{ and } G_g^t \text{ repeated } T \text{ times. This joint distribution exists in our set because we are not placing any restrictions on the dynamics besides ergodicity. Stationary distribution are clearly ergodic. Hence, this equals}
\]

\[
= \frac{T}{3} \sum_{t=1}^{T} \overline{h}_{avg}^2 \left( \int_{G_{f_{sup}}^t} \phi \left( x_t \bigg| \delta_{f_{sup}}^t \right) dG_{f_{sup}}^t(\delta_{f_{sup}}^t), \int_{G_{g_{sup}}^t} \phi \left( x_t \bigg| \delta_{g_{sup}}^t \right) dG_{g_{sup}}^t(\delta_{g_{sup}}^t) \right).
\]

\(^{21}\)To map his notation into ours, take his \( z = 0 \), and take his measure \( R \) equal to \( P \). Equation (130) is obvious then, and (131) follows by taking the exponential of both sides in the inequality inside the probability and rearranging.
The terms inside the sum are all the same:

\[
\begin{align*}
&= T h^2 \left( \int_{G_{\sup}^f} \phi \left( x_t \mid \delta_t^f \right) dG_{\sup}^f(\delta_t^f), \int_{G_{\sup}^g} \phi \left( x_t \mid \delta_t^g \right) dG_{\sup}^g(\delta_t^g) \right) \\
&= T \sup_{(G_t^f, G_t^g)} h^2 \left( \int_{G_t^f} \phi \left( x_t \mid \delta_t^f \right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi \left( x_t \mid \delta_t^g \right) dG_t^g(\delta_t^g) \right) \\
&= T h^2 \left( \int_{G_t^f} \phi \left( x_t \mid \delta_t^f \right) dG_t^f(\delta_t^f), \int_{G_t^g} \phi \left( x_t \mid \delta_t^g \right) dG_t^g(\delta_t^g) \right).
\end{align*}
\]

Taking the supremum over $G_t^f$ and $G_t^g$ is equivalent to taking supremum over $F_{t-1}^f$ and $F_{t-1}^g$ 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 2 hold. \qed

### C.2 Bounding the Posterior Divergence

**Proposition 6** (Bounding the Posterior Divergence). Let $X_T$ satisfy Assumption 1 and Assumption 2. Let $p_T := \sum_k \Pi_{k,t} \phi(x_t \mid \mu_t, \Sigma_t)$ denote the true density. Let $\Xi_T \subset \Xi$ and $T \to \infty$. Let $Q_T$ be a mixture approximation with $K_T$ components. Assume the following condition holds with probability $1 - 2\delta$ for $\delta \in (0, 1/2)$ and constants $C$ and $i \in \mathbb{N}$:

\[
\sup_{\epsilon_i} \epsilon_i \left( q_T \left( x_t \mid F_{t-1}^Q \right), p_T \left( x_t \mid F_{t-1}^P \right) \right) < C \eta_T. \tag{19}
\]

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

\[
\sup_{\epsilon_i \geq \epsilon_{T,i}} \log N \left( \epsilon_i, \left\{ \xi \in \Xi_T \mid h_{\infty}(\xi, \xi^P) \leq \epsilon_i \right\}, h_{\infty} \right) \leq T \epsilon_{T,i}^2, \tag{20}
\]

and

\[
Q_T \left( B_T \left( \xi^P, \epsilon_{T,i}, 2 \right) \mid X_T \right) \geq C \exp \left( -C_0 T \epsilon_{T,i}^2 \right). \tag{21}
\]

**Proof.** We are looking at locally asymptotically mixed normal models, as discussed in Lemma 8, and we bind the Hellinger distance and Kullback-Leibler divergence in terms of $(x_t - \mu_t)^T \Sigma_t^{-1} (x_t - \mu_t)$. In addition, the supremum of the deviations is
clearly greater than the average of the deviations, and so the $h_\infty$-norm forms smaller balls than both $D_{KL}(f \mid \mid g)$ and $V_{k,0}$. Consequently, we can replace $B_T(\xi_0, \epsilon_T, 2)$ with \( \{ \xi \in \Xi \mid h_\infty^2(\xi, \xi_0) < T \epsilon_T^2 \} \). We use 2 as the last argument of $B$ because we are using $V_{2,0}$, i.e., effectively the 2nd-moment of the Kullback-Leibler divergence.

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

$$
\sup_{\epsilon_i > \epsilon_{T,i}} \log N (\epsilon_i, \{ \xi \in \Xi_T \mid h_\infty(\xi, \xi_0) \leq \epsilon_i \}, h_\infty) \leq T \epsilon_{T,i}^2 \tag{137}
$$

and

$$
\mathcal{Q}_T (\{ \xi \in \Xi \mid h_\infty^2(\xi, \xi_0) < \epsilon_{T,i} \}) \geq C \exp (-T \epsilon_{T,i}^2) \tag{138}
$$

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

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

$$
\sup_t h \left( q_T (x_t \mid \mathcal{F}_{t-1}^Q), p_T (x_t \mid \mathcal{F}_{t-1}^P) \right) < C \eta_T. \tag{139}
$$

Equation (138) is satisfied if

$$
\eta_T^2 \geq C \exp (-T \epsilon_{T,i}^2) \propto \exp \left( -T \frac{\log(T)^i}{T} \right) = \frac{1}{T^i}. \tag{140}
$$

To satisfy (137), $h_\infty^2$ must be bounded below and decline exponentially fast. The expressions above hold for any $\eta_T^2 \geq \eta_T$. Let $\eta_T^* = \frac{\log(T)^n}{T^n}$. We know there exists a covering with $K_T = \frac{\log(T)^n}{\eta_T^*}$ components. This implies that

$$
K_T = \frac{\log(T)^i}{\eta_T^*} = \frac{\log(T)^i}{\log(T)^i/T^i} = T^i. \tag{141}
$$

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

$$
N (\epsilon_n, \{ \xi \in \Xi \mid h_\infty^2(\xi, \xi_0) \leq \epsilon_i, h_\infty^2 \leq T^i \} = \exp (\log (T^i)) = \exp (T \epsilon_{T,i}^2). \tag{142}
$$
Taking logarithms of both sides of (142) finishes the proof.

\[ \square \]

### C.3 Contraction Rate of the Marginal Density

**Theorem 8** (Contraction Rate of the Marginal Density). Let \( X_T \) satisfy Assumption 1 and assume that the \( X_T \) are independent across \( t \). 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 \( \epsilon_T = \sqrt{\frac{\log(T)}{T}} \) and 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 above and the true transition density satisfies

\[
P_T (Q_T (h(p_T (x_t), q_T (x_t)) \geq C\epsilon_T | X_T)) \to 0.
\]

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

Having done that we can verify the conditions in Proposition 6. If we take \( i = 1 \) in (19), Theorem 3 implies the necessary bound on the sieve complexity exists. In addition, since \( h_{\infty} \) is bounded above by the Hellinger distance, \( h \), the conclusions of Proposition 6 trivially go through in Hellinger’s weaker topology.

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

\[ \square \]

### C.4 Contraction Rate of the Transition Density

**Theorem 7** (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 \( \epsilon_T := \sqrt{\frac{\log(T)^2}{T}} \) 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 above and the true transition density
satisfies

$$P_T \left( Q_T \left( \sup_{\mathcal{F}_{t-1}^P, \mathcal{F}_{t-1}^Q} h \left( p_T \left( x_t \mid \mathcal{F}_{t-1}^P \right), q_T \left( x_t \mid \mathcal{F}_{t-1}^Q \right) \right) \geq C \epsilon_T \right) \right) \rightarrow 0.$$

Proof. The proof of this is essentially identical to the marginal density case, mutatis mutandis. Lemma 2 implies the that $h_\infty$ has the required exponentially consistent tests. We verify the conditions in Proposition 6. If we take $i = 2$ in (19), Theorem 4 implies the necessary bound on the sieve complexity exists.

This verifies the three conditions in Theorem 5 on a set with with probability $1 - 2\delta$ with respect to the prior. This then gives us the posterior contraction rate

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

$\Box$
D.1 Component Coefficient Posterior

Let \( X_k \) be the \( T_k \times N \) vector and \( Y_k \) be the \( T_k \times D \) vector of data in component \( K \). This implies that \( \Sigma_k \) is a \( D \times D \) matrix and \( \beta_k \) is an \( N \times D \) matrix.\(^{22}\) Meanwhile, \( V \) is a \( D \times D \) matrix and \( U \) is a \( N \times N \) matrix.

The joint density is

\[
\Pr(Y_k, \beta_k, \Sigma_k | X_k) = \exp \left( -\frac{1}{2} \text{tr} \left\{ V_k^{-1} (\beta_k - \bar{\beta})' U^{-1} (\beta_k - \bar{\beta}) \right\} \right) \exp \left( -\frac{1}{2} \text{tr} \left\{ (Y_k - X_k\beta_k) \Sigma_k^{-1} (Y_k - X_k\beta_k)' \right\} \right) \\
\frac{\left| \Sigma_k \right|^{-T_k/2}}{(2\pi)^{T_k/2}} \frac{1}{\left| V \right|^N U^D} \frac{\left| (\mu_1 - 2) \Omega \right|^{\nu/2}}{\sqrt{2^{\nu} \Gamma(D(\nu/2))}} \left| \Sigma_k \right|^{-\nu+D+1} \exp \left( -\frac{1}{2} \text{tr} \left\{ (\mu_1 - 2) \Omega \Sigma_k^{-1} \right\} \right) 
\]

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

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

\(^{22}\)The likelihood in (143) is correct because the trace is the sum of the diagonal elements.
Combining the two kernels of $\beta_k$ and expanding gives

\[ \alpha |\Sigma_k|^{-\frac{\nu + D + T + 1}{2}} \exp \left( -\frac{1}{2} \text{tr} \left\{ V^{-1} \left( \left( \beta_k - \bar{\beta} \right)' U^{-1} \left( \beta_k - \bar{\beta} \right) \right) + \left( (Y_k - X_k \beta_k)' (Y_k - X_k \beta_k) + (\mu_1 - 2)\Omega \right) \Sigma_k^{-1} \right\} \right) \]  

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

(145)

Isolating the terms that have a $\beta_k$ in them:

\[ = \exp \left( -\frac{1}{2} \text{tr} \left\{ V_k^{-1} (-2\beta_k' U^{-1} \bar{\beta} + \beta_k' U^{-1} \beta_k) + \Sigma_k^{-1} (-2\beta_k' X_k' Y_k + \beta_k' X_k' X_k \beta_k) + V_k^{-1} \bar{\beta}' U^{-1} \bar{\beta} + \Sigma_k^{-1} (Y_k' Y_k + (\mu_1 - 2)\Omega) \right\} \right). \]  

(146)

Rewriting the traces in terms of the vectorization operator:

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

(147)

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

\[ \alpha \exp \left( \text{tr} \left\{ \beta_k' (U^{-1} \bar{\beta} V_k^{-1} + X_k' Y_k \Sigma_k^{-1}) \right\} - \frac{1}{2} \text{tr} \left\{ \left( (V_k^{-1} \otimes U^{-1}) + (\Sigma_k^{-1} \otimes X_k' X_k) \right) \text{vec} \{ \beta_k \} \text{vec} \{ \beta_k \} \right\} \right) \exp \left( -\frac{1}{2} \text{tr} \left\{ V_k^{-1} \bar{\beta}' U^{-1} \bar{\beta} + \Sigma_k^{-1} (Y_k' Y_k + (\mu_1 - 2)\Omega) \right\} \right) |\Sigma_k|^{-\frac{\nu + D + T + 1}{2}}. \]  

(148)
If we assume that \( V_k = \Sigma_k \), we can simplify this as

\[
\exp \left( \text{tr} \left\{ \beta_k' \left( U^{-1} \beta + X_k' Y_k \right) \Sigma_k^{-1} \right\} - \frac{1}{2} \text{tr} \left\{ \left( \Sigma_k^{-1} \otimes (U^{-1} + X_k' X_k) \right) \text{vec} \{ \beta_k \} \text{vec} \{ \beta_k' \} \right\} \right)
\]

\[
\exp \left( -\frac{1}{2} \text{tr} \left\{ \Sigma_k^{-1} \left( \beta' U^{-1} \beta + Y_k' Y_k + (\mu_1 - 2)\Omega \right) \right\} \right) |\Sigma_k|^{-\frac{\nu + D + T + 1}{2}}
\]

\[
= \exp \left( \text{vec} \{ \beta_k \} \text{vec} \{ (U^{-1} \beta + X_k' Y_k) \Sigma_k^{-1} \} - \frac{1}{2} \text{vec} \{ \beta_k \} \text{vec} \{ \Sigma_k^{-1} \otimes (U^{-1} + X_k' X_k) \} \text{vec} \{ \beta_k \} \right)
\]

\[
\exp \left( -\frac{1}{2} \text{tr} \left\{ \Sigma_k^{-1} \left( \beta' U^{-1} \beta + Y_k' Y_k + (\mu_1 - 2)\Omega \right) \right\} \right) |\Sigma_k|^{-\frac{\nu + D + T + 1}{2}}.
\] (150)

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 \( Z_k := (U^{-1} \beta + X_k' Y_k) \) and \( W_k := (U^{-1} + X_k' X_k) \). We can rewrite (150) as

\[
= \exp \left( -\frac{1}{2} \left( \text{vec} \{ \beta_k \} - \left( \Sigma_k^{-1} \otimes W_k \right)^{-1} Z_k \Sigma_k^{-1} \right)' \left( \Sigma_k^{-1} \otimes W_k \right) \left( \text{vec} \{ \beta_k \} - \left( \Sigma_k^{-1} \otimes W_k \right)^{-1} \Sigma_k^{-1} \right) \right)
\]

\[
\exp \left( \frac{1}{2} \Sigma_k^{-1} Z_k \left( \Sigma_k^{-1} \otimes W_k \right)^{-1} Z_k \Sigma_k^{-1} \right) \exp \left( -\frac{1}{2} \text{tr} \left\{ \Sigma_k^{-1} \left( \beta' U^{-1} \beta + Y_k' Y_k + (\mu_1 - 2)\Omega \right) \right\} \right) |\Sigma_k|^{-\frac{\nu + D + T + 1}{2}}.
\] (151)

I now eliminate all of the Kronecker products:

\[
= \exp \left( -\frac{1}{2} \text{vec} \{ \beta_k - W_k^{-1} Z_k \} \text{vec} \{ W_k (\beta_k - W_k^{-1} Z_k) \Sigma_k^{-1} \} \right)
\]

\[
\exp \left( \frac{1}{2} \text{vec} \{ (U^{-1} \beta + Z_k) \Sigma_k^{-1} \} \text{vec} \{ W_k^{-1} Z_k \} - \frac{1}{2} \text{tr} \left\{ \Sigma_k^{-1} \left( \beta' U^{-1} \beta + Y_k' Y_k + (\mu_1 - 2)\Omega \right) \right\} \right) |\Sigma_k|^{-\frac{\nu + D + T + 1}{2}}.
\] (153)

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:

\[
= \exp \left( -\frac{1}{2} \text{tr} \left\{ \Sigma_k^{-1} \left( \beta_k - (U^{-1} + X'_k X_k)^{-1} (U^{-1} \bar{y} + X'_k Y_k) \right)' (U^{-1} + X'_k X_k) \left( \beta_k - (U^{-1} + X'_k X_k)^{-1} (U^{-1} \bar{y} + X'_k Y_k) \right) \right\} \right)
\]

\[
\exp \left( -\frac{1}{2} \text{tr} \left\{ \Sigma_k^{-1} \left( (\bar{y}'U^{-1}\bar{y} + (\mu_1 - 2)\Omega) - (U^{-1} \bar{y} + X'_k Y_k)' (U^{-1} + X'_k X_k)^{-1} (U^{-1} \bar{y} + X'_k Y_k) \right) \right\} \right) \left| \Sigma_k \right|^{-\nu_0 + D + T + 1 \slash 2}.
\]

The first expression is kernel of a matrix-normal distribution. The mean is \((U^{-1} + X'_k X_k)^{-1} (U^{-1} \bar{y} + X'_k Y_k)\), and the two covariance parameters are \(\Sigma_k\), and \((U^{-1} + X'_k X_k)^{-1}\). The second expression is the kernel of a Inverse-Wishart distribution. Its scale parameter is \((\bar{y}'U^{-1}\bar{y} + (\mu_1 - 2)\Omega) - (U^{-1} \bar{y} + X'_k Y_k)' (U^{-1} + X'_k X_k)^{-1} (U^{-1} \bar{y} + X'_k Y_k)\). It has \(\mu_1 + D - 1 + T_k\) degrees of freedom. To see the intuition behind this, note that if \(U^{-1}\) and \(\Omega\) 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 \(\beta_k\) parameter does not show up in the second expression, we can draw from the posterior by drawing the \(\Sigma_k\) from its marginal posterior, and then drawing from the posterior of \(\beta_k\) conditional on \(\Sigma_k\).

### D.2 Hypermean Posterior with Heteroskedastic Data

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

\[
\Pr \left( \{\beta\}^K_{k=1}, \bar{\beta}, \{\Sigma\}^K_{k=1} \right) = \exp \left( -\frac{1}{2} \text{tr} \left\{ V^{-1} (\bar{\beta} - \beta^t)' U^{-1} (\bar{\beta} - \beta^t) \right\} \right) \exp \left( \sum_{k=1}^{K} -\frac{1}{2} \text{tr} \left\{ \Sigma_k^{-1} (\beta_k - \bar{\beta})' U^{-1} (\beta_k - \bar{\beta}) \right\} \right)
\]

\[
\sqrt{(2\pi)^{NK}} |U|^D |U|^{-\nu_U + N + 1 \slash 2} \exp \left( -\frac{1}{2} \text{tr} \left\{ \Psi_U U^{-1} \right\} \right) \prod_{k=1}^{K} \frac{1}{\sqrt{(2\pi)^{NK} |\Sigma_k|^N |U|^D}}
\]

\[(155)\]
Dropping all of the terms that contain neither $\beta$ nor $U$:

$$
|U|^{-\frac{\nu U + N + (K+1)D+1}{2}} \exp \left( -\frac{1}{2} \text{tr} \left\{ V^{-1} (\beta - \beta^t)'U^{-1}(\beta - \beta^t) + \sum_{k=1}^{K} \Sigma_k^{-1}(\beta_k - \beta^t)'U^{-1}(\beta_k - \beta^t) \right\} \right) \exp \left( -\frac{1}{2} \text{tr} \{ \Psi_U U^{-1} \} \right).
$$

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

$$
\propto \exp \left( -\frac{1}{2} \text{tr} \left\{ V^{-1}\beta'U^{-1}\beta - 2V^{-1}\beta^tU^{-1}\beta + V^{-1}\beta^tU^{-1}\beta^t + \sum_{k=1}^{K} \Sigma_k^{-1}(\beta'^tU^{-1}\beta - 2\beta'_kU^{-1}\beta + \beta'_kU^{-1}\beta_k) \right\} \right) \quad (156)
$$

Exploiting properties of the trace and vectorization, where $B := \text{vec}\{\beta\}$:

$$
\propto \exp \left( -\frac{1}{2} \text{vec}\{\beta^t\}'(V^{-1} \otimes W^{-1})B + \text{vec}\{W^{-1}\beta^tV^{-1}\}'B - \frac{1}{2} \sum_{k=1}^{K} \text{tr}\{(\Sigma_k^{-1} \otimes U^{-1})BB'\} + \text{vec}\left\{ \sum_{k=1}^{K} U^{-1}\beta_k\Sigma_k^{-1} \right\}'B \right) \\
|U|^{-\frac{\nu U + N + (K+1)D+1}{2}} \exp \left( -\frac{1}{2} \text{tr} \left\{ V^{-1}\beta'^tU^{-1}\beta^t + \sum_{k=1}^{K} \Sigma_k^{-1}\beta_k'^tU^{-1}\beta_k + \Psi_U U^{-1} \right\} \right). \quad (158)
$$
We can simplify using the circular commutativity of the trace:

\[
\propto \exp \left( -\frac{1}{2} \text{vec}\{\vec{\beta}\}' \left( \left( \sum_{k=1}^{K} \Sigma_k^{-1} \right) \otimes U^{-1} + V^{-1} \otimes U^{-1} \right) \text{vec}\{\vec{\beta}\} + \text{vec}\left\{ U^{-1} \beta V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \right\}' \text{vec}\{\vec{\beta}\} \right) \] (159)

\[
|U|^{-\frac{\nu_l + N + (K+1)D + 1}{2}} \exp \left( -\frac{1}{2} \text{tr} \left\{ \beta V^{-1} \beta' U^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \beta_k' U^{-1} + \Psi_U U^{-1} \right\} \right).
\]

Collecting terms:

\[
\propto \exp \left( -\frac{1}{2} \text{vec}\{\vec{\beta}\}' \left( \left( \sum_{k=1}^{K} \Sigma_k^{-1} + V^{-1} \right) \otimes U^{-1} \right) \text{vec}\{\vec{\beta}\} + \text{vec}\left\{ U^{-1} \left( \beta V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \right) \right\}' \text{vec}\{\vec{\beta}\} \right) \] (160)

\[
|U|^{-\frac{\nu_l + N + (K+1)D + 1}{2}} \exp \left( -\frac{1}{2} \text{tr} \left\{ \left( \sum_{k=1}^{K} \Sigma_k^{-1} + V^{-1} \right) \beta' U^{-1} \beta + \left( \beta V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \right)' U^{-1} \beta \right\} \right)
\]

\[
\propto \exp \left( -\frac{1}{2} \text{tr} \left\{ \left( \sum_{k=1}^{K} \Sigma_k^{-1} + V^{-1} \right) \beta' U^{-1} \beta + \left( \beta V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \right)' U^{-1} \beta \right\} \right) \] (161)

\[
|U|^{-\frac{\nu_l + N + (K+1)D + 1}{2}} \exp \left( -\frac{1}{2} \text{tr} \left\{ \left( \beta V^{-1} \beta' + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \beta_k + \Psi_U \right) U^{-1} \right\} \right).
\]

We now vectorize the first line of (161) 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

\[
\exp \left( -\frac{1}{2} \text{vec}\{\vec{\beta}\}' \left( \left( \sum_{k=1}^{K} \Sigma_k^{-1} + V^{-1} \right) \otimes U^{-1} \right) \text{vec}\{\vec{\beta}\} - 2 \text{vec}\left\{ U^{-1} \left( \beta V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \right) \right\}' \text{vec}\{\vec{\beta}\} \right) \] (162)
We then apply the multivariate equation of squares, and let $Z := (\beta^t V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1})$ and $W := (\sum_{k=1}^{K} \Sigma_k^{-1} + V^{-1})$:

$$
\exp \left( -\frac{1}{2} \left( \text{vec} \{ \hat{\beta} \} - (W \otimes U^{-1})^{-1} \text{vec} \{ U^{-1} Z \} \right) (W \otimes U^{-1}) \left( \text{vec} \{ \hat{\beta} \} - (W \otimes U^{-1})^{-1} \text{vec} \{ U^{-1} Z \} \right)^t \right)
\exp \left( \frac{1}{2} \text{vec} \{ U^{-1} Z \}^t (Z \otimes U^{-1})^{-1} \text{vec} \{ U^{-1} Z \} \right)
$$

(163)

We can simplify the vectorization.

$$
\exp \left( -\frac{1}{2} \text{vec} \{ \hat{\beta} - Z W^{-1} \} (W \otimes U^{-1}) \text{vec} \{ \hat{\beta} - Z W^{-1} \} \right) \exp \left( \frac{1}{2} \text{tr} \{ U^{-1} Z W^{-1} Z' \} \right)
$$

(164)

We can replace the vectorizations with traces.

$$
\exp \left( -\frac{1}{2} \text{tr} \{ U^{-1} (\hat{\beta} - Z W^{-1}) W (\hat{\beta} - Z W^{-1}) \} \right) \exp \left( \frac{1}{2} \text{tr} \{ U^{-1} Z W^{-1} Z' \} \right)
$$

(165)

Equation (165) 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 $(\beta^t V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1})(\sum_{k=1}^{K} \Sigma_k^{-1} + V^{-1})^{-1}$ Note, there is no reason here that $\beta_k$ cannot itself be a matrix.

To compute the distribution of $U$, we combine the last lines of (161) and (165). This gives

$$
|U|^{-\frac{D+1}{2}} \exp \left( -\frac{1}{2} \text{tr} \left\{ U^{-1} \left( \beta^t V^{-1} \beta' + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \beta_k' + \Psi_U \right) - \left( \beta^t V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \right) \left( \sum_{k=1}^{K} \Sigma_k^{-1} + V^{-1} \right)^{-1} \left( \beta^t V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \right)' \right\} \right)
$$

(166)
Clearly, $U$ is marginally inverse-Wishart. It has $\nu_U + (K + 1)D$ degrees of freedom, and its scale matrix equals

$$
\beta^\top V^{-1} \beta' + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1} \beta'_k + \Psi_U - (\beta^\top V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1})(\sum_{k=1}^{K} \Sigma_k^{-1} + V^{-1})^{-1}(\beta^\top V^{-1} + \sum_{k=1}^{K} \beta_k \Sigma_k^{-1})'.
$$

### D.3 Derivation of the Posterior of the Innovation Covariances’ Mean

The product of the relevant likelihood and prior is

$$
\Omega \mid \{\Sigma_k\}_{k=1}^{K} \propto \prod_{k=1}^{K} \Omega\left| \mu_1 + D - 1 \right| \exp\left( -\frac{\mu_1 - 2}{2} \text{tr}\{\Omega \Sigma_k^{-1}\} \right) \cdot |\Omega|^{\frac{\nu_2}{2}} \exp\left( -\frac{1}{2} \text{tr}\{\text{diag}(a_1, \ldots, a_D)^{-1}\Omega\} \right).
$$

This is the kernel of a Wishart distribution. That is

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