UNOBSERVED HETEROGENEITY, GROUPED RANDOM EFFECTS AND THE EAMP ALGORITHM

KARUN ADUSUMILLI

Abstract. We propose Grouped Random Effects as a new approach for nonlinear panel data models with unobservables. This posits that observations can be separated into groups, each having its own prior on unobserved heterogeneity, and with the parameters of the prior unknown and differing across groups. Both random effects and grouped fixed effects are special cases of this. The model can be estimated by jointly maximizing over group assignments, common and prior parameters. We propose a novel and fast algorithm to carry out this maximization, termed EAMP, that augments the standard EM algorithm with two additional steps: Assignment (A) for group assignment and Propagation (P) for updating the prior. We further show that the steps in the EAMP algorithm are closely related to those for mean-field Variational-Bayes inference in Dirichlet mixture models. Advantages of the Grouped Random Effects approach include automatic first order bias correction, need for fewer number of groups, and greater flexibility in modeling unobserved dynamics. We illustrate our methods using two examples: the first studies heterogeneity in income dynamics using PSID data. The second studies heterogeneity in the effect of union status on wages. In the case of income dynamics, we find large heterogeneity in the magnitudes and persistence of income shocks. This heterogeneity can explain the observed non-normality and nonlinear persistence of income shocks.

This version: March 1, 2021 .

1. Introduction

Many econometric models feature rich time varying patterns of unobserved heterogeneity. The two common approaches to tackling unobserved heterogeneity in panel data models are random and fixed effects. Recently, discrete approaches, such as Grouped-fixed effects (Hahn and Moon, 2010; Bonhomme and Manresa, 2015; Bonhomme et al., 2017), have been proposed as a way to increase the precision of fixed effect estimates by grouping observations into a finite number of clusters.

In this paper, we propose Grouped Random Effects (GRE) as a new approach for nonlinear panel data models with unobserved heterogeneity. This nests both Grouped-Fixed Effects (GFE) and Random Effects (RE) as special cases. It posits that observations can be separated into groups, each having its own prior on unobserved heterogeneity, and with the parameters of the prior unknown and differing across groups. The additional flexibility afforded by the prior results in a number of advantages over existing methods:

First, it requires fewer assumptions and number of groups than GFE. In GFE, observations are clustered so that the unobserved heterogeneity is approximately constant within in each group. By contrast, the GRE approach only requires the unobserved heterogeneity to be approximately independent of the covariates within each group. This is a weaker requirement, and it allows for smaller bias with the same number of groups (or fewer groups for the same bias).

Second, GRE is better at capturing the distribution of unobserved heterogeneity. Our algorithm for GRE computes the posterior distribution of the unobserved heterogeneity for each observation, which can be used to calculate marginal effects using posterior averaging (Bonhomme and Weidner [2019]). Due to posterior averaging, the GRE estimator achieves automatic first-order bias correction: we show that under time invariant heterogeneity, GRE estimators for common parameters and marginal effects have a bias of the order $T^{-3/2}$ (where T is the number of time periods), whereas GFE and standard random effects typically produce a T^{-1} bias.

Third, GRE enables greater flexibility in modeling time-varying unobserved heterogeneity. It allows for both priors that restrict the correlation between unobserved heterogeneity, e.g., an auto-regressive (AR) specification, and also priors where the covariance structure is completely unrestricted. Many settings such as the persistent-transitory model for income dynamics naturally feature an AR specification for unobserved heterogeneity. Furthermore, with a flexible specification, GRE can adapt

to the underlying dynamics of the unobserved heterogeneity since it estimates the co-variances between them from the data. This can result in improved estimates of marginal effects, if, e.g., the unobserved heterogeneity was really time invariant, but this was not known beforehand.

We estimate the GRE model by jointly maximizing the likelihood of the observations over group assignments, prior and structural parameters. This is the Maximum-Likelihood (MLE) approach. For the computation of MLE estimates, we propose a novel variant of the EM algorithm, called EAMP. It consists of four steps: Expectation (E) for calculating the posterior; Assignment (A) for assigning individuals to a group; Maximization (M) for maximizing the expected log-likelihood with respect to the structural parameters; and Propagation (P) for propagating posterior moments back to the prior. The E and M steps are the same as in a standard EM algorithm, while A and P are new. The derivation of this algorithm is based on the interpretation of EM as variational optimization (Neal and Hinton [1998]), which we generalize and adapt to our context. All of the steps in EAMP can be computed very efficiently when the panel likelihood admits a conjugate prior, with the computational burden being comparable to Lloyd's algorithm used in GFE (in fact this can be considered a special case of EAMP). For non-conjugate likelihoods one can employ approximate ways to perform the E step, using Laplace or local variational approximations.

We emphasize that the EAMP algorithm is distinct from the EM algorithm for mixture models. Following the seminal contribution of Heckman and Singer [1984], there has been a large literature on *frequentist* mixture models for unobserved heterogeneity. The aim in this literature is to flexibly model the random effects distribution using mixtures of distribution. However, the mixture probabilities are assumed to be independent of covariates, so the random effects distribution is common to all observations. By contrast, the GRE employs binary mixture weights (group assignments) that are intrinsic to each observation (and so, naturally correlated with covariates). The goal of the A step in EAMP is precisely to extract these weights. In fact, the A step does not have a equivalent counterpart in the standard EM algorithm.

There is, however, a close connection between GRE and *Bayesian* mixture models for unobserved heterogeneity, where the mixture weights are binary, latent variables with an attendant Dirichlet hyper-prior (but in frequentist mixtures, these weights are probabilities). The EAMP algorithm for GRE is closely related to mean-field

Variational-Bayes (VB) inference in the Bayesian model. The connection also extends to Dirichlet Process (DP) priors where the number of groups is allowed to be infinity. Constructing the VB algorithm, and highlighting this connection is another contribution of this paper. Our analysis shows that GFE, GRE and Variational-Bayes differ only in terms what variables are treated as random or as parameters. For instance, GRE treats unobserved heterogeneity as latent variables, while the group assignments are estimated as fixed parameters. Variational-Bayes treats both as random. Our VB algorithm is also practically useful in that it is much less computationally expensive than MCMC based methods commonly employed in Bayesian panel data models, e.g., Hirano [2002], Liu [2018] and Liu et al. [2019].

We study the theoretical properties of the GRE estimator under both a discrete setting, with finite number of groups and time-varying unobserved heterogeneity, and a continuous one with time invariant unobserved heterogeneity. Our analysis of the discrete setup is very similar to that for GFE (see, e.g., Bonhomme and Manresa [2015] and Cheng et al. [2019]), and we show that perfect group assignment is possible in settings with large T.

Our analysis of continuous unobserved heterogeneity shows some interesting results. Bonhomme et al. [2017] study the properties of clustering with continuous unobserved heterogeneity under a two step approach, where the observations are grouped in the first step and the parameters estimated in the second. They posit low dimensional unobserved heterogeneity in both covariates and outcomes. By contrast, we analyze the properties of one-step methods (i.e., which jointly estimate both parameters and group assignments) and we find that the latter do not require any restriction on the distribution of the covariates. Thus - and this is true for GFE estimation as well - one-step methods can achieve significantly lower discretization bias. But the benefit of the GRE approach goes much further: we show that for estimation of common parameters and marginal effects, the bias is of the order $T^{-3/2}$. This in contrast to GFE, which only achieves T^{-1} bias, even in a one-step setting. The reason for this, as mentioned earlier, is that GRE employs the posterior distribution of unobservables to estimate common parameters and marginal effects.

Our work builds on a large literature on modeling latent variables in both econometrics and computer-science. On the econometrics side, it is related to the growing literature on classifying individuals into groups based on latent characteristics. These estimators were introduced in Hahn and Moon [2010] and Bonhomme and

Manresa [2015], and have been generalized by Vogt and Linton [2020], Ando and Bai [2016]), Bonhomme et al. [2017] and Cheng et al. [2019]. While all these papers discretize unobserved heterogeneity, our work differs in specifying a continuous, group specific prior. Therefore, this paper is also related to previous work on random effects, particularly that of Lancaster [2002] and Arellano and Bonhomme [2009]. In the special case of linear panel data models with time-invariant α_i, Liu et al. [2020] propose an Empirical-Bayes estimator for the posterior mean of α_i, and show that it is ratio optimal. The methods here do not come with such guarantees, but are however more general in that they allow α_{it} to be time varying, can be used to estimate features of the distribution of α_i beyond the posterior mean, and also apply to nonlinear panel data models.

On the computational side, our paper is related to the vast and growing literature on estimation of latent variables and Variational Inference. We refer to Bishop [2006] and Wainwright et al. [2008] for textbook treatments. Our VB algorithm is a variant of Coordinate Ascent Variational Inference (CAVI), itself a form of variational message passing (Winn and Bishop, 2005). The treatment, however, differs from standard mean field analysis in that the posterior distribution of the latent variables is not fully factorized.

We illustrate our methodology using two empirical examples: The first studies heterogeneity in income dynamics using PSID data. Identification of the nature and persistence of income shocks is important for answering a wider range of economic questions including the effectiveness of self-insurance (Blundell et al. [2008]), the determinants of inequality (Huggett et al. [2011]), and effectiveness of stabilization or stimulus policies, among others. For instance, consumption responds much more strongly to persistent income shocks as opposed to transitory ones, so the relative magnitudes of these shocks has important implications for consumption inequality. The presence of a potential random walk in earnings is also important for understanding asset accumulation and welfare. Previously, Alan et al. [2018] found strong evidence for unobserved heterogeneity in the magnitude and persistence of income shocks. However their methods do not distinguish between transitory and permanent shocks. Using the GRE approach, we find substantial heterogeneity in all three quantities. We find that the persistence of income shocks is substantially far from a unit root (as in Alan et al. [2018], we find the median persistence to be 0.85). We also show that the heterogeneity in the magnitude of income shocks can explain

their observed non-normality. Furthermore, we find persistence is lower among individuals facing smaller magnitudes of shocks. When aggregated across households, this would show up as nonlinear persistence. TBC

2. Grouped random-effects

We work within a nonlinear panel data setting. The observed variables are $\{(\mathbf{y}_i, \mathbf{x}_i) : i = 1, \ldots, n\}$, where $\mathbf{y}_i := (y_{i1}, \ldots, y_{iT})$ and $\mathbf{x}_i := (x_{i1}, \ldots, x_{iT})$ are both vectors corresponding to the time series data of each observation i until T time periods. The conditional density of $\mathbf{y}_i$ given $\mathbf{x}_i$ is parametrized by the likelihood function $p(\mathbf{y}_i|\mathbf{x}_i, \alpha_i, \theta)$, where α_i is an unobserved/latent variable that vector valued and potentially time varying. We do not restrict the dependence of y_{it} on past observations, so the likelihood could potentially be dynamic. The parameter θ is unknown and finite dimensional. In general, we shall suppose that the distribution of the unobservables, α_i, in the population is continuous. Our parameter of interest can be θ or some marginal effect.

Arellano and Bonhomme [2009] propose a class of estimators by maximizing the integrated log-likelihood

$$\sum_i \ln f_i(\theta) = \sum_i \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha)d\alpha.$$

Here $\pi(\alpha)$ is a weight, or prior, over the support of α. In random effects estimation, we would parametrize the prior by $\pi(\alpha|\gamma)$ and jointly maximize over θ and γ. In this paper, we generalize the random-effects approach by letting the parameter γ in $\pi(\alpha|\gamma)$ vary by group. Let $\mathcal{S} = \{1, \ldots, S\}$ denote the groups and $w_i(s)$ an indicator variable that takes the value of 1 when observation i is in group s. Fixing the class of the prior, we propose to maximize the integrated log-likelihood

$$\begin{aligned}\sum_i \ln f_i\left(\theta, \{\gamma_s\}_s, \{w_i(s)\}_{i,s}\right) &= \sum_i \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta) \left\{\prod_s \pi(\alpha|\gamma_s)^{w_i(s)}\right\} d\alpha \\ &= \sum_i \sum_s w_i(s) \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha, \qquad (2.1)\end{aligned}$$

where the second equality follows by the fact that $w_i(s)$ only takes the values 0 or 1. Note that the maximization is carried out jointly over $\theta, \{\gamma_s\}_s$ and $\{w_i(s)\}_{i,s}$. We call the resulting estimator, the Grouped Random Effects (GRE) estimator.

The prior, $\pi(\alpha|\gamma)$, can be chosen to be any member of the exponential family

$$\pi(\alpha|\gamma) = h(\alpha)g(\gamma)\exp(\gamma^\intercal u(\alpha)). \qquad (2.2)$$

This family includes many of the most commonly used distributions such as the Gaussian, gamma and beta distributions, among others. Following standard terminology, we refer to γ as the natural parameter, and $u(\cdot)$ as the sufficient statistic of the exponential family. The restriction to exponential families is convenient for computation but not strictly necessary (see, for instance, our example below with a Gaussian AR(1) specification of the prior). Very often the conditional likelihood $p(\mathbf{y}_i|\mathbf{x}_i, \alpha_i, \theta)$ will have a conjugate prior, in which case it will turn out to be computationally attractive to choose this for $\pi(\alpha|\gamma)$.

The parameter of interest could be θ, or, more generally, some marginal effect

$$\tau_0 = E\left[\psi(\alpha_i, \theta_0)\right]. \tag{2.3}$$

For instance, τ could be a measure of the average or variance of heterogeneity in random coefficient panel data models (Chamberlain [1992]; Arellano and Bonhomme, 2009). We allow α_i to be vector valued, so τ could also be a measure of correlation between different components of α_i. The EAMP algorithm that we propose in the next section automatically computes group-specific posterior distributions $q_{si}(\alpha_i)$ for each observation i. These can be used to estimate τ using posterior-averaging (Bonhomme and Weidner, 2019)

$$\hat{\tau} = \frac{1}{n}\sum_i E_{q_i(\alpha_i)}\left[\psi(\alpha_i, \hat{\theta})\right]; \quad q_i(\alpha_i) := \prod_s q_{si}(\alpha)^{w_i(s)}. \tag{2.4}$$

In the above expression $w_i(s) \in \{0, 1\}$ are also computed from our EAMP algorithm for maximizing (2.1). In the Bayesian setting, w_i are treated as latent variables, similar to α_i, and we would therefore need a further expectation with respect to the posterior $q(w_i)$ of w_i; see Section (4.2).

The GFE estimator of Bonhomme and Manresa [2015] is a special case of our approach, obtained by setting $\pi(\alpha|\gamma) = \delta(\alpha|\gamma)$, where $\delta(\alpha|\gamma)$ denotes the Dirac delta function at the point γ. In particular, the authors proceed by classifying the observations into a finite number of groups, and letting the value of α_i be constant within each group. The GRE estimator is also closely related to finite mixture models which posit $\pi(a) = \sum_s w(s)\pi(\alpha|\gamma_s)$ where $w(s)$ are mixture probabilities that do not vary with i (or the covariates). In some contexts, mixture models suffer from an 'zero variance' problem, wherein the likelihood goes to ∞ when the prior collapses to a point mass (e.g., Gaussian mixtures for clustering). This is not a

concern with the GRE estimator since it just becomes GFE when the prior becomes a point mass.

2.1. **General estimating equations.** Our procedure can also be applied in cases where the parameters are obtained from an estimating equation rather than a conditional likelihood. Specifically, suppose the true value θ^* satisfies the moment condition

$$\theta^* = \arg\max_{\theta} E_{(y,x,\alpha)} \left[m(\mathbf{y}_i, \mathbf{x}_i, \alpha; \theta) \right],$$

where the expectation is joint over (y, x, α). Note that the above is equivalent to

$$\theta^* = \arg\max_{\theta} E \left[\int q_i(\alpha|\mathbf{y}_i, \mathbf{x}_i) m(\mathbf{y}_i, \mathbf{x}_i, \alpha; \theta) d\alpha \right], \tag{2.5}$$

where $q^*(\alpha|\mathbf{y}, \mathbf{x})$ is the conditional likelihood of α given $(\mathbf{y}, \mathbf{x})$, and the expectation is over the observables $(\mathbf{y}, \mathbf{x})$.

We can estimate θ in this setting using a GRE estimator. The estimation proceeds by jointly maximizing the criterion function

$$\bar{m}\left(\theta, \{\gamma_s\}_s, \{w_i(s)\}_{i,s}\right) = \sum_{i=1}^{n} \sum_{s} w_i(s) \ln \int \exp\left\{m(\mathbf{y}_i, \mathbf{x}_i, \alpha, \theta)\right\} \pi(\alpha|\gamma_s) d\alpha$$

over $\theta, \{\gamma_s\}, \{w_i(s)\}$. The reason this is a valid approach is that the resulting estimate $\hat{\theta}$, of θ^* can be interpreted as

$$\hat{\theta} = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \int \hat{q}_i(\alpha|\mathbf{y}_i, \mathbf{x}_i) m(\mathbf{y}_i, \mathbf{x}_i, \alpha; \theta) d\alpha, \tag{2.6}$$

where $\hat{q}_i(\alpha|\mathbf{y}_i, \mathbf{x}_i)$ is an estimate of the posterior distribution of α given $(\mathbf{y}_i, \mathbf{x}_i)$. This interpretation is easiest seen from the EM based algorithm that we propose in the next section to compute $\hat{\theta}$, and which the maximization of $\bar{m}\left(\theta, \{\gamma_s\}_s, \{w_i(s)\}_{i,s}\right)$ is equivalent to - for now we just note that one can think of (2.6) as the 'last' M-step in this algorithm.

Despite the generality, for the rest of the paper, we shall continue working with $p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)$ for simplicity.

2.2. **The GRE approach in various panel data settings.**

2.2.1. *Linear random coefficient panel data models.* Our method covers linear panel data models of the form

$$y_{it} = \alpha_{it}^{\intercal} w_{1it} + \theta^{\intercal} w_{2it} + \epsilon_{it},$$

where α_{it} is a time-varying heterogeneity term. Here $(w_{1it}, w_{2it}) = f(\mathbf{x}_i, y_{it-1}, \ldots, y_{i0})$ can be any known functions of covariates and past outcomes, assumed to be uncorrelated with the current and future values of ϵ_{it}.[1] The linear random coefficients model is widely used in economics, including in models of earnings and consumption dynamics (Alan et al. [2018]; Lee [2019]), estimation of production functions with firm specific technology (Cheng et al. [2019]), panel forecasts (Liu et al. [2019]), estimation of heterogenous treatment effects (Arellano and Bonhomme [2012]) and difference-in-difference estimation (Bonhomme and Manresa [2015]). For instance, the standard fixed effects model is a special case, obtained by setting $w_{1it} = 1$.

We can adopt a (pseudo -) conditional likelihood approach by specifying[2]

$$p(\mathbf{y}_i|\mathbf{x}_i, \alpha_i, \theta) \propto \prod_{t=1}^{T} \exp\left\{-\left(y_{it} - \alpha_{it}^{\intercal} w_{1it} - \theta^{\intercal} w_{2it}\right)^2\right\}.$$

The conjugate prior for this likelihood family is

$$\pi(\alpha|\gamma) \equiv N(\mu, \Sigma); \quad \gamma := (\Sigma^{-1}\mu, -\Sigma^{-1}/2), u(\alpha) := (\alpha, \alpha\alpha^{\intercal}).$$

The parameters of interest may be θ or any functional of α_{it}, e.g., $E[\alpha_{it_0}]$.

The difference between GRE and GFE in the linear setting is that the former incorporates a variance term Σ (letting $\Sigma \to 0$ gives the GFE estimator). It is better able adapt to the correlation structure of α since Σ can be estimated from the data.

Nevertheless, estimating Σ in its entirety may result in substantial variance due to estimating T^2 parameters. It is often preferable to restrict the covariance further. In the simplest instance, we may assume there is no correlation across time. This is equivalent to setting $N(\mu, \Sigma) = \prod_t N(\mu_t, \Sigma_t)$.

Alternatively, we may posit an AR(1) specification for unobserved heterogeneity of the form $\alpha_{it} = \lambda_t + \rho\alpha_{it-1} + v_{it}$ for $t > 1$ with $v_{it} \sim N(0, \sigma^2)$, and $\alpha_{i1} \sim N(c_1, \sigma^2/(1-\rho^2))$. Here $\{v_{i2}, \ldots, v_{iT}\}$ are assumed to be iid, and $\{\lambda_1, \ldots, \lambda_T\}$ are unknown constants. This implies $\pi(\alpha|\gamma) \equiv N(\mu, \Sigma(\sigma, \rho))$, where the (t, t')-th element of $\Sigma(\sigma, \rho)$ is given by $\sigma^2\rho^{-|t-t'|}/(1-\rho^2)$, and μ is a T-dimensional vector representing $E[\alpha_i]$. There is a one-to-one transformation between μ and $[\lambda_1, \ldots, \lambda_T]^{\intercal}$, but we will have little use for the latter, so we focus on computing μ. This prior is

[1] In the dynamic setting, the likelihood conditions on the first period outcome y_{i0}, which is reasonable if T is moderately large.

[2] This is a pseudo-likelihood since we do not actually know that ϵ_{it} is distributed as a standard normal. However, as Bonhomme and Manresa [2015] demonstrate, this often still results in consistent estimates of marginal effects.

not a part of the exponential family (atleast in terms of the parameters μ, σ, ρ), but we show in Appendix A that one can still efficiently compute the GRE estimator.

2.2.2. *Fixed effect panel data models when the likelihood is from an exponential family.* When the conditional likelihood is from an exponential family, we can choose $\pi(\alpha|\gamma)$ from the family of priors conjugate to it. Examples include Poisson, Exponential, Pareto and Weibull panel data models, among others. For illustration, consider the Poisson panel data model with time varying fixed effects

$$y_{it} \sim \text{Poisson}\left(\alpha_{it} \exp\left\{\theta^\intercal x_{it}\right\}\right).$$

The conjugate prior to this likelihood family is

$$\pi(\alpha|\gamma) \equiv \prod_{t=1}^{T} \Gamma(k_t, \beta_t); \quad \gamma := (k_1 - 1, \ldots, k_T - 1, \beta_1, \ldots, \beta_T),$$

where $\Gamma(k, \beta)$ denotes the Gamma distribution with shape parameter k and inverse scale parameter β. The natural statistics of the prior are $(\alpha_{i1}, \ldots, \alpha_{iT}, \ln \alpha_{i1}, \ldots, \ln \alpha_{iT})$. Unlike the linear panel data model, allowing for interactive effects, or correlations between the fixed effects will typically break conjugacy.

2.2.3. *Non-conjugate likelihoods.* In cases where the likelihood does not admit a conjugate prior, we suggest using the normal distributions as the prior family:

$$\pi(\alpha|\gamma) \equiv N(\mu, \Sigma); \quad \gamma := (\Sigma^{-1}\mu, -\Sigma^{-1}/2), u(\alpha) := (\alpha, \alpha\alpha^\intercal).$$

Computing the exact posterior is now more involved. However, we can often approximate the posterior fairly quickly using techniques such as Laplace and local-variational approximations (cf. Section 3.4).

3. EAMP: Generalizing the EM Algorithm

3.1. **EM as Variational Optimization.** Our algorithm makes extensive use of the interpretation of EM as variational optimization, a view first espoused by Neal and Hinton [1998]. Consider the problem of maximizing $\sum_{i=1}^{n} \ln \int q(\mathbf{y}_i, \alpha|\mathbf{x}_i, \theta) d\alpha$ over θ. Neal and Hinton [1998] show that we can rewrite the term inside the summation as

$$\ln \int q(\mathbf{y}_i, \alpha|\mathbf{x}_i, \theta) d\alpha = \max_{q(\cdot)} \left\{ E_{q(\alpha_i)} \left[\ln q(\mathbf{y}_i, \alpha_i|\mathbf{x}_i, \theta)\right] + H(q(\alpha_i)) \right\},$$

where the optimization is a variational one that maximizes over all possible distributions $q(\cdot)$ of α, and $H(q) := -E_{q(\cdot)}\left[\ln q(\alpha)\right]$ denotes the entropy. The value of $q(\cdot)$

that maximizes the above is $q^*(\alpha) = p(\alpha|\mathbf{y}_i, \mathbf{x}_i, \theta)$. We thus obtain

$$\max_{\theta} \sum_{i=1}^{n} \ln \int p(\mathbf{y}_i, \alpha|\mathbf{x}_i, \theta) d\alpha = \max_{q(\cdot),\theta} \sum_{i=1}^{n} \left\{ E_{q_i(\alpha_i)} \left[\ln p(\mathbf{y}_i, \alpha_i|\mathbf{x}_i, \theta) \right] + H(q(\alpha_i)) \right\}.$$

Maximizing the right hand side of the above display equation repeatedly over $\theta, q(\cdot)$ is thus equivalent to the standard EM algorithm.

For our setting, we shall restate the results of Neal and Hinton [1998] in a slightly different form. Let $\mathrm{KL}(p_1||p_2)$ denote the Kullback-Leibler divergence between any two distributions p_1, p_2. Our interest is in integrals of the form $\int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha)d\alpha$, for some prior $\pi(\alpha)$. By the Donsker-Varadhan variational formula, it follows

$$\ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha)d\alpha = \max_{q(\cdot)} \left\{ E_{q(\cdot)} \left[\ln p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta) \right] - \mathrm{KL}(q \mid\mid \pi) \right\}. \tag{3.1}$$

Furthermore, assuming that $p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)$ is upper bounded in the support of π, the value of $q(\cdot)$ that maximizes the right hand side of (3.1) is given by

$$q^*(\alpha) = \frac{p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha)}{\int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha)d\alpha}. \tag{3.2}$$

The above holds for any bounded non-negative function $p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)$.

3.2. **The EAMP Algorithm.** This section describes the EAMP algorithm, focusing on the general approach. Appendix A provides the computational details and pseudo-codes for the various examples in Section 2.2.

Invoking (3.1), we can rewrite the optimization problem of maximizing the integrated likelihood in (2.1) as

$$\begin{aligned} &\max_{\substack{\{w_i(s)\},\\ \theta,\{\gamma_s\}}} \sum_{i=1}^{n} \sum_{s=1}^{S} w_i(s) \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha \\ &= \max_{\substack{\{q_{si}(\cdot)\},\\ \{w_i(s)\},\theta,\{\gamma_s\}}} \sum_{i=1}^{n} \sum_{s=1}^{S} w_i(s) \left\{ E_{q_{si}(\cdot)} \left[\ln p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta) \right] - \mathrm{KL}\left(q_{si}(\alpha) \mid\mid \pi(\alpha|\gamma_s) \right) \right\}. \end{aligned} \tag{3.3}$$

Our extension of the EM algorithm proceeds by maximizing the right hand side of (3.3) repeatedly over $q_{si}(\cdot), w_i(s), \theta, \gamma_s$ - in that order. This is done by cycling repeatedly through the following sequence of steps:

Step E: Expectation. In this step, we update the value of $q_{si}(\cdot)$ by finding the value that maximizes (3.3), conditional on all the other variables. In view of (3.2),

this update is given by

$$q_{si}(\alpha) \leftarrow \frac{p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)}{\int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha}. \tag{3.4}$$

This computation is instantaneous if $\pi(\alpha|\gamma)$ is a conjugate prior for $p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)$.

Step A: Assignment. Next, we assign each observation into one of the groups $\mathcal{S} = \{1, \ldots, S\}$ by maximizing (3.3) with respect to $w_i(s)$. Since we run this step right after updating $q_{si}(\cdot)$, we have that

$$E_{q_{si}(\cdot)}\left[\ln p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\right] - \mathrm{KL}\left(q_{si}(\alpha) \,||\, \pi(\alpha|\gamma_s)\right) = \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha,$$

for each s. Hence, we would assign observation i to the group $s(i)$ by setting

$$s(i) \leftarrow \arg\max_s \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha. \tag{3.5}$$

Computing the above integral directly is computationally expensive, but we can exploit equation (3.2), which ensures

$$\int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha = \frac{p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)}{q_{si}(\alpha)}.$$

Thus, for likelihoods with conjugate priors, we can obtain an analytical expression for this integral using the right hand side of the above equality, and the assignment step is just a matter of evaluating this expression.

Step M: Maximization. We next update the value of θ as

$$\theta \leftarrow \arg\max_\theta \sum_{i=1}^{n}\sum_{s=1}^{S} w_i(s) E_{q_{si}(\cdot)}\left[\ln p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\right]. \tag{3.6}$$

This requires solving an optimization problem. In many examples the conditional likelihood $p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)$ is convex, so computation is typically very fast.

Step P: (Expectation) Propagation. Finally, we update γ_s for each s by choosing the value that maximizes (3.3). This is the same as setting

$$\begin{aligned}\gamma_s &\leftarrow \arg\min_\gamma \sum_{i=1}^{n} w_i(s)\mathrm{KL}\left(q_{si}(\alpha) \,||\, \pi(\alpha|\gamma)\right) \\ &= \arg\min_\gamma -\int \left(\frac{1}{n_s}\sum_{i=1}^{n} w_i(s) q_{si}(\alpha)\right) \ln \pi(\alpha|\gamma)d\alpha,\end{aligned} \tag{3.7}$$

where n_s is the number of observations currently assigned to group s. Define $\bar{q}_s(\alpha) = n_s^{-1}\sum_{i=1}^{n} w_i(s)q_{si}(\alpha)$ as the mixture distribution obtained by averaging the

densities, $q_{si}(\alpha)$, of all observations assigned to the group s. Then,

$$\gamma_s \leftarrow \arg\min_{\gamma} \mathrm{KL}\left(\bar{q}_s(\alpha) \;||\; \pi(\alpha|\gamma)\right). \tag{3.8}$$

Thus the optimal γ_s minimizes the KL divergence between the average posterior and the prior. The solution to this is straightforward when $\pi(\alpha|\gamma)$ is from an exponential family. In this case, the optimal value of γ_s is the one that matches the moments of the sufficient statistics, $u(\cdot)$, of the exponential family - see, e.g., Bishop [2006]:

$$E_{\pi(\cdot|\gamma_s)}[u(\alpha)] = E_{\bar{q}_s}[u(\alpha)].$$

Recalling the definition of $\bar{q}_s(\alpha)$, we thus obtain

$$E_{\pi(\cdot|\gamma_s)}[u(\alpha)] = \frac{1}{n_s}\sum_{i=1}^{n} w_i(s) E_{q_{si}}[u(\alpha)]. \tag{3.9}$$

In other words, this step 'propagates' the moments of $q_{si}(\alpha)$ onto the 'prior' $\pi(\alpha|\gamma_s)$. Computing the expectations $E_{q_{si}}[u(\alpha)]$ and $E_{\pi(\cdot|\gamma_s)}[u(\alpha)]$ is also quite straightforward when the prior and posterior are from the same exponential family (which would be the case if $\pi(\alpha|\gamma)$ were a conjugate prior). Let γ_{is} denote the current 'posterior' values of γ, as implied by $q_{is}(\cdot)$. Then $E_{q_{si}}[u(\alpha)] = \partial_\gamma \ln g(\gamma_{is})$ - the function $g(\cdot)$ is defined in (2.2) - and the updated values of γ_s are the ones that solve

$$\partial_\gamma \ln g(\gamma_s) = \frac{1}{n_s}\sum_{i=1}^{n} w_i(s)\partial_\gamma \ln g(\gamma_{is}).$$

3.3. **Convergence of EAMP.** By construction, each of the steps in EAMP increase the energy functional that is the right hand side expression of (3.3). Furthermore, the functional has an upper bound given by the upper bound of $p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)$. Hence, as long as $p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta) \leq C$ for some $C < \infty$ independent of $(\mathbf{y}_i, \mathbf{x}_i, \alpha_i, \theta)$, the EAMP algorithm will converge to a stationary point. As with the usual EM algorithm, the convergence is only local, and one would need to initialize the algorithm at multiple points to obtain a global solution.

To get good initial estimates, we suggest running GFE first. The GFE estimates and standard errors can then be used as initializations for the prior parameters in GRE.

3.4. **Non-conjugate likelihoods.** When the conditional likelihood is from a non-conjugate family, computing the posterior for the E step is more involved. In such cases, we recommend choosing the prior from the Gaussian family, and approximating the posterior with another Gaussian distribution.

3.4.1. *Laplace approximation.* The Laplace approximation aims to approximate the update of $q_{si}(\cdot)$ with a Gaussian distribution centered around its mode. Recall that $q_{si}(\alpha) \propto p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)$. The mode of this distribution is given by

$$\alpha_{si} = \arg\max_{\alpha} \ln \left\{ p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s) \right\}.$$

Denote the curvature around the mode by

$$A_{si} = -\nabla^2 \ln \left\{ p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s) \right\} \Big|_{\alpha=\alpha_{si}}.$$

Then the Laplace approximation to $q_{si}(\cdot)$ is $\tilde{q}_{si} = N(\alpha_{si}, A_{si})$. The difference between the Laplace approximation and the true posterior is of the order $O(T^{-1})$ when α_i is time-invariant.

3.4.2. *Variational approximations.* The main limitation of the Laplace approximation is that it based solely on the properties of the target distribution around its mode, and fails to capture its global behavior. Variational methods also lead to a Gaussian approximation to the posterior distribution, but offer better accuracy since the approximation uses global information. There exist a number of variational algorithms for this purpose, including local variational approximations (Jaakkola and Jordan [1997]) mean-field Gaussian Variational-Bayes, non-conjugate variational message passing (Knowles and Minka [2011]) and partially non-centered parameterizations (Tan et al. [2013]). In Appendix A, we illustrate these methods using local variational approximations for random coefficients panel Logistic regression, following Jaakkola and Jordan [1997].

3.5. **Selecting the number of groups.** Suppose that we have $\bar{S}$ different models $\mathcal{M}_1, \ldots, \mathcal{M}_{\bar{S}}$ where $\mathcal{M}_S$ denotes the model with S groups. We can use the BIC criterion to select the preferred one as

$$S^* = \arg\min_{S} \left\{ Sq \ln(nT) - 2 \ln \hat{p}(\mathbf{y}|\mathbf{x}, \mathcal{M}_S) \right\},$$

where q is the number of estimated parameters per group, and $\hat{p}(\mathbf{y}|\mathbf{x}, \mathcal{M}_S)$ is the maximized likelihood function

$$\hat{p}(\mathbf{y}|\mathbf{x}, \mathcal{M}_S) := \sum_{i=1}^{n} \sum_{s=1}^{S} w_i(s) \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s) d\alpha.$$

4. Relationship to mean-field Variational-Bayes inference

It is well known that the EM algorithm is closely connected to Variational Inference (Blei et al., 2017). In this section, we extend this notion to our context by showing that the steps in the EAMP algorithm closely resemble the steps for mean-field Variational-Bayes (VB) inference in a probabilistic model with a mixture prior over unobserved heterogeneity. We start with a setting of fixed S, and discuss extensions for infinite number of groups in Appendix B.

To simplify the exposition, we shall suppose that θ is known, and the quantities of interest are the posterior distributions over α_i. From a Bayesian perspective, the assignment vector w_i is a latent variable comprising a 1-of-S binary vector with elements $w_{si} \in \{0, 1\}$ and $\sum_s w_{si} = 1$. Thus we have two latent variables in our setup: (α_i, w_i). We specify a group specific multinomial prior for $\mathbf{w} := \{w_{s1}, \ldots, w_{sn}\}_s$ as

$$p(\mathbf{w}|\mu) = \prod_i \prod_s \mu_s^{w_{si}},$$

where $\{\mu_s : s = 1, \ldots, S\}$ are hyper-parameters denoting mixing coefficients. Conditional on w_i and hyper-parameters $\gamma = (\gamma_1, \ldots, \gamma_S)$, the distribution of $\alpha := (\alpha_1, \ldots, \alpha_n)$ is given by

$$p(\alpha|\mathbf{w}, \gamma) = \prod_i \prod_s \pi(\alpha_i|\gamma_s)^{w_{si}}.$$

Conditional on the latent variables, the distribution of the observed quantities $\mathbf{y} := (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is given by

$$p(\mathbf{y}|\mathbf{x}, \alpha, w) \equiv p(\mathbf{y}|\mathbf{x}, \alpha) = \prod_i p(\mathbf{y}_i|\mathbf{x}_i, \alpha_i).$$

We now introduce hyper-priors over the hyper-parameters μ, γ. Since both $p(w_i|\mu)$ and $\pi(\alpha_i|\gamma_s)$ are members of the exponential family, it is convenient to specify conjugate priors for the hyper-parameters. Hence, we specify a Dirichlet prior for μ

$$p(\mu) = \text{Dir}(\mu|\beta_0) = C(\beta_0) \prod_s \mu_s^{\beta_0 - 1},$$

where β_0 is a pre-specified quantity. Similarly, the distribution of γ_s is specified as

$$p(\gamma) = \prod_s p(\gamma_s); \quad p(\gamma_s) := f(\chi_0, \nu_0) g(\gamma_s)^{\nu_0} \exp(\nu_0 \gamma_s^\intercal \chi_0),$$

where (ν_0, χ_0) are pre-specified quantities.

Based on the above, the joint distribution of all the random variables is given by

$$p(\mathbf{y}, \alpha, w, \mu, \gamma|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}, \alpha)p(\alpha|w, \gamma)p(w|\mu)p(\mu)p(\gamma).$$

The only observed variables are $(\mathbf{y}, \mathbf{x})$. Our interest is in calculating the posterior distribution of the unobservables, denoted as $q^*(\alpha, \mathbf{w}, \mu, \gamma)$. We consider a mean-field approximation to the posterior which factorizes between the latent variables and hyper-parameters $q(\alpha, w, \mu, \gamma) = q(\alpha, w)q(\mu, \gamma)$. The aim then is to find the function $q(\cdot)$ that minimizes the KL divergence, $\text{KL}\left(q(\alpha, w)q(\mu, \gamma) \mid\mid q^*(\alpha, w, \mu, \gamma)\right)$, between the factorized and true posterior. This is in turn equivalent to minimizing the KL divergence from the joint probability $\text{KL}\left(\tilde{q}(\alpha, w)\tilde{q}(\mu, \gamma) \mid\mid p(\mathbf{y}, \alpha, w, \mu, \gamma|\mathbf{x})\right)$. The form of the joint probability and the factorization imply a further, induced, factorization over $q(\mu, \gamma)$, given by $q(\mu, \gamma) = q(\mu)\prod_s q(\gamma_s)$.[3] In contrast to standard mean field methods, we will not further factorize the posterior distribution over the latent variables. However, note that we can always write $q(\alpha, w) = q(\alpha|w)q(w)$. Combining the above, our variational problem is

$$q = \arg\min_{\tilde{q}} \text{KL}\left(\tilde{q}(\alpha|w) \cdot \tilde{q}(w) \cdot \tilde{q}(\mu) \cdot \prod_s \tilde{q}(\gamma_s) \mid\mid p(\mathbf{y}, \alpha, w, \mu, \gamma|\mathbf{x})\right). \tag{4.1}$$

The factorization between latent variables and hyper-parameters follows a well studied idea in the Variational-Bayes literature. Wang and Blei [2019] show that if there exists some 'true' set of hyper-parameters, (μ^*, γ^*), from which the data is generated, then the factorized posterior, $q(\mu, \gamma)$, concentrates around (μ^*, γ^*) at $\sqrt{n}$ rates. This also implies that the factorized posterior $q(\alpha, w)$ is close to the 'true' posterior of these quantities if one knew (μ^*, γ^*).

We maximize (4.1) by sequentially maximizing over $\tilde{q}(\alpha|w)$, $\tilde{q}(w)$, $\tilde{q}(\mu)$ and $\{\tilde{q}(\gamma_s)\}_s$ holding the other quantities fixed. These correspond almost exactly to the steps E, A, P in the EAMP algorithm (we leave out the M step as our model omitted the common parameters θ).

4.1. **The VB-EAMP Algorithm.** We describe the steps of the Variational-Bayes EAMP algorithm, relegating the derivation to Appendix B. The derivation of the P step is based on standard mean-field VB methods (see, e.g., Winn and Bishop [2005] and Blei et al. [2017]), but the derivation of the E and A steps is new.

[3]This is easily verified using graphical methods, see Bishop (2006).

Step E: Updating $q(\alpha|w)$. The update rule for $q(\alpha|w)$ is given by:

$$q(\alpha|w) \leftarrow \prod_i \prod_s q_{si}(\alpha_i)^{w_{si}}, \tag{4.2}$$

where

$$q_{si}(\alpha) := \frac{p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\bar{\pi}_s(\alpha)}{\int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\bar{\pi}_s(\alpha)d\alpha}, \text{ and} \tag{4.3}$$

$$\bar{\pi}_s(\alpha) = h(\alpha)g(\bar{\gamma}_s)\exp\left(\bar{\gamma}_s^\intercal u(\alpha)\right); \ \bar{\gamma}_s := E_{q(\gamma_s)}\left[\gamma_s\right]. \tag{4.4}$$

The Variational-Bayes E step is the same as the E-step in EAMP, but with γ_s replaced by its posterior mean $\bar{\gamma}_s$.

Step A: Updating $q(w)$. The update rule to $q(w)$ is given by

$$q(w) \leftarrow \prod_i \prod_s r_{si}^{w_{si}}, \tag{4.5}$$

where r_{si} are the 'responsibilities'

$$r_{si} = \frac{\exp\left\{E_{q(\mu_s)}\left[\ln \mu_s\right] + \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha)\bar{\pi}_s(\alpha)d\alpha\right\}}{\sum_s \exp\left\{E_{q(\mu_s)}\left[\ln \mu_s\right] + \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha)\bar{\pi}_s(\alpha)d\alpha\right\}}. \tag{4.6}$$

As in the MLE procedure, it is not necessary to evaluate the integrals in (4.6) directly since we can obtain analytical expressions for them using the knowledge of $q_{si}(\alpha)$ from (4.3). Computation of $E_{q(\mu_s)}\left[\ln \mu_s\right]$ will also be similarly straightforward once we derive the form of $q(\mu_s)$ below. We observe that the Variational-Bayes A step induces a soft, i.e probabilistic, assignment of individuals into groups, as opposed to the hard assignment in MLE.

Step P: Updating $q(\mu)$ and $q(\gamma)$. The update to $q(\mu)$ is given by

$$q(\mu) = \text{Dir}(\mu|\boldsymbol{\beta}), \tag{4.7}$$

where $\boldsymbol{\beta}$ is a $S \times 1$ vector with $\beta_s = \beta_0 + \sum_i r_{si}$. The above allows us to obtain an analytical expression for $E_{q(\mu_s)}\left[\ln \mu_s\right]$ used in Step A as $E_{q(\mu_s)}\left[\ln \mu_s\right] = \psi(\beta_s) - \psi(\sum_s \beta_s)$, where $\psi(\cdot)$ denotes the Digamma function. Finally, $q(\gamma_s)$ is updated as

$$q(\gamma_s) = f(\chi_s, \nu_s)g(\gamma_s)^{\nu_s}\exp(\nu_s\gamma_s^\intercal\chi_s), \tag{4.8}$$

where

$$\nu_s = \nu_0 + \sum_i r_{si}, \text{ and } \nu_s\chi_s = \chi_0 + \sum_i r_{si}E_{q_{si}(\alpha)}\left[u(\alpha_i)\right].$$

This has clearly the same form as the Propagation step in (3.9) except for having (ν_0, χ_0) as an additional pseudo-observation.

4.2. **MLE vs VB inference.** The primary difference between MLE and mean-field VB inference lies in the treatment of w_i.[4] The MLE approach treats $\{w_i\}_i$ as n incidental parameters, estimated from the time series of each observation. Bayesian inference allows group assignment to be probabilistic, which can be seen as effectively regularizing w_i. The resulting estimates for marginal effects are also less sensitive to the exact estimation of group structure. Indeed, the posterior average estimators in the VB setting are given by

$$\hat{\tau}_b = \frac{1}{n}\sum_i \sum_s r_{si} E_{q_{si}(\alpha_i)}\left[\psi(\mathbf{y}_i, \mathbf{x}_i, \alpha_i)\right] = \frac{1}{n}\sum_i E_{q(\alpha_i, w_i)}\left[\psi(\mathbf{y}_i, \mathbf{x}_i, \alpha_i)\right], \tag{4.9}$$

where r_{si} and $q_{si}(\cdot)$ are defined in (4.6) and (4.3). Compared to the estimator (2.4) based on MLE, we see that $\hat{\tau}_b$ takes a weighted average of the contribution across all possible groups, instead of focusing only on the best one. This is a more accurate characterization of observations that lie on the boundaries between groups.

5. Theoretical properties

To motivate the GRE approach, we start by noting that we can always represent the conditional density $p(\alpha_i|\mathbf{x}_i)$ as a continuous mixture of the form

$$p(\alpha_i|\mathbf{x}_i) = \int_\Xi \pi\left(\alpha_i|\gamma(\xi)\right) p(\xi|\mathbf{x}_i) d\xi. \tag{5.1}$$

Here, $\pi(\alpha|\cdot)$ denotes the prior family used in our GRE specification, while $p(\xi|\mathbf{x}_i)$ denotes some unknown and un-restricted mixing weights. We represent the latent variable by ξ_i instead of $\gamma_i \equiv \gamma(\xi_i)$ as this allows the dimension, d, of ξ_i to be potentially smaller than that of $\gamma(\cdot)$. Our theoretical results below show that the GRE estimator adapts to the unknown value of d.

The mixture representation implies that α_i is independent of $\mathbf{x}_i$ conditional on ξ_i. As currently written, equation (5.1) always holds (for a reasonable choice of prior family, e.g., Gaussian), but in our formal assumption below, we will require ξ_i to lie in a compact set:

Assumption 1. *There exists an individual specific latent variable ξ_i such that (i) α_i is independent of $\mathbf{x}_i$ conditional on ξ_i; (ii) the probability distribution, $p(\alpha_i|\xi_i)$, of α_i*

[4]Bayesian inference also adds a hyper-prior on γ_s, but this only serves to add an additional pseudo-observation. Thus the effect of this is negligible for large n.

given ξ_i is a member of the exponential family $\pi(\alpha|\cdot)$, i.e, $p(\alpha_i|\xi_i) = \pi(\alpha_i|\gamma(\xi_i))$ for some Lipschitz continuous mapping $\gamma(\cdot)$; and (iii) the domain Ξ of ξ_i is a compact set of fixed dimension d and there exists $C < \infty$ such that $\sup_{\xi\in\Xi} |\partial^2_\gamma \ln \pi(\alpha|\gamma(\xi))| \leq C$.

The only additional requirement imposed by Assumption 1 over (5.1) is for the domain of ξ to be compact. Consider, for instance, the setting where $\pi(a|\gamma) \equiv N(\mu, \Sigma)$. The requirement of $\sup_{\xi\in\Xi} |\partial^2_\gamma \ln \pi(\alpha|\gamma(\xi))| \leq C$ is equivalent to imposing that the eigenvalues of Σ are bounded away from 0 and ∞. This implies that the conditional density $p(\alpha_i|\mathbf{x}_i)$ will be super-smooth with a bounded range for the standard-deviation (Ghosal et al., 2007). While somewhat restrictive, we note that it is in fact possible to approximate any twice differentiable $\pi(\alpha|\mathbf{x})$ with a continuous mixture of normals by letting $\dim(\xi) = \dim(\gamma(\cdot))$, and $\lambda_{\min}(\Sigma) \to 0$ (here $\lambda_{\min}(A)$ denotes the lowest eigenvalue of A). Thus, the scope of Assumption 1 can be potentially extended, though we leave this as an avenue for future research.

Note that $\alpha_i, \mathbf{x}_i$ are, in general, vectors that grow with T, so ξ_i could be potentially changing with T as well (but we leave this dependence implicit). However, the assumption of compact support for ξ is made uniformly for all T.

Assumption 1 imposes significantly fewer restrictions on the distribution of covariates $\mathbf{x}_i$ than those imposed by Bonhomme et al. [2017] for two-stage fixed-effects. In that paper, the authors specify a hidden Markov structure for $\mathbf{x}_i$ of the form $p(\mathbf{x}_i) = \prod_t p(x_{it}|x_{it-1}, \mu_{it})$, where μ_{it} is a latent variable, and assume that the latent variables can be written as $\alpha_{it} = \alpha(\xi_i, \lambda_t)$ and $\mu_{it} = \mu(\xi_i, \lambda_t)$ for some random variable λ_t. This implies low dimensional heterogeneity in both α_i and $\mathbf{x}_i$. By contrast, Assumption 1 does not impose any restriction on the heterogeneity of $\mathbf{x}_i$; note also that we make no restrictions on the distribution $p(\xi_i|\mathbf{x}_i)$. Therefore, it is less restrictive in settings where x_{it} is high dimensional.

We study the properties of the GRE estimator in two regimes:

5.1. **Discrete ξ_i .** Suppose Ξ is a discrete set of cardinality S. In view of (5.1), this implies that $p(\alpha_i|\mathbf{x}_i)$ can be represented as a finite mixture of densities, with an unrestricted form for the mixture weights. In fact, this assumption is relatively mild, since Norets and Pelenis [2014] show that a finite mixture of normals with unrestricted mixing weights can approximate a large class of conditional densities. We note also that α_i still has continuous support; it is only the support of ξ_i that is discrete.

We suppose that the number of groups (i.e., the number of mixture components) S is known beforehand, and denote the group specific hyper-parameters by $\gamma_1^*, \ldots, \gamma_S^*$. To simplify the analysis, we shall also suppose that we may partition any candidate γ_s into time-specific elements $\gamma_s = (\gamma_s^{(1)}, \ldots, \gamma_s^{(T)})$ such that

$$\pi\left(\alpha_i|\gamma_s\right) = \prod_t \pi(\alpha_{it}|\gamma_s^{(t)}).$$

The above ensures that the treatment of γ is no different than that of α in GFE; hence we can reuse the theoretical results that have already been derived for the latter (see, e.g., Bonhomme and Manresa [2015], Cheng et al. [2019]). Indeed, the only difference is that we now work with the integrated likelihood

$$l_i(\gamma, \theta) := \frac{1}{T}\sum_t \ln \int p(y_{it}|x_{it}, \alpha, \theta)\pi(\alpha|\gamma^{(t)})d\alpha$$

instead of $l_i(\alpha, \theta) := T^{-1}\sum_t \ln p(y_{it}|x_{it}, \alpha)$ for GFE. We also note that the techniques can be easily extended to allow correlation in the prior across time at the expense of more involved proofs.

We state the relevant regularity conditions, which are similar to that for GFE, in Appendix D. Let $\hat{\theta}$ denote the parameter estimate from GRE, and θ_0 the true value. We have the following result:

Theorem 1. *Suppose Assumption 1 and Assumption C in Appendix D hold. Then,*

$$\sqrt{nT}(\hat{\theta} - \theta_0) \Longrightarrow N(0, V),$$

where the covariance matrix V is described in Appendix D. Furthermore, letting $s^*(i)$ and $\hat{s}_i$ denote the true and estimated group assignments,

$$\max_{1\leq i\leq n} \mathbb{P}\left(\hat{s}_i \neq s_i^*\right) = o(n^{-1}).$$

The proof of Theorem 1 is a straightforward application of the methods of Cheng et al. [2019]. We omit the details. The GRE estimator gives rise to the same rate of estimation of θ_0 as GFE. However, it bears emphasizing that it does so under much weaker conditions.

5.2. **Continuous ξ_i.** We now turn to a setting where the unobserved heterogeneity ξ_i is continuous, but assume that the mapping $\gamma(\xi_i)$ is of a fixed dimension. The GRE estimator can then be thought of as discretizing the space, Ξ, of ξ.

We employ the following notation: Let $p(\xi_i)$ denote the population distribution of ξ_i; $p(\mathbf{x}_i|\xi_i)$ the conditional distribution of $\mathbf{x}_i$ given ξ_i and $p(\mathbf{y}_i|\mathbf{x}_i, \xi_i, \theta) :=$

$\int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma(\xi_i))d\alpha$. We shall also let $\mathbb{P}$ and $\mathbb{E}[\cdot]$ denote the probability measure and expectation corresponding to the joint distribution $p(\mathbf{y}_i|\mathbf{x}_i, \xi_i, \theta)p(\mathbf{x}_i|\xi_i)p(\xi_i)$ over $(\mathbf{y}_i, \mathbf{x}_i, \xi_i)$. The joint probability over $(\mathbf{y}_i, \mathbf{x}_i)$ for observation i (i.e, conditional on ξ_i) is given by $p(\mathbf{y}_i, \mathbf{x}_i|\xi_i, \theta) \equiv p(\mathbf{y}_i|\mathbf{x}_i, \xi_i, \theta)p(\mathbf{x}_i|\xi_i)$, shorthanded as $p(\cdot|\xi_i, \theta)$. The integrated log-likelihood is denoted by

$$l_i(\gamma, \theta) := \frac{1}{T} \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma)d\alpha.$$

Note that $l_i(\gamma(\xi_i), \theta) = \ln p(\mathbf{y}_i|\mathbf{x}_i, \xi_i, \theta)$. Let $\Gamma_S := (\gamma_1, \ldots, \gamma_S)$ denote a vector of group-specific parameters. We then define

$$\begin{aligned} \hat{\Gamma}_S(\theta) &:= \arg\max_{\Gamma_S} \sum_i \max_{\gamma \in \Gamma_S} l_i(\gamma, \theta), \text{ and} \\ \hat{\gamma}_{i,S}(\theta) &:= \arg\max_{\gamma \in \hat{\Gamma}_S(\theta)} l_i(\gamma, \theta). \end{aligned} \tag{5.2}$$

Our first result concerns the closeness of the estimates $\hat{\gamma}_{i,S}(\hat{\theta})$ to $\gamma(\xi_i)$. The relevant regularity conditions are detailed in Appendix E.

Lemma 1. *Suppose Assumption 1 and Assumption D1 in Appendix E hold. Then, for $S, T, n \to \infty$,*

$$\left\|\hat{\theta} - \theta_0\right\|^2 + \frac{1}{n}\sum_i \left\|\hat{\gamma}_{i,S}(\hat{\theta}) - \gamma(\xi_i)\right\|^2 = O_{\mathbb{P}}\left(\frac{1}{T} + S^{-2/d} + \frac{S^{-1/d}}{\sqrt{nT}}\right).$$

Lemma 1 also provides a rate for estimation of θ_0, but this is not sharp, as we will see shortly. The rate in Lemma 1 is similar in form to that obtained in Bonhomme et al. [2017]. This is not too surprising since (as in the previous sub-section) γ plays the same role as α in GFE. This superficial similarity in rates however masks two important differences:

First, as discussed earlier, Assumption 1 does not restrict heterogeneity in $\mathbf{x}_i$. This implies that the dimension d of ξ_i can be potentially smaller than that required in Bonhomme et al. [2017]. This weaker requirement is allowed due to the joint estimation of groups and common parameters. An analysis of the proof of Lemma 1 shows that this advantage applies to 'one-step' GFE as well.

Second, the dimension d is smaller for GRE as compared to one-step GFE. In GFE, observations with the same value of ξ_i must have the same α_i, whereas the GRE approach allows for variability in α_i conditional on ξ_i. Hence, the random effects specification soaks up some of the variability in α_i, which would otherwise increase the dimension, d, of ξ_i in GFE.

We now turn to the estimation of marginal effects, τ_0, using (2.4). The estimation of the common parameters θ_0 is a special case since τ_0 is a function of both θ_0 and α_i. To obtain a sharp bound on the rate of estimation of τ_0 , we shall assume that α_i is time-invariant. Unfortunately, our current techniques cannot handle time-varying α_i as they rely on Laplace approximation arguments.

Assumption 2. *The unobserved heterogeneity α_i is time-invariant.*

We also require additional regularity conditions that are detailed in Appendix E.

Theorem 2. *Suppose Assumptions 1, 2 and D1 - D3 in Appendix E hold. Then, for $S, T, n \to \infty$,*

$$\hat{\tau} - \tau_0 = \frac{1}{n} \sum_i s_i^{(\tau)} + O_{\mathbb{P}}\left(\frac{1}{T^{3/2}} + \frac{S^{-1/d}}{T}\right) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{nT}}\right),$$

where $s_i^{(\tau)}$, described in Appendix E, is a score function satisfying $E[s_i^{(\tau)}] = 0$.

With a suitable choice of S, we can make $S^{-1/d} \asymp T$. Hence the bias from GRE is of the order $T^{-3/2}$. This is faster than the T^{-1} rate achieved by Bonhomme et al. [2017] for two-stage GFE. It is interesting to observe that the bias is not affected by n (in fact a similar situation occurs in GFE as well, see Bonhomme et al. [2017]).

Even though our theory does not suggest an upper bound on S, we do not recommend taking $S = n$. If there are really a finite number of groups as in Section 5.1, then it is more efficient to set S to that number. Our analysis also does not consider second order terms that might become relevant with large S. For these reasons, we suggest choosing S according to some information criterion, such as BIC.

The intuition behind Theorem 2 is as follows: Suppose that τ_0 does not depend on θ_0. Then, by a Laplace approximation argument as in Arellano and Bonhomme [2009], we can show

$$\hat{\tau} - \tau_0 = \frac{1}{n} \sum_i s_i^{(\tau)} + \frac{1}{nT} \sum_i B_i,$$

where B_i is a bias term that depends on the discrepancy between $\pi\left(\cdot|\hat{\gamma}_{i,S}(\theta_0)\right)$ and $\pi(\cdot|\gamma(\xi_i))$. The result then follows from Lemma 1. Obtaining the rates for estimation of θ_0, however, requires new techniques since the maximizers $\hat{\theta}, \{\hat{\gamma}_s\}_s, \{\hat{w}_i(s)\}_{i,s}$ of the integrated log-likelihood (2.1) are potentially discontinuous with respect to the data. Our proof exploits the structure of the EAMP algorithm, which implies $\hat{\theta}$ can

be characterized as the solution (in terms of θ) to:

$$\frac{1}{n}\sum_i E_{\hat{q}_i(\alpha)}\left[\partial_\theta \ln p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta)\right]=0,$$

where $\hat{q}_i(\cdot) := q(\cdot|\hat{\gamma}_{i,S}(\hat{\theta}),\hat{\theta})$ with $q_i(\alpha|\gamma,\theta)$ defined as

$$q_i(\alpha|\gamma,\theta) := \frac{p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta)\pi(\alpha|\gamma)}{\int p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta)\pi(\alpha|\gamma)d\alpha}.$$

We then use a Laplace approximation argument to compare $\hat{\theta}$ to the solution $\tilde{\theta}$ of the infeasible equation

$$\frac{1}{n}\sum_i E_{q_i(\alpha)}\left[\partial_\theta \ln p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta)\right]=0,$$

where $q_i(\alpha) := q(\cdot|\gamma(\xi_i),\theta_0)$ is the infeasible posterior.

6. Empirical examples

We illustrate our methodology by applying it on two datasets: The first studies heterogeneity in income dynamics using PSID data. The second studies heterogeneity in the effect of union status on hourly wages.

6.1. **Income dynamics.**

In analogy with the standard model of income dynamics, we decompose income (y_{it}) into deterministic ($\theta_0+\theta_1\{t-1\}$), persistent (η_{it}) and transitory (ε_{it}) components:

$$y_{it}=\theta_0+\theta_1(t-1)+\alpha_{it}+\varepsilon_{it}. \tag{6.1}$$

The persistent component is further decomposed as

$$\alpha_{it}=\rho_i\alpha_{it-1}+v_{it}.$$

where v_{it} is a persistent shock and ρ_i determines the degree of persistence. We assume that ε_{it} is conditionally independent of all the past income and income shocks. Similar to Storesletten et al. [2004], we specify a normal distribution for the transitory and permanent shocks: $\varepsilon_{it}\sim N(0,\sigma^2_{\varepsilon i})$, $v_{it}\sim N(0,\sigma^2_{vi})$. For the initial condition on α_{it}, we impose the stationary distribution $\alpha_{i1}\sim N(0,\sigma^2_{vi}/(1-\rho_i^2))$. This is a strong assumption, but it does simplify the analysis considerably.[5] In the standard model, $\rho_i,\sigma^2_{\varepsilon i},\sigma^2_{vi}$ are assumed to be constant across individuals (the

[5]Alan et al. [2018] employ an initial condition that roughly translates to $\alpha_{1i}\sim b_0+\exp(b_1)N(0,\sigma^2_{vi}/(1-\rho_i^2))$ in our setting, where b_0,b_1 are common parameters estimated from the data. They find that the hypothesis of $b_1=0$ cannot be rejected.

canonical permanent-transitory income model further assumes $\rho_i = 1$). Here, we substantially generalize this by allowing these parameters to vary by household.

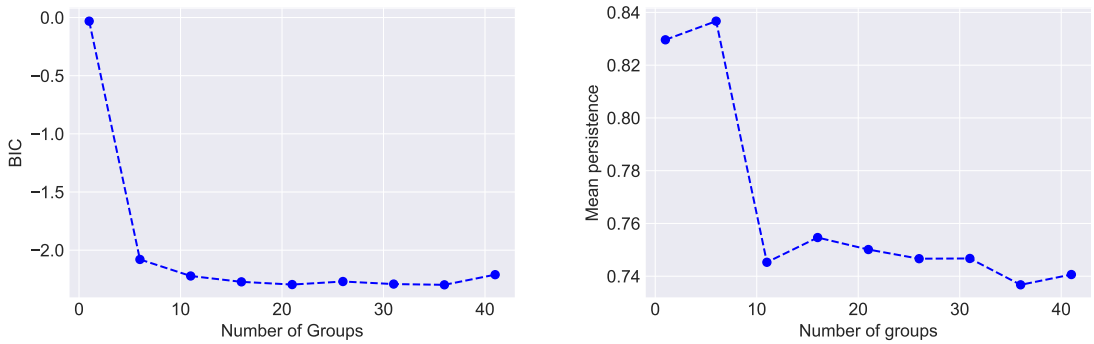
Our methodology does not allow us to distinguish transitory income shocks and measurement error, so the ε_{it} terms should be treated as incorporating both. In the literature, it is common to employ an MA(1) structure for ε_{it}, but as noted by Arellano et al. [2017], the biennial nature of PSID data would appear to make the conditional independence assumption plausible.

Clearly, equation (6.1) falls into our setup, with the model implying an AR(1) prior on unobserved heterogeneity as described in Section 2.2.1. The GRE approach groups observations so that $\rho_i, \sigma^2_{\varepsilon i}, \sigma^2_{vi}$ are approximately constant within each group. Note that the objects of interest here are the prior parameters themselves. For this reason we will not be able to achieve bias reduction in the continuous setting, though the bias would still be negligible if the true number of groups were really finite.

It has been noted by a number of authors (e.g., Arellano et al. [2017], Arellano and Bonhomme [2019]) that the assumption of normal errors is not supported by the data when $S = 1$. However, as we increase the number of groups, we can expect that the GRE would capture more of the variability in the data, and the normality assumption would be more tenable; the rationale is similar to how a mixture of normals can approximate arbitrary densities. Thus, even though we use normal priors, our procedure allows the overall density of the income shocks v_{it}, ε_{it} to be substantially different from a normal and our results below will demonstrate this.

6.1.1. *Data.* We use PSID data that samples individuals every two years between 1999 and 2017. Earnings is defined as the sum of earnings of the married couples divided by the price index. To avoid oversampling low-income households, those from the Survey of Economic Opportunity are removed. We also only consider the subset of married households with male household heads whose age is between 20 and 65. From this we further construct a balanced panel by only keeping individuals with observations on earnings over all waves. The earnings data is then purged of observed heterogeneity by taking the residuals from the regression of log-earnings on the ages of the head and wife, education of the head and wife, and indicators for state, family size, number of children, race of head, and child support.[6] The final panel consists of $n = 604$ individuals and $T = 10$ waves.

[6]Taking first stage residuals is a common approach in this literature, though it is not uncontroversial, see Alan et al. [2018]. We do not, however, employ time dummies, or fixed effects in this regression as that would bias the measurement of heterogeneity.



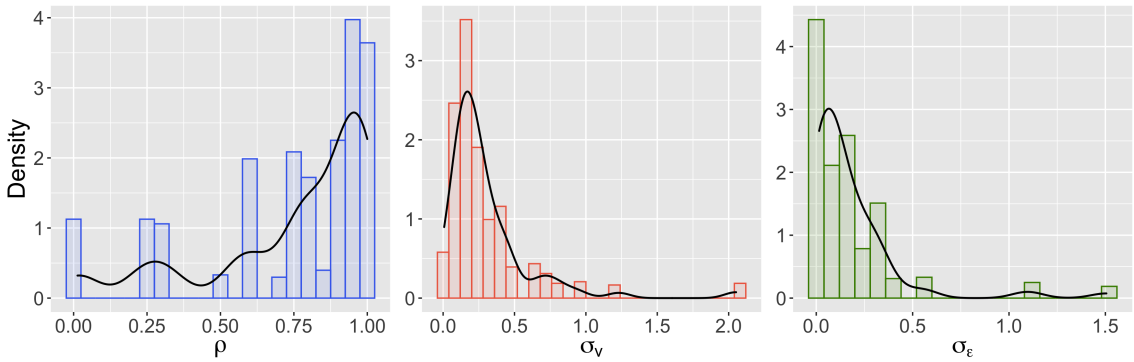
Note: Each point corresponds to running EAMP for 1500 iterations with 10 random initializations.

FIGURE 6.1. Choice of number of groups

6.1.2. *Results.* The number of groups for the EAMP algorithm was estimated using BIC, see Figure 6.1. This results in around 21 groups. Our estimates of the marginal effects however seem to be quite stable for all values of S beyond $S = 16$, as illustrated in the second panel of the figure using $\mathbb{E}[\rho_i]$; similar plots may be obtained for all the other estimates in this section as well.

Figure 6.2 displays the distribution of $\rho_i, \sigma_{vi}, \sigma_{\varepsilon i}$ that results from running the EAMP algorithm with $S = 21$ groups. Table 1 reports the corresponding summary statistics. We find a high amount of heterogeneity in both the magnitude and persistence of income shocks. The median value of the AR parameter is 0.85, which is exactly the same as the estimate of Alan et al. [2018]. However, we find greater heterogeneity: our estimate for the 10th quantile is 0.237, while Alan et al. [2018] find this to be 0.51. It appears that while ρ_i is close to unity for many households (more than 25% have a value greater than 0.95), a majority actually face much less persistent shocks. This suggests that the canonical model, which takes $\rho_i = 1$, may be not be a good approximation of the data. We also find substantial heterogeneity in the standard deviations of persistent and transitory income shocks. This implies that the welfare impact of social insurance policies should vary substantially across households, as social insurance is more beneficial to households facing larger shocks.

From our GRE estimates, we can compute the overall densities of income shocks by averaging the priors, e.g., $f_v^{(\text{prior})}(\cdot) = n^{-1} \sum_s n_s N(\cdot|0, \sigma_{vs}^2)$, where n_s is the number of observations in group s and σ_{sv} is the group-specific standard-deviation of v_{it}. In fact, better density estimates can be obtained using posterior averaging. For each i, t, we can estimate the posterior density, $q_{it}(v_{it})$, of v_{it} by writing $v_{it} = \alpha_{it+1} - \rho_{s^*(i)}\alpha_{it}$ and using the posterior distribution of α_i. Averaging these posteriors gives another



Note: The bars represent density histograms, and the dark line is the kernel smoothed density estimate.

FIGURE 6.2. Distribution of household-specific parameters

TABLE 1. Estimation results

1. Marginal effects						
		Mean	Std. dev	q_{10}	q_{50}	q_{90}
ρ	AR Parameter	0.749	0.284	0.237	0.851	0.978
σ_v	Std. dev of persistent shock	0.303	0.311	0.089	0.189	0.660
σ_ε	Std. dev of temporary shock	0.171	0.245	0.015	0.134	0.304
2. Common parameters						
θ_0	Intercept	-0.087				
θ_1	Trend	-0.008				

Notes: The group size is $S = 21$. The results were obtained after running EAMP algorithm for 2000 iterations with 20 random initializations.

estimate of the density of the persistent shock: $f_v^{(\text{post})}(\cdot) = (n(T-1))^{-1} \sum_i \sum_{t>1} q_{it}(\cdot)$. Similarly, the posterior density estimate of ε can be computed using $\varepsilon_{it} = y_{it} - \theta_0 - \theta_1(t-1) - \alpha_{it}$. Figure 6.3 plots the density estimates of v, ε computed using both prior and posterior averaging. There is not much difference between the two, which is indicative of mis-specification not being a major issue. The overall densities appear far from normal, and the shape is also qualitatively similar to the estimates of Arellano and Bonhomme [2019], even though the authors there impose $\rho = 1$. The figures suggest that heterogeneity in the distribution of income shocks can explain their observed non-normal distributions.

Heterogeneity may also play a role in the observed nonlinear persistence of income shocks. Arellano et al. [2017] report that different magnitudes of income shocks are associated with different degrees of persistence. However, in a heterogenous setting,

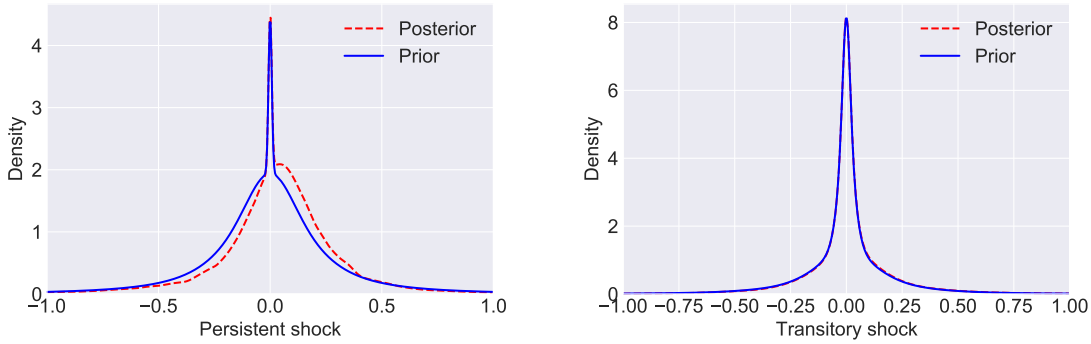


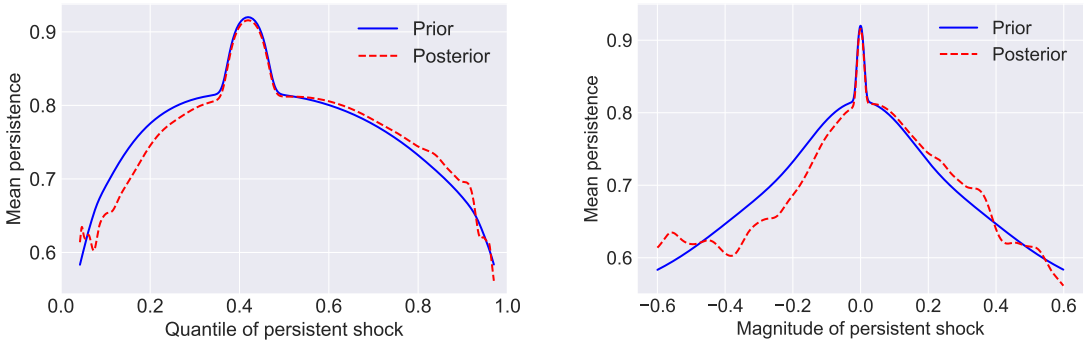
FIGURE 6.3. Densities of income shocks

this could also be a result of aggregation across households. Suppose, for instance, that households with higher standard deviations of income shocks also face lower persistence. Then, in the aggregate, we would find low persistence for high quantiles of income shocks, even though all shocks really have the same persistence within a household. Our results suggest that this may indeed be a possibility. Following Arellano et al. [2017], we introduce a measure of aggregate persistence,

$$W(\tau) := \mathbb{E}\left[\frac{\partial \alpha_{it+1}}{\partial \alpha_{it}}\middle| v_{it} = F_v^{-1}(\tau)\right] = \mathbb{E}\left[\rho_i | v_{it} = F_v^{-1}(\tau)\right],$$

defined as the expected persistence of an income shock at the τ-th quantile. We then estimate $W(\cdot)$ using prior and posterior averaging. For instance, the posterior average estimate is given by $W^{(\text{post})}(\tau) = \sum_{i,t>1} \rho_{s^*(i)} \left\{ q_{it}(\hat{V}_\tau) / \sum_{i,t>1} q_{it}(\hat{V}_\tau) \right\}$, where $\hat{V}(\tau)$ is the τ-th quantile of income shocks, computed using the aggregate density function $f_v^{(\text{post})}(\cdot)$ defined above. The estimates are displayed in Figure 6.4, where we also plot expected persistence versus the magnitude of income shocks. We find that, in the aggregate, larger shocks (corresponding to households with higher σ_v) are less persistent.[7] Our posterior estimates also suggest that positive income shocks are slightly more persistent than negative ones. These results suggest a possible role for heterogeneity in explaining nonlinear persistence. However, with short panels, we cannot definitively rule out nonlinear persistence to shocks within a household as well. A formal test of the two hypotheses would be useful, but we do not pursue it here.

[7] Arellano et al. [2017, Figure 2(d)] report average persistence at various quantiles of both v_{it} and α_{it}. While do we not vary α_{it}, our results are qualitatively similar to their plots for values of α_{it} in the middle quantiles.



Note: The quantiles in the first panel are computed using the posterior average density. A shock of magnitude 0 corresponds to the 42nd quantile (due to the skewness of the posterior).

FIGURE 6.4. Nonlinear persistence of income shocks

6.1.3. *Alternative specifications.* An alternate view of income dynamics (e.g., Guvenen [2007]) posits that individuals face very different life-cycle profiles, so that

$$y_{it} = \theta_{it} + \alpha_{it} + \varepsilon_{it},$$

where θ_{it} denotes time and individual specific growth rates to income. We estimate this model using a GRE specification where $\theta_{it}, \rho_i, \sigma^2_{vi}, \sigma^2_{\varepsilon i}$ are constant within a group. Prior literature often assumed $\theta_{it} := \theta_{0i} + \theta_{1i}(t-1)$, but our clustering techniques allow for time varying terms as well, so we will keep the most general specification. The results are provided in Appendix ??.

6.2. **Effect of union status on hourly wages.** Our next empirical example studies heterogeneity in the effect of collective bargaining coverage on hourly wages. A number of earlier studies, including Chamberlain [1994], Card [1996], Chernozhukov et al. [2013] and Graham et al. [2018] have found evidence for substantial heterogeneity in this setting. We use panel data on hourly wages to estimate the following model:

$$y_{it} = \alpha_{0i} + \alpha_{1i} s_{it} + \theta^{\intercal} x_{it} + \epsilon_{it}.$$

Here, y_{it} denotes the hourly wage of individual i at time t, s_{it} is a binary variable indicating union status, and x_{it} is a collection of covariates. The terms α_{0i}, α_{1i} are individual-specific. The first term is an individual fixed effect for wages, while the second term, α_{1i} captures the heterogeneous response of union status on wages. We assume that ϵ_{it} is uncorrelated with all past values of y_{it}, s_{it}, x_{it} .

6.2.1. *Data.* We employ data on male respondents from the 1979 cohort of the NLSY survey. We use the same sample selection as in Graham et al. [2018]. In

particular, we exclude individuals from supplementary samples of poor whites and military personnel, self-employed individuals, and individuals with reported hourly wages less than 1$ or greater than 1000$. From this we further extract a balanced panel of respondents who had complete wage and union coverage information, and who changed union status at some point in the years 1987-1992. The hourly wages and union status are obtained from CPS data on these respondents. The final panel consists of 634 individuals over $T = 6$ waves. We did not increase the time periods further as this results in substantial attrition (however we do use data from another additional year, 1987, as compared to Graham et al. [2018]). We use the following variables as covariates: time dummies for each year, indicators for Black and Hispanic, years of education, and quantiles of AFQT test scores.

6.2.2. *Results.* The number of groups for the EAMP algorithm was estimated using BIC. This results in 7 groups. TBC

References

Sule Alan, Martin Browning, and Mette Ejrnæs. Income and consumption: A micro semistructural analysis with pervasive heterogeneity. *Journal of Political Economy*, 126(5):1827–1864, 2018.

Tomohiro Ando and Jushan Bai. Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191, 2016.

Manuel Arellano and Stéphane Bonhomme. Robust priors in nonlinear panel data models. *Econometrica*, 77(2):489–536, 2009.

Manuel Arellano and Stéphane Bonhomme. Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, 79 (3):987–1020, 2012.

Manuel Arellano and Stéphane Bonhomme. Recovering latent variables by matching. *arXiv preprint arXiv:1912.13081*, 2019.

Manuel Arellano, Richard Blundell, and Stéphane Bonhomme. Earnings and consumption dynamics: a nonlinear panel data framework. *Econometrica*, 85(3): 693–734, 2017.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877, 2017.

Richard Blundell, Luigi Pistaferri, and Ian Preston. Consumption inequality and partial insurance. *American Economic Review*, 98(5):1887–1921, 2008.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

Stéphane Bonhomme and Martin Weidner. Posterior average effects. *arXiv preprint arXiv:1906.06360*, 2019.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2019-16), 2017.

David Card. The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica: Journal of the Econometric Society*, pages 957–979, 1996.

Gary Chamberlain. Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics*, 10(1):20–26, 1992.

Gary Chamberlain. Quantile regression, censoring, and the structure of wages. 1994.

Xu Cheng, Frank Schorfheide, and Peng Shao. Clustering for multi-dimensional heterogeneity, 2019.

Victor Chernozhukov, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2): 535–580, 2013.

Subhashis Ghosal, Aad Van Der Vaart, et al. Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.

Bryan S Graham, Jinyong Hahn, Alexandre Poirier, and James L Powell. A quantile correlated random coefficients panel data model. *Journal of Econometrics*, 206 (2):305–335, 2018.

Fatih Guvenen. Learning your earning: Are labor income shocks really very persistent? *American Economic Review*, 97(3):687–712, 2007.

Jinyong Hahn and Hyungsik Roger Moon. Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(3):863–881, 2010.

James Heckman and Burton Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320, 1984.

Keisuke Hirano. Semiparametric bayesian inference in autoregressive panel data models. *Econometrica*, 70(2):781–799, 2002.

Mark Huggett, Gustavo Ventura, and Amir Yaron. Sources of lifetime inequality. *American Economic Review*, 101(7):2923–54, 2011.

Tommi Jaakkola and Michael Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, volume 82, 1997.

David A Knowles and Tom Minka. Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*, pages 1701–1709, 2011.

Tony Lancaster. Orthogonal parameters and panel data. *The Review of Economic Studies*, 69(3):647–666, 2002.

Wooyong Lee. Identification and estimation of dynamic random coefficient models. 2019.

Laura Liu. Density forecasts in panel data models: A semiparametric bayesian perspective. 2018.

Laura Liu, Hyungsik Roger Moon, and Frank Schorfheide. Forecasting with a panel tobit model. Technical report, National Bureau of Economic Research, 2019.

Laura Liu, Hyungsik Roger Moon, and Frank Schorfheide. Forecasting with dynamic panel data models. *Econometrica*, 88(1):171–201, 2020.

Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

KW Newey and Daniel McFadden. Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pages 2112–2245, 1994.

Andriy Norets and Justinas Pelenis. Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory*, pages 606–646, 2014.

Kjetil Storesletten, Chris I Telmer, and Amir Yaron. Cyclical dynamics in idiosyncratic labor market risk. *Journal of political Economy*, 112(3):695–717, 2004.

Linda SL Tan, David J Nott, et al. Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28(2): 168–188, 2013.

Luke Tierney, Robert E Kass, and Joseph B Kadane. Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716, 1989.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

Michael Vogt and Oliver Linton. Multiscale clustering of nonparametric regression curves. *Journal of Econometrics*, 2020.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1 (1–2):1–305, 2008.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

John Winn and Christopher M Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694, 2005.

Appendix A. Details of EAMP

A.1. **Linear panel data models.** For each observation i, the linear mixed-effects panel data model can be written in matrix form as

$$\mathbf{y}_i = W_{1i}\alpha_i + W_{2i}\theta + \epsilon_i.$$

As noted earlier, we shall consider a normal prior family of the form $\pi(\alpha|\gamma) = N(\mu, \Sigma)$. The EAMP algorithm proceeds in the following steps:

Step E. Since the specified prior is normal and conjugate to the likelihood, the group-specific posterior will also be normally distributed. In particular,

$$q_{si}(\alpha) \leftarrow N(\alpha|\mu_{si}, \Sigma_{si}),$$

where

$$\begin{aligned}
\Sigma_{si} &\leftarrow \left(\Sigma_s^{-1} + W_{1i}^\intercal W_{1i}\right)^{-1}, \\
\mu_{si} &\leftarrow \Sigma_{si}\left(\Sigma_s^{-1}\mu_s + W_{1i}^\intercal(\mathbf{y}_i - W_{2i}\theta)\right).
\end{aligned}$$

Here (μ_s, Σ_s) denote the current values of the prior parameters for group s.

Step A. A closed form expression for $\int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha$ is given by

$$\begin{aligned}
\ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha \leftarrow{}& \frac{1}{2}\left(\ln|\Sigma_{si}| - \ln|\Sigma_s|\right) \\
&- \frac{1}{2}\left\{(y_i - W_{2i}\theta)^\intercal(y_i - W_{2i}\theta) + \mu_s^\intercal\Sigma_s^{-1}\mu_s - \mu_{si}^\intercal\Sigma_{si}^{-1}\mu_{si}\right\} + \text{const} \\
\equiv{}& I_{si} + \text{const},
\end{aligned}$$

where the constant does not depend on s. We would therefore assign observation i to the group s that maximizes I_{si}:

$$s(i) \leftarrow \arg\max_s I_{si}.$$

Step M. The update to θ is given by

$$\begin{aligned}
\theta &\leftarrow \left(\sum_i W_{2i}W_{2i}^\intercal\right)^{-1}\sum w_i(s)W_{2i}\left(\mathbf{y}_i - W_{1i}E_{q_{si}(\cdot)}[\alpha]\right) \\
&\equiv \left(\sum_i W_{2i}W_{2i}^\intercal\right)^{-1}\sum_i W_{2i}\left(\mathbf{y}_i - W_{1i}\sum_s w_i(s)\mu_{si}\right).
\end{aligned}$$

Step P. Suppose that we impose no restrictions on the prior means and variances, (μ, Σ). In this case, the natural statistics of the prior family are α_i and $\alpha_i\alpha_i^\intercal$. This

implies that the prior parameters may be updated as

$$\mu_s \leftarrow \frac{1}{n_s}\sum_i w_i(s)\mu_{si},$$
$$\Sigma_s \leftarrow \frac{1}{n_s}\sum_i w_i(s)\left(\Sigma_{si} + \mu_{si}\mu_{si}^\intercal\right) - \mu_s\mu_s^\intercal,$$

for each s, with $n_s = \sum_i w_i(s)$ denoting the number of observations in group s.

Alternatively, we may wish to restrict the dependence structure in the prior. As a first example, suppose that we can partition $\alpha_i \equiv (\alpha_{1i}^\intercal, \alpha_{2i}^\intercal)^\intercal$, where it is known that $\mathrm{Cov}[\alpha_{1i}, \alpha_{2i}] = 0$. Then the natural statistics are $\alpha_i, \mathrm{vec}(\alpha_{1i}\alpha_{1i}^\intercal)$ and $\mathrm{vec}(\alpha_{2i}\alpha_{2i}^\intercal)$. Thus Σ_s is a block diagonal matrix comprised of Σ_{1s}, Σ_{2s}, each of which can be updated as

$$\Sigma_{1s} \leftarrow \frac{1}{n_s}\sum_i w_i(s)\left(\Sigma_{1si} + \mu_{1si}\mu_{1si}^\intercal\right) - \mu_{1s}\mu_{1s}^\intercal,$$
$$\Sigma_{2s} \leftarrow \frac{1}{n_s}\sum_i w_i(s)\left(\Sigma_{2si} + \mu_{2si}\mu_{2si}^\intercal\right) - \mu_{2s}\mu_{2s}^\intercal,$$

where $(\Sigma_{1si}, \mu_{1si}, \mu_{1s})$ denote the partitions of $(\Sigma_{si}, \mu_{si}, \mu_s)$ corresponding to α_{1i} etc.

A stronger restriction, when α_{it} is univariate and time stationary, is to specify $\pi(\alpha|\gamma) = N(\mathbf{e}\mu, \sigma^2 I)$, where $\mathbf{e}$ denotes a T dimensional vector of ones. The natural statistics in this instance are $\sum_t \alpha_{it}$ and $\sum_t \alpha_{it}^2$. Let μ_{sit} denote the t-th component of the posterior mean μ_{si} from Step E. Similarly, let σ_{sit}^2 denote the t-th diagonal entry of Σ_{si}. Then the updates to the prior parameters are given by

$$\mu_s \leftarrow \frac{1}{n_s T}\sum_{i,t} w_i(s) E_{q_{si}(\cdot)}[\alpha_{it}] \equiv \frac{1}{n_s T}\sum_{i,t} w_i(s)\mu_{sit},$$

and

$$\sigma_s^2 \leftarrow \frac{1}{n_s T}\sum_{i,t} w_i(s) E_{q_{si}(\cdot)}[\alpha_{it}^2] - \mu_s^2$$
$$\equiv \frac{1}{n_s T}\sum_{i,t} w_i(s)\left(\sigma_{sit}^2 + \mu_{sit}^2\right) - \mu_s^2.$$

As yet a fourth possibility, we consider an AR(1) specification for α_i by setting $\pi(\alpha|\gamma) \equiv N\left(\mu, \Sigma(\sigma^2, \rho)\right)$, where the (t, t')-th element of $\Sigma(\sigma, \rho)$ is given by $\sigma^2\rho^{-|t-t'|}/(1-\rho^2)$. This prior is not a member of the exponential family. However, we can still obtain a tractable algorithm by expanding the objective function in the

minimization problem (see equation 3.7 in the main text) for $\gamma_s := (\mu_s, \sigma_s^2, \rho_s)$:

$$\begin{aligned} &\frac{1}{n_s}\sum_{i=1}^{n} w_i(s)\mathrm{KL}\left(q_{si}(\alpha) \,||\, \pi(\alpha|\gamma_s)\right) \qquad (A.1)\\ &= \frac{1}{2n_s}\sum_{i=1}^{n} w_i(s)\left\{\mathrm{Tr}\left(\Sigma_s^{-1}\Sigma_{si}\right) + (\mu_s-\mu_{si})^\intercal\Sigma_s^{-1}(\mu_s-\mu_{si}) + \ln|\Sigma_s|\right\} + \text{const.}, \end{aligned}$$

where $\Sigma_s := \Sigma(\sigma_s^2, \rho_s)$, and the constant comprises of terms that are independent of γ_s. Minimizing (A.1) over μ_s gives

$$\mu_s = \frac{1}{n_s}\sum_i w_i(s)\mu_{si}. \qquad (A.2)$$

Substituting the above expression into (A.1), we find that the updated values of (σ_s^2, ρ_s) can be obtained by minimizing

$$\begin{aligned} Q(\sigma_s^2, \rho_s) &:= \mathrm{Tr}\left(\Sigma_s^{-1}\tilde{\Sigma}_{si}\right) + \ln|\Sigma_s|, \text{ where}\\ \tilde{\Sigma}_{si} &:= \frac{1}{n_s}\sum_i w_i(s)\left(\Sigma_{si} + \mu_{si}\mu_{si}^\intercal\right) - \mu_s\mu_s^\intercal. \end{aligned}$$

Let $\tilde{\sigma}_{sit}^2$ denote the t-th diagonal entry of $\tilde{\Sigma}_{si}$, and $\tilde{\sigma}_{sitt'}$ its (t,t')-th entry for $t \neq t'$. We note the following two properties of Σ_s: First, $|\Sigma_s| := \det(\Sigma_s) = \sigma_s^{2T}/(1-\rho_s^2)$. Second, Σ_s^{-1} is of the form

$$\Sigma_s^{-1} = \frac{1}{\sigma_s^2}\begin{bmatrix} 1 & -\rho_s & & & \\ -\rho_s & 1+\rho_s^2 & -\rho_s & & \\ & -\rho_s & \ddots & \ddots & \\ & & \ddots & 1+\rho_s^2 & -\rho_s \\ & & & -\rho_s & 1 \end{bmatrix}.$$

Using these two properties, we can expand $Q(\sigma_s^2, \rho_s)$ as

$$Q(\sigma_s^2, \rho_s) = \ln(\sigma_s^{2T}) - \ln(1-\rho_s^2) + \frac{1+\rho_s^2}{\sigma_s^2}\sum_{t=1}^{T}\tilde{\sigma}_{sit}^2 - \frac{\rho_s^2}{\sigma_s^2}\left\{\tilde{\sigma}_{si1}^2 + \tilde{\sigma}_{siT}^2\right\} - \frac{2\rho_s}{\sigma_s^2}\sum_{t=1}^{T-1}\tilde{\sigma}_{sit(t+1)}.$$

Minimizing the above expression over (ρ_s, σ_s^2), we obtain

$$\rho_s\sum_{t=2}^{T-1}\tilde{\sigma}_{sit}^2 - \sum_{t=1}^{T-1}\tilde{\sigma}_{sit(t+1)} - \frac{\rho_s\sigma_s^2}{1-\rho_s^2} = 0, \qquad (A.3)$$

and

$$\sigma_s^2 = \frac{(1+\rho_s^2)}{T}\sum_{t=1}^{T}\tilde{\sigma}_{sit}^2 - \frac{\rho_s^2}{T}\left\{\tilde{\sigma}_{si1}^2 + \tilde{\sigma}_{siT}^2\right\} - \frac{2\rho_s}{T}\sum_{t=1}^{T-1}\tilde{\sigma}_{sit(t+1)}. \qquad (A.4)$$

Algorithm 1 EAMP algorithm for linear panel data models

Initialize $\theta, (\mu_s, \Sigma_s)_s$ to random values

Repeat:

For all i, s**:**

$\Sigma_{si} \leftarrow (\Sigma_s^{-1} + W_{1i}^\intercal W_{1i})^{-1}$

$\mu_{si} \leftarrow \Sigma_{si} \left(\Sigma_s^{-1}\mu_s + W_{1i}^\intercal(\mathbf{y}_i - W_{2i}\theta)\right)$

$I_{si} \leftarrow \frac{1}{2}\left(\ln|\Sigma_{si}| - \ln|\Sigma_s|\right) - \frac{1}{2}\left\{(y_i - W_{2i}\theta)^\intercal(y_i - W_{2i}\theta) + \mu_s^\intercal\Sigma_s^{-1}\mu_s - \mu_{si}^\intercal\Sigma_{si}^{-1}\mu_{si}\right\}$

$r_{si} \leftarrow (I_{si} == \max_s I_{si})$

End For

$\theta \leftarrow \left(\sum_i W_{2i}W_{2i}^\intercal\right)^{-1}\sum_i W_{2i}\left(\mathbf{y}_i - W_{1i}\sum_s r_{si}\mu_{si}\right)$

For all s**:**

$n_s \leftarrow \sum_i r_{si}$

$\mu_s \leftarrow \frac{1}{n_s}\sum_i r_{si}\mu_{si}$

$\Sigma_s \leftarrow \frac{1}{n_s}\sum_i r_{si}\left(\Sigma_{si} + \mu_{si}\mu_{si}^\intercal\right) - \mu_s\mu_s^\intercal$

End For

log_likelihood $\leftarrow \sum_{i,s} r_{si} I_{si}$

Until convergence criteria for 'log_likelihood' are satisfied

Equations (A.3) and (A.4) imply a cubic equation for ρ_s, which can be solved analytically. This can then be plugged into (A.4) to update σ_s^2. To summarize, the P step under an AR(1) specification of the prior consists of solving the equations (A.2)-(A.4).

In general, the EAMP algorithm is only guaranteed to find a local solution. Hence it is important to run the algorithm with multiple random initializations, and pick the solution that achieves the highest in-sample likelihood

$$\ln L := \sum_i \sum_s w_i(s) \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma_s)d\alpha \equiv \sum_i \sum_s w_i(s) I_{si}.$$

The pseudo-code for the resulting algorithm (with unrestricted covariance structure) is presented in Algorithm 1.

A.2. **Poisson panel data models with time varying fixed effects.** Consider the fixed effects Poisson panel data model

$$p(\mathbf{y}_i|\mathbf{x}_i, \alpha_i, \theta) \propto \prod_{t=1}^{T} (\alpha_{it} \exp(\theta^\intercal x_{it}))^{y_{it}} \exp(-\alpha_{it} \exp(\theta^\intercal x_{it})).$$

As noted earlier, the conjugate prior to this likelihood family is

$$\pi(\alpha|\gamma) \sim \prod_{t=1}^{T} \Gamma(k_t, \beta_t); \quad \gamma := (k_1 - 1, \ldots, k_T - 1, \beta_1, \ldots, \beta_T),$$

where $\Gamma(k, \beta)$ denotes the Gamma distribution with shape parameter k and inverse scale parameter β. The natural statistics of this prior distribution are $u(\alpha_i) = [\ln \alpha_{i1}, \ldots, \ln \alpha_{iT}, \alpha_{i1}, \ldots \alpha_{iT}]$. Define the group specific prior parameters as $(\mathbf{k}_s, \boldsymbol{\beta}_s) \equiv (k_{s1}, \ldots, k_{sT}, \beta_{s1}, \ldots, \beta_{sT})$. The EAMP algorithm then proceeds in the following steps:

Step E. Since the specified prior is conjugate to the likelihood, the group-specific posterior also has a Gamma distribution:

$$q_{si}(\alpha) \leftarrow \prod_{t=1}^{T} \Gamma(k_{sit}, \beta_{sit}) \equiv \prod_{t=1}^{T} q_{sit}(\alpha),$$

where for all s, i, t

$$k_{sit} \leftarrow y_{it} + k_{st} - 1,$$
$$\beta_{sit} \leftarrow \exp(\theta^\intercal x_{it}) + \beta_{st}.$$

Note from the above that $E_{q_{sit}(\alpha)}[\alpha] = k_{sit}/\beta_{sit}$, and $E_{q_{sit}(\alpha)}[\ln \alpha] = \psi(k_{sit}) - \ln(\beta_{sit})$, where $\psi(\cdot)$ denotes the Digamma function.

Step A. Denote the log integrated density by I_{si}. Then

$$I_{si} \leftarrow \sum_t \left(\ln \Gamma(k_{sit}) - k_{sit} \ln \beta_{sit} - \ln \Gamma(k_{st}) + k_{st} \ln \beta_{st} + y_{it}(\theta^\intercal x_{it}) - \ln y_{it}!\right),$$

where $\Gamma(\cdot)$ denotes the gamma function. We then assign observation i to group s as:

$$s(i) \leftarrow \arg\max_s I_{si}.$$

Step M. Plugging in the relevant functional forms in the general M step (3.6), we get

$$\begin{aligned}\theta &\leftarrow \arg\max_{\tilde{\theta}} \sum_{s,i,t} w_i(s) \left\{ y_{it}(\tilde{\theta}^\intercal x_{it}) - E_{q_{sit}(\alpha)}[\alpha] \exp(\tilde{\theta}^\intercal x_{it}) \right\} \\ &\equiv \arg\max_{\tilde{\theta}} \sum_{s,i,t} w_i(s) \left\{ y_{it}(\tilde{\theta}^\intercal x_{it}) - (k_{sit}/\beta_{sit}) \exp(\tilde{\theta}^\intercal x_{it}) \right\}.\end{aligned}$$

This is a convex optimization problem, equivalent to Poisson regression with known individual, time and group specific intercepts $\ln(k_{sit}/\beta_{sit})$. Thus one can compute θ through standard gradient descent methods.

Step P. Recall that the natural statistics of the prior family are α_i and $\ln \alpha_i$. This implies that the prior parameters need to solve

$$\begin{aligned}\frac{k_{st}}{\beta_{st}} &\leftarrow \frac{1}{n_s} \sum_i w_i(s) \frac{k_{sit}}{\beta_{sit}}, \\ \psi(k_{st}) - \ln(\beta_{st}) &\leftarrow \frac{1}{n_s} \sum_i w_i(s) \left(\psi(k_{sit}) - \ln(\beta_{sit}) \right).\end{aligned}$$

for each s, t. Unfortunately, the above system of equations does not have a closed form solution. However one can use standard root finding algorithms to solve for k_{st} via

$$\ln k_{st} - \psi(k_{st}) \leftarrow \ln \left(\frac{1}{n_s} \sum_i w_i(s) \frac{k_{sit}}{\beta_{sit}} \right) - \frac{1}{n_s} \sum_i w_i(s) \left(\psi(k_{sit}) - \ln(\beta_{sit}) \right).$$

There exists a unique solution to the above since the function $\ln x - \psi(x)$ is non-zero and monotonically decreasing on the positive real line.

A.3. **Logistic panel data models.** The random coefficients Logistic panel data model is given by

$$p(y_{it}|x_{it}, \alpha_{it}, \theta) = \sigma\left((2y_{it} - 1) \{ \alpha_{it}^\intercal w_{1it} + \theta^\intercal w_{2it} \} \right),$$

where $\sigma(x) = e^x/(1 + e^x)$ is the Logistic function. We shall assume that the data is stacked in such a way that the model is generated by

$$\begin{aligned}y_{it} &\sim \sigma(y_{it}^*); \text{ where} \\ \mathbf{y}_i^* &= W_{1i}\alpha_i + W_{2i}\theta.\end{aligned}$$

In this section, we show how one can employ the Jaakkola and Jordan [1997] device for estimation of logistic panel data models with normal priors.

Following Jaakkola and Jordan [1997], we employ the variational approximation

$$p(y_{it}|x_{it},\alpha_{it},\theta) \geq \sigma(\xi)\exp\left\{(y_{it}-1/2)X_{it}-\lambda(\xi)(X_{it}^2-\xi^2)-\xi/2\right\}$$
$$:= \tilde{p}(y_{it}|x_{it},\alpha_{it},\theta,\xi)$$

for all ξ, where $X_{it} := \alpha_{it}^{\intercal}w_{1it}+\theta^{\intercal}w_{2it}$ and $\lambda(\xi) := (1/2-\sigma(\xi))/2\xi$. Note that the inequality in the above expression becomes exact if and only if $\xi = -X_{it}$. Denote $\tilde{p}(\mathbf{y}_i|\mathbf{x}_i,\alpha_i,\theta,\boldsymbol{\xi}_{si}) := \prod_{t=1}^{T}\tilde{p}(y_{it}|x_{it},\alpha_{it},\theta,\xi_{sit})$, where $\boldsymbol{\xi}_{si} := (\xi_{si1},\ldots,\xi_{siT})$. For a given value of $\boldsymbol{\xi}_{si}$, for each (i,s), the variational inference approach replaces the conditional likelihood $p(\mathbf{y}_i|\mathbf{x}_i,\alpha_i,\theta)$ with its best estimate from $\tilde{p}(\mathbf{y}_i|\mathbf{x}_i,\alpha_i,\theta,\boldsymbol{\xi}_{si})$. The resulting maximization problem is therefore

$$\max_{\substack{\{w_i(s)\},\theta,\\ \{\gamma_s\},\{\boldsymbol{\xi}_{si}\}}}\sum_{i=1}^{n}\sum_{s=1}^{S}w_i(s)\ln\int\tilde{p}(\mathbf{y}_i|\mathbf{x}_i,\alpha_i,\theta,\boldsymbol{\xi}_{si})\pi(\alpha|\gamma_s)d\alpha$$
$$= \max_{\substack{\{q_{si}(\cdot)\},\{w_i(s)\},\\ \theta,\{\gamma_s\},\{\boldsymbol{\xi}_{si}\}}}\sum_{i=1}^{n}\sum_{s=1}^{S}w_i(s)\left\{E_{q_{si}(\cdot)}\left[\ln\tilde{p}(\mathbf{y}_i|\mathbf{x}_i,\alpha_i,\theta,\boldsymbol{\xi}_{si})\right]-\mathrm{KL}\left(q_{si}(\alpha)\;||\;\pi(\alpha|\gamma_s)\right)\right\}. \tag{A.5}$$

Note that (A.5) is only a lower bound to (3.3) since we do not allow $\boldsymbol{\xi}_{si}$ to be dependent on α_i. However, as noted by Jaakkola and Jordan [1997], the additional flexibility provided by $\boldsymbol{\xi}_{si}$ results in improved accuracy as compared to Laplace approximation. The advantage of the variational approach is that posterior distribution, $q_{si}(\cdot)$, is Gaussian as the form of $\tilde{p}(\cdot)$ ensures the term α_i only enters quadratically, and the prior is Gaussian. Hence the parameters of $q_{si}(\cdot)$ can be obtained after completing the square, in analogy with linear panel data models.

Let $\mathbf{v}_i := (y_{i1}-0.5,\ldots,y_{iT}-0.5)^{\intercal}$ and $\Lambda_{si} := \mathrm{diag}(\lambda(\xi_{si1}),\ldots,\lambda(\xi_{siT}))$, where the second term denotes a $T\times T$ diagonal matrix.

The variational algorithm proceeds by maximizing (A.5) repeatedly over $q_{si}(\cdot), w_i(s), \theta, \gamma_s, \boldsymbol{\xi}_{si}$, the details of which are given below:

Step E. As noted in the main text, the posterior distribution under the approximate likelihood $\tilde{p}(\mathbf{y}_i|\mathbf{x}_i,\alpha_i,\theta,\boldsymbol{\xi}_{si})$ is given by

$$\tilde{q}_{si}(\alpha) \propto \tilde{p}(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta,\boldsymbol{\xi}_{si})\pi(\alpha|\gamma_s).$$

Note that α_i enters quadratically in $\tilde{p}(\cdot)$. Since the prior is Gaussian, it therefore follows that the posterior distribution $\tilde{q}_{si}(\cdot)$ is also Gaussian, the parameters of which

are obtained as:

$$\tilde{q}_{si}(\alpha) \leftarrow N(\mu_{si}, \Sigma_{si}),$$

where

$$\begin{aligned}
\Sigma_{si} &\leftarrow \left(\Sigma_s^{-1} + 2W_{1i}^{\intercal}\Lambda_{si}W_{1i}\right)^{-1}, \\
\mu_{si} &\leftarrow \Sigma_{si}\left(\Sigma_s^{-1}\mu_s + W_{1i}^{\intercal}\left(\boldsymbol{v}_i - 2\Lambda_{si}W_{2i}^{\intercal}\theta\right)\right).
\end{aligned}$$

Step A. Denote the log integrated density by I_{si}. Then

$$\begin{aligned}
I_{si} \leftarrow \frac{1}{2}\left(\ln|\Sigma_{si}| - \ln|\Sigma_s|\right) - \frac{1}{2}\left(\mu_s^{\intercal}\Sigma_s^{-1}\mu_s - \mu_{si}^{\intercal}\Sigma_{si}^{-1}\mu_{si}\right) \\
+ \sum_t \left(\ln\sigma(\xi_{sit}) - \xi_{sit}/2 + \lambda(\xi_{sit})\xi_{sit}^2\right) + \text{const.}
\end{aligned}$$

We then assign observation i to the group s that maximizes I_{si}:

$$s(i) \leftarrow \arg\max_s I_{si}.$$

Step M. We can expand $\ln\tilde{p}(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta, \boldsymbol{\xi})$ for terms involving θ as

$$\ln\tilde{p}(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta, \boldsymbol{\xi}_{si}) = (y_{it} - 1/2)w_{2it}^{\intercal}\theta - \lambda(\xi_{sit})(w_{1it}^{\intercal}\alpha_{it} + w_{2it}^{\intercal}\theta)^2 + \text{const.}.$$

Hence,

$$\begin{aligned}
\theta &\leftarrow \arg\max_{\tilde{\theta}} \sum_{s,i,t} w_i(s) E_{\tilde{q}_{si}(\alpha)}\left[\ln\tilde{p}(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta, \boldsymbol{\xi}_{si})\right] \\
&\equiv \arg\max_{\tilde{\theta}} \sum_{s,i,t} w_i(s) E_{\tilde{q}_{si}(\alpha)}\left[(y_{it} - 1/2)w_{2it}^{\intercal}\theta - \lambda(\xi_{sit})(w_{1it}^{\intercal}\alpha_{it} + w_{2it}^{\intercal}\theta)^2\right].
\end{aligned}$$

Taking the first order condition with respect to θ and exploiting the form of $\tilde{q}_{si}(\alpha)$ obtained above, we get after some algebra that

$$\theta \leftarrow \left(\sum_{s,i} w_i(s) W_{2i}^{\intercal}\Lambda_{si}W_{2i}\right)^{-1} \sum_{s,i} w_i(s) W_{2i}^{\intercal}\left(\boldsymbol{v}_i - 2\Lambda_{si}W_{1i}^{\intercal}\mu_{si}\right).$$

Step P. Since the prior is Gaussian, the updates are the same as in linear panel data models

$$\begin{aligned}
\mu_s &\leftarrow \frac{1}{n_s}\sum_i w_i(s)\mu_{si}, \\
\Sigma_s &\leftarrow \frac{1}{n_s}\sum_i w_i(s)\left(\Sigma_{si} + \mu_{si}\mu_{si}^{\intercal}\right) - \mu_s\mu_s^{\intercal}.
\end{aligned}$$

Step M-inner. In this step we update the values of $\{\boldsymbol{\xi}_{si}\}$ for each i, s. As noted in the main text, this involves solving

$$\boldsymbol{\xi}_{si} \leftarrow \arg\max_{\boldsymbol{\xi}} E_{\tilde{q}_{si}(\cdot)} \left[\ln \tilde{p}(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta, \boldsymbol{\xi})\right]. \tag{A.6}$$

Now, we may expand $\ln \tilde{p}(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta, \boldsymbol{\xi})$ for terms involving $\boldsymbol{\xi}$ as

$$\ln \tilde{p}(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta, \boldsymbol{\xi}) = \sum_t \left\{\ln \sigma(\xi_t) - \xi_t/2 - \lambda(\xi_t)\left((w_{1it}^\intercal \alpha_{it} + w_{2it}^\intercal \theta)^2 - \xi_t^2\right)\right\} + \text{const.}$$

Plugging in the above in (A.6) and taking the first order condition with respect to $\boldsymbol{\xi}_{is}$, we get

$$E_{\tilde{q}_{si}(\alpha)}\left[\lambda'(\xi_{sit})\left((w_{1it}^\intercal \alpha_{it} + w_{2it}^\intercal \theta)^2 - \xi_{sit}^2\right)\right] = 0,$$

for each s, i, t. Making use of our knowledge of $\tilde{q}_{si}(\alpha)$, we thus obtain

$$\xi_{sit} \leftarrow \left\{w_{1it}^\intercal \left(\Sigma_{sit} + \mu_{sit}\mu_{sit}^\intercal\right) w_{1it} + 2\theta^\intercal w_{2it} w_{1it}^\intercal \mu_{sit} + (w_{2it}^\intercal \theta)^2\right\}^{1/2}$$

for each s, i, t, where $(\mu_{sit}, \Sigma_{sit}) = (E_{\tilde{q}_{si}(\alpha)}[\alpha_{it}], \text{Var}_{\tilde{q}_{si}(\alpha)}[\alpha_{it}])$. Note that the quantities $(\mu_{sit}, \Sigma_{sit})$ are just sub-compenents of (μ_{si}, Σ_{si}).

Appendix B. Details of VB-EAMP

B.1. **Derivation of the VB-EAMP algorithm.** In a standard mean-field VB setting where all the posterior quantities are fully factorized, the update rules are a form of message passing (Blei et al., 2017). This is also true in our setting for the updates of $q(\mu), \{q(\gamma_s)\}_s$. For instance, the value of $q(\mu)$ that optimizes

$$\mathcal{L}(q) := -\text{KL}\left(\tilde{q}(\alpha|w) \cdot \tilde{q}(w) \cdot \tilde{q}(\mu) \cdot \prod_s \tilde{q}(\gamma_s) \;||\; p(\mathbf{y}, \alpha, w, \mu, \gamma|\mathbf{x})\right)$$

from (4.1) - when all the other quantities are held fixed - is given by[8]

$$\ln q(\mu) = E_{q(\mathbf{z}_{-\mu})}\left[\ln p(\mathbf{y}, \mathbf{z}|\mathbf{x})\right] + \text{const.} \tag{B.1}$$

This form of the updates also holds for $\{q(\gamma_s)\}_s$, but the updates for $q(\alpha|w), q(w)$ requires new derivations specific to our setting.

Step A: Updating $q(w)$. Due to the form of the joint likelihood and the factorization of $q(\cdot)$, we have (recall that $H(\cdot)$ denotes the entropy)

$$\mathcal{L}(q) = \int q(\bar{w}) \left\{\int q(\alpha|\bar{w}) E_{q(\mu,\gamma)}\left[\ln p(\mathbf{y}, \alpha, \bar{w}, \mu, \gamma|\mathbf{x})\right] d\alpha + H(q(\alpha|\bar{w}))\right\} d\bar{w} + \text{const},$$

[8]For the derivation of the messages in the standard mean-field setting, see Winn and Bishop [2005] and Blei et al. [2017].

where the constant does not depend on $q(\alpha|w)$, and we use $\bar{w}$ to emphasize the fact that it represents a point in the support of w (and not the random variable w). Using the above expression, we can optimize the value of $q(\alpha|\bar{w})$ for each given $\bar{w}$. This is given by

$$\begin{aligned}\ln q(\alpha|\bar{w}) &= E_{q(\mu,\gamma)}\left[\ln p(\mathbf{y},\alpha,\bar{w},\mu,\gamma|\mathbf{x})\right] + \text{const} \\ &= \sum_i \ln p(\mathbf{y}_i|\mathbf{x}_i,\alpha_i) + \sum_i \sum_s \bar{w}_{si} E_{q(\gamma_s)}\left[\ln \pi(\alpha_i|\gamma_s)\right] + \text{const} \\ &= \sum_i \sum_s \bar{w}_{si}\left\{\ln p(\mathbf{y}_i|\mathbf{x}_i,\alpha_i) + E_{q(\gamma_s)}\left[\ln \pi(\alpha_i|\gamma_s)\right]\right\} + \text{const},\end{aligned}$$

where the last equality follows from $\sum_s \bar{w}_{si} = 1$. Hence we find

$$q(\alpha|\bar{w}) \leftarrow \prod_i \prod_s q_{si}(\alpha_i)^{\bar{w}_{si}};\ \text{where}$$

$$\ln q_{si}(\alpha) := \ln p(\mathbf{y}_i|\mathbf{x}_i,\alpha) + E_{q(\gamma_s)}\left[\ln \pi(\alpha|\gamma_s)\right] + \text{const}.$$

Completing the density function gives the update rule

$$q_{si}(\alpha) = \frac{p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta)\bar{\pi}_s(\alpha)}{\int p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta)\bar{\pi}_s(\alpha)d\alpha};\quad \bar{\pi}_s(\alpha) \propto \exp E_{q(\gamma_s)}\left[\ln \pi(\alpha|\gamma_s)\right].$$

Since $\pi(\alpha|\gamma)$ has an exponential distribution, we can obtain an explicit expression for $\bar{\pi}_s(\alpha)$:

$$\bar{\pi}_s(\alpha) = h(\alpha)g(\bar{\gamma}_s)\exp\left(\bar{\gamma}_s^{\intercal}u(\alpha)\right);\ \bar{\gamma}_s := E_{q(\gamma_s)}\left[\gamma_s\right].$$

Step A: Updating $q(w)$. We can write $\mathcal{L}(q)$ in a form that highlights the role of $q(w)$ as

$$\mathcal{L}(q) = \int q(w)\left\{E_{q(\alpha,\mu,\gamma|w)}\left[\ln p(\mathbf{y},\mathbf{z}|\mathbf{x})\right] + H\left(q(\alpha|w)\right)\right\}dw + H(q(w)) + \text{const}.$$

Hence the optimized value of $q(w)$ is

$$\ln q(w) = E_{q(\alpha,\mu,\gamma|w)}\left[\ln p(\mathbf{y},\mathbf{z}|\mathbf{x})\right] + H\left(q(\alpha|w)\right) + \text{const}.$$

Now, in view of (4.2), we have $H\left(q(\alpha|w)\right) = \sum_{i,s} w_{si}H(q_{si}(\alpha_i))$. Furthermore, using again the form of $q(\alpha|w)$ from (4.2) and the fact $\sum_s w_{si} = 1$, some straightforward algebra gives

$$E_{q(\alpha,\mu,\gamma|w)}\left[\ln p(\mathbf{y},\mathbf{z}|\mathbf{x})\right] = \sum_{i,s} w_{si}\left\{E_{q(\mu_s)}\left[\ln \mu_s\right] + E_{q_{si}(\alpha_i)}\left[\ln p(\mathbf{y}_i|\mathbf{x}_i,\alpha_i)\right] + E_{q_{si}(\alpha_i)}\left[\ln \bar{\pi}_s(\alpha_i)\right]\right\},$$

where $\bar{\pi}_s(\alpha)$ has been defined earlier in (4.3).

Combining the above, it can be seen that the update for $q(w)$ is given by

$$q(w) = \prod_i \prod_s r_{si}^{w_{si}},$$

where

$$\begin{aligned}\ln r_{si} &= E_{q(\mu_s)}\left[\ln \mu_s\right] + E_{q_{si}(\alpha_i)}\left[\ln p(\mathbf{y}_i|\mathbf{x}_i, \alpha_i)\right] + E_{q_{si}(\alpha_i)}\left[\ln \bar{\pi}_s(\alpha_i)\right] + H\left(q_{si}(\alpha_i)\right) + \text{const} \\ &= E_{q(\mu_s)}\left[\ln \mu_s\right] + \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha)\bar{\pi}_s(\alpha)d\alpha + \text{const}\end{aligned}$$

and the last step follows from (4.2) and the variational formula (3.1) since $E_{q_{si}(\alpha)}\left[\ln \bar{\pi}_s(\alpha)\right] + H\left(q_{si}(\alpha)\right) = -\mathrm{KL}(q_{si}(\alpha) \,||\, \bar{\pi}_s(\alpha))$. As we require $\sum_s r_{si} = 1$, it follows

$$r_{si} = \frac{\exp\left\{E_{q(\mu_s)}\left[\ln \mu_s\right] + \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha)\bar{\pi}_s(\alpha)d\alpha\right\}}{\sum_s \exp\left\{E_{q(\mu_s)}\left[\ln \mu_s\right] + \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha)\bar{\pi}_s(\alpha)d\alpha\right\}}.$$

Step P: Updating $q(\mu)$ & $q(\gamma)$. As described in equation (B.1), the update to $q(\mu)$ is given by

$$\begin{aligned}\ln q(\mu) &= E_{q(\alpha,w,\gamma)}\left[\ln p(\mathbf{y}, \mathbf{z}|\mathbf{x})\right] + \text{const} \\ &= (\beta_0 - 1)\sum_s \ln \mu_s + \sum_i \sum_s r_{si} \ln \mu_s + \text{const},\end{aligned}$$

where we have made use of the expression (4.5) for $q(w)$ in the second equality. From the form of the density, we recognize

$$q(\mu) = \mathrm{Dir}(\mu|\boldsymbol{\beta}), \tag{B.2}$$

where $\boldsymbol{\beta}$ is a $S \times 1$ vector with $\beta_s = \beta_0 + \sum_i r_{si}$.

The update to $q(\gamma_s)$ is given by the message $\ln q(\gamma_s) = E_{q(\alpha,w,\mu)}\left[\ln p(\mathbf{y}, \mathbf{z}|\mathbf{x})\right] + \text{const}$. Evaluating this expectation, and making use of (4.5), some straightforward algebra gives

$$\ln q(\gamma_s) = \left(\nu_0 + \sum_i r_{si}\right) \ln g(\gamma_s) + \gamma_s^\intercal \left(\chi_0 + \sum_i r_{si} E_{q_{si}(\alpha)}\left[u(\alpha_i)\right]\right) + \text{const}.$$

We recognize that this has the form

$$q(\gamma_s) = f(\chi_s, \nu_s) g(\gamma_s)^{\nu_s} \exp(\nu_s \gamma_s^\intercal \chi_s), \tag{B.3}$$

where

$$\nu_s = \nu_0 + \sum_i r_{si}, \text{ and } \nu_s \chi_s = \chi_0 + \sum_i r_{si} E_{q_{si}(\alpha)}\left[u(\alpha_i)\right].$$

B.2. **Dirichlet Process priors and infinite number of groups.** We can let the number of groups be infinity and avoid the need to select S. This is achieved by specifying a Dirichlet Process (DP) mixture model for the prior parameter α. Indeed, DP mixtures have a stick-breaking representation that is closely related to our probabilistic model with fixed S. Under this representation, instead of a multinomial prior for w (as was the case for fixed S), we now have

$$p(w|\mathbf{v}) = \prod_i \prod_s \left\{ v_s \prod_{j=1}^{s} (1 - v_j) \right\}^{w_{si}},$$

where $\mathbf{v} := (v_1, v_2, \ldots, v_\infty)$ and the index s now ranges over $1, 2, \ldots, \infty$. The hyper-parameters $\mathbf{v}$ are modeled as independent draws from a beta distribution, i.e.,

$$p(\mathbf{v}|a_0) \equiv \prod_s p(v_s|a_0) = \prod_s \text{Beta}(1, a_0).$$

The rest of the probability model is the same as before.

We consider a mean-field approximation which factorizes between the latent variables and hyper-parameters $q(\alpha, \mathbf{w}, \mathbf{v}, \gamma) = q(\alpha, \mathbf{w})q(\mathbf{v}, \gamma)$. This induces a further factorization for $q(\mathbf{v}, \gamma)$, and the final factorization is given by

$$q(\alpha, \mathbf{w}, \mathbf{v}, \gamma) = q(\alpha|w) \cdot q(w) \cdot \prod_s q(v_s) \cdot \prod_s q(\gamma_s).$$

In practice, we truncate the support of s at a sufficiently large number $\bar{S}$.

We obtain the modified VB-EAMP algorithm by optimizing over each of the above factors. Here, we describe the result, skipping the derivation which is similar to before: The updates for $q(\alpha|w)$ and $q(\gamma_s)$ remain unchanged. For the update to $q(w)$, we have $q(w) = \prod_i \prod_s r_{si}^{w_{si}}$ as before, but the value of r_{si} is now

$$r_{si} = \frac{\exp\left\{ E_{q(v_s)}\left[\ln v_s\right] + \sum_{j=1}^{s-1} E_{q(v_j)}\left[\ln(1 - v_j)\right] + \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha)\bar{\pi}_s(\alpha)d\alpha \right\}}{\sum_s \exp\left\{ E_{q(v_s)}\left[\ln v_s\right] + \sum_{j=1}^{s-1} E_{q(v_s)}\left[\ln(1 - v_s)\right] + \ln \int p(\mathbf{y}_i|\mathbf{x}_i, \alpha)\bar{\pi}_s(\alpha)d\alpha \right\}}. \tag{B.4}$$

In addition, the update to $q(v_s); s = 1, 2, \ldots, \bar{S}$ is given by

$$\begin{aligned} q(v_s) &= \text{Beta}(b_s, a_s), \text{ where} \\ b_s &= 1 + \sum_i r_{si}, \text{ and} \\ a_s &= a_0 + \sum_i \sum_{j=s+1}^{\bar{S}} r_{ji}. \end{aligned} \tag{B.5}$$

Based on the above, we can obtain closed form expressions for $E_{q(v_s)}[\ln v_s]$, $E_{q(v_s)}[\ln(1-v_s)]$ used in (B.4) as $E_{q(v_s)}[\ln v_s] = \psi(a_s) - \psi(a_s + b_s)$ and $E_{q(v_s)}[\ln(1-v_s)] = \psi(b_s) - \psi(a_s + b_s)$, where $\psi(\cdot)$ is the Digamma function.

B.3. **Linear panel data models.** We specify a Normal-Inverse-Wishart (NIW) distribution as the hyper-prior for $\gamma_s \equiv (\mu_s, \Sigma_s)$:

$$p(\mu_s, \Sigma_s) \equiv \text{NIW}(\mathbf{m}_0, \lambda_0, \mathbf{\Psi}_0, \nu_0).$$

The terms $(\mathbf{m}_0, \lambda_0, \mathbf{\Psi}_0, \nu_0)$ are hyper-parameters, and will be updated in the course of the algorithm. Many of the steps of VB-EAMP algorithm are the same as in Appendix A, so we employ much of the same terminology. Also, as in Section 4, we shall suppose for simplicity that there are no common parameters θ.

Step E. Recall that in calculating the group-specific quantity, we make use of quantity $\pi_s(\alpha|\bar{\gamma}_s)$, where $\bar{\gamma}_s = E_{q(\gamma_s)}[\gamma_s]$. We shall show below that $\pi_s(\alpha|\bar{\gamma}_s) \equiv N(\mathbf{m}_s, \nu_s^{-1}\mathbf{\Psi}_s)$, where the values of $(\mathbf{m}_s, \lambda_s, \mathbf{\Psi}_s, \nu_s)$ are first initialized to random values and then updated in Step P. Since $\pi_s(\alpha|\bar{\gamma}_s)$ is normal, the group-specific posterior will also be normally distributed and given by

$$q_{si}(\alpha) \leftarrow N(\alpha|\mu_{si}, \Sigma_{si}),$$

where

$$\begin{aligned}
\Sigma_{si} &\leftarrow \left(\nu_s \mathbf{\Psi}_s^{-1} + W_{1i}^\intercal W_{1i}\right)^{-1}, \\
\mu_{si} &\leftarrow \Sigma_{si}\left(\nu_s \mathbf{\Psi}_s^{-1}\mu_s + W_{1i}^\intercal(\mathbf{y}_i - W_{2i}\theta)\right).
\end{aligned}$$

Step A. Recall the definition of I_{si} in Appendix A. The responsibilities r_{si} are updated as

$$r_{si} \leftarrow \frac{\exp\left(\psi(\beta_s) + I_{si}\right)}{\sum_s \exp\left(\psi(\beta_s) + I_{si}\right)},$$

for each s, i, where $\psi(\cdot)$ denotes the Digamma function.

Step M. The hyper-parameters $\boldsymbol{\beta}$ for w are updated as

$$\beta_s \leftarrow \beta_0 + \sum_i r_{si}.$$

Step P. The updates to the hyper-parameters $(\mathbf{m}_s, \lambda_s, \mathbf{\Psi}_s, \nu_s)$ are given by

$$\begin{aligned}
\lambda_s &\leftarrow \lambda_0 + \sum_i r_{si}, \\
\mathbf{m}_s &\leftarrow \frac{\lambda_0 \mathbf{m}_0 + \sum_i r_{si}\mu_{si}}{\lambda_s}, \\
\nu_s &\leftarrow \nu_0 + \sum_i r_{si}, \\
\mathbf{\Psi}_s &\leftarrow \mathbf{\Psi}_0 + \sum_i r_{si}\left(\Sigma_{si} + \mu_{si}\mu_{si}^\intercal\right) + \lambda_0 \mathbf{m}_0 \mathbf{m}_0^\intercal - \lambda_s \mathbf{m}_s \mathbf{m}_s^\intercal,
\end{aligned}$$

for each s. Based on the above, the posterior distribution for (μ_s, Σ_s) is

$$q(\mu_s, \Sigma_s) \equiv \text{NIW}(\mathbf{m}_s, \lambda_s, \mathbf{\Psi}_s, \nu_s).$$

Recall that the natural parameters for the multivariate normal are $\gamma_s := (\Sigma_s^{-1}\mu_s, \text{vec}(\Sigma_s^{-1}))$. Now from the properties of the NIW distribution, we have that

$$E_{q(\mu_s,\Sigma_s)}\left[\Sigma_s^{-1}\mu_s\right] = \nu_s \mathbf{\Psi}_s^{-1}\mathbf{m}_s, \text{ and } E_{q(\mu_s,\Sigma_s)}\left[\Sigma_s^{-1}\right] = \nu_s \mathbf{\Psi}_s^{-1}.$$

Hence it follows $\bar{\gamma}_s = (\nu_s \mathbf{\Psi}_s^{-1}\mathbf{m}_s, \nu_s \mathbf{\Psi}_s^{-1})$. It is then straightforward to observe that $\pi(\alpha|\bar{\gamma}_s) \equiv N(\mathbf{m}_s, \nu_s^{-1}\mathbf{\Psi}_s)$ as noted in Step E above.

Appendix C. Simulation

We provide a small simulation study focusing on the setting with continuous unobserved heterogeneity. The outcomes are given by

$$y_{it} = \alpha_{0i} + \alpha_{1i}x_{1it} + \theta_0 x_{2it} + \epsilon_{it},$$

where $\boldsymbol{\alpha}_i := (\alpha_{0i}, \alpha_{1i})$ are individual specific unobserved heterogeneity terms, $\theta_0 = 1$ and $\epsilon_{it} \sim$ i.i.d $N(0,1)$. We jointly model $(\boldsymbol{\alpha}_i, \boldsymbol{x}_i)$ as follows: First we draw individual specific values of unobserved heterogeneity as

$$\begin{pmatrix} \alpha_{0i} \\ \alpha_{1i} \\ \mu_{1i} \\ \mu_{2i} \end{pmatrix} \sim N\left(\begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0.9 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.9 & 0 & 0 & 1 \end{bmatrix}\right).$$

Conditional on (μ_{1i}, μ_{2i}), we obtain $(\mathbf{x}_{1i}, \mathbf{x}_{2i})$ for each individual as

$$x_{1it} \sim \text{i.i.d } N(\mu_{1i}, 1);\ x_{2it} \sim \text{i.i.d } N\left(\mu_{2i}, 1\right), \text{for } t = 1, \ldots, T.$$

Thus the model incorporates correlation between the individual specific intercept and $\mathbf{x}_{2i}$ corresponding to the common parameter. In employing latent variables

Algorithm 2 VB-EAMP algorithm for linear panel data models

Initialize $(\mathbf{m}_s, \lambda_s, \mathbf{\Psi}_s, \nu_s)_s$

Repeat:

For all i, s**:**

$\Sigma_{si} \leftarrow (\nu_s \mathbf{\Psi}_s^{-1} + W_{1i}^\intercal W_{1i})^{-1}$

$\mu_{si} \leftarrow \Sigma_{si} (\nu_s \mathbf{\Psi}_s^{-1} \mu_s + W_{1i}^\intercal (\mathbf{y}_i - W_{2i}\theta))$

$I_{si} \leftarrow \frac{1}{2} (\ln |\Sigma_{si}| - \ln |\Sigma_s|) - \frac{1}{2} \left\{ (y_i - W_{2i}\theta)^\intercal (y_i - W_{2i}\theta) + \mu_s^\intercal \Sigma_s^{-1} \mu_s - \mu_{si}^\intercal \Sigma_{si}^{-1} \mu_{si} \right\}$

$r_{si} \leftarrow \exp(\psi(\beta_s) + I_{si}) \,/\, \sum_s \exp(\psi(\beta_s) + I_{si})$

End For

For all s**:**

$\beta_s \leftarrow \beta_0 + \sum_i r_{si}$

$\mathbf{m}_s \leftarrow (\lambda_0 \mathbf{m}_0 + \sum_i r_{si} \mu_{si}) / \lambda_s$

$\nu_s \leftarrow \nu_0 + \sum_i r_{si}$

$\mathbf{\Psi}_s \leftarrow \mathbf{\Psi}_0 + \sum_i r_{si} (\Sigma_{si} + \mu_{si} \mu_{si}^\intercal) + \lambda_0 \mathbf{m}_0 \mathbf{m}_0^\intercal - \lambda_s \mathbf{m}_s \mathbf{m}_s^\intercal$

End For

Until convergence criteria for $(\mathbf{m}_s, \mathbf{\Psi}_s)_s$ **are satisfied**

for the construction of the covariates, the DGP is similar to model proposed by Bonhomme et al. [2017] for GFE.

Table 2 shows the results from employing the GRE approach, as well as comparisons with GFE, Random Effects (RE) and Fixed Effects (FE), the last of which are equivalent to GRE and GFE with 1 group. We see that for the common parameter, θ_0, the GRE approach dominates GFE in terms of MSE, for any group size. As suggested by the theory, this difference is driven almost entirely by the much lower bias of GRE. The same is true (albeit to a lower extent) for the marginal effect $E[\alpha_{0i}]$ as well. On the other hand, both GRE and GFE work equally well for estimation of $E[\alpha_{1i}]$. This is not altogether surprising, since the model introduced a high amount of correlation between $\mathbf{x}_{2i}$ (which corresponds to the common parameter) and α_{0i}, while α_{1i} is independent of all the other covariates, and therefore functions as a pure random effect. For this reason, the estimation of $E[\alpha_{1i}]$ works best under standard Random Effects, and adding more groups only serves to degrade the MSE.

TABLE 2. Simulation results

		$\theta_0 = 1$		$E[\alpha_{0i}] = 2$	$E[\alpha_{1i}] = 3$
Estimator	# of groups	Bias	RMSE	RMSE	RMSE
GRE	1	0.1573	0.169	0.190	0.054
	2	0.1067	0.126	0.179	0.066
	12	0.0711	0.104	0.169	0.075
	25	0.0421	0.093	0.180	0.084
GFE	1	0.4566	0.469	0.502	0.112
	2	0.2638	0.275	0.287	0.073
	12	0.2112	0.225	0.242	0.076
	25	0.1858	0.200	0.218	0.084

Notes: The table reports results for 1000 simulations. Both the EAMP and GFE algorithms were initialized with 10 random initial values, but changing this did not have any discernible effect. The sample size is $n = 100$ and number of time periods is $T = 5$.

TABLE 3. Simulation results by sample size

		$\theta_0 = 1$		$E[\alpha_{0i}] = 2$	$E[\alpha_{1i}] = 3$
Estimator	Time periods	Bias	RMSE	RMSE	RMSE
GRE	5	0.0711	0.104	0.169	0.075
	10	0.0535	0.071	0.103	0.042
	15	0.0332	0.046	0.072	0.032
GFE	5	0.2112	0.225	0.242	0.276
	10	0.2112	0.219	0.229	0.057
	15	0.2100	0.217	0.225	0.047

Notes: The table reports results for 1000 simulations. The sample size is $n = 100$ and number of groups is $S = 12$.

The number of time periods of $T = 5$ is small. Table 3 illustrates that increasing the number of time periods leads to a substantial reduction in MSE for GRE, but only leads to a marginal improvement for GFE.

APPENDIX D. THEORETICAL PROPERTIES UNDER DISCRETE UNOBSERVED HETEROGENEITY

Let $\boldsymbol{\gamma}_0 := (\gamma_{10}, \ldots, \gamma_{S0})$ denote the true values of the prior parameters corresponding to each of the discrete groups. Denote by s_{i0} the actual group to which

observation i belongs to, and by γ_{i0} the corresponding value of unobserved heterogeneity. Define $\omega := (\gamma, \theta)$ and $\omega_{i0} := (\gamma_{i0}, \theta_0)$.

We use the notation $p(\cdot|\omega_i)$ as a shorthand for the conditional distribution, $p(\mathbf{y}_i, \mathbf{x}_i|\omega_{i0})$, of the data given ω_{i0} (i.e., given the group index s_{i0} for observation i). The corresponding expectation is denoted by $E_{p(\cdot|\omega_i)}[\cdot]$. We also denote the unconditional expectation by $E[\cdot]$; compared to $E_{p(\cdot|\omega_i)}[\cdot]$, this additionally takes the expectation with respect to ω_{i0}. The limiting covariance matrix is given by

$$V = E\left[\partial_{\theta,\theta^\intercal} l_i(\omega_{i0})\right]^{-1} E\left[\partial_\theta l_i(\omega_{i0}) \partial_\theta l_i(\omega_{i0})^\intercal\right] E\left[\partial_{\theta,\theta^\intercal} l_i(\omega_{i0})\right]^{-1}.$$

As noted in the main text, the properties of GRE in this setting are closely related to that of GFE. Thus, Theorem 1 follows by similar argument as in Cheng et al. [2019]. The relevant regularity conditions, adapted to our setting, are given below, and we refer to Cheng et al. [2019] for a discussion of these assumptions:

Assumption C. *(i) The random variables $\{\mathbf{y}_i, \mathbf{x}_i, \xi_i\}$ are iid in i. For each i, $\{y_{it}, x_{it}, \xi_{it}\}_{t=1}^T$ are stationary strong mixing with mixing coefficients $a(\tau)$ such that $a(\tau) \leq c_\alpha r^\tau$ for some $c_\alpha < \infty$ and $\tau \in (0,1)$.*

(ii) The domain of ω_{i0} is a compact set $\Gamma \times \Theta$.

(iii) For any $\eta > 0$,

$$\min_{i=1,\ldots,N} \inf_{\|\omega_i - \omega_{i0}\| > \eta} E_{p(\cdot|\omega_{i0})}[l_i(\omega_i)] > \varepsilon > 0.$$

(iv) $\sup_i E_{p(\cdot|\omega_{i0})}\left[\sup_{\omega_i \in \Gamma\times\Theta} \|l_i(\omega_i)\|^q\right] < \infty$ and $\sup_i E_{p(\cdot|\omega_{i0})}\left[\sup_{\omega_i \in \Gamma\times\Theta} \|\partial_\omega l_i(\omega_i)\|^q\right] < \infty$ for some $q \geq 6$.

(v) There exists $c > 0$ such that $\|\gamma_{is} - \gamma_{is'}\| > \eta$ for all $s \neq s'$. Furthermore, $P(s_{i0} = s) = \pi_s > 0$ for all s.

(vi) The terms $E[\partial_\omega l_i(\omega_{i0})]$ and $E[\partial_{\omega,\omega^\intercal} l_i(\omega_{i0})]$ are both full rank.

(vii) $n, T \to \infty$ and $n = O(T^{q/4 - 1/2})$, where $q \geq 6$ is the same quantity as in part (iv) above.

Appendix E. Theoretical properties under continuous unobserved heterogeneity

E.1. **Assumptions for Lemma 1.** Denote $\omega := (\gamma, \theta)$ and $\omega_{i0} := (\gamma(\xi_i), \theta_0)$. We impose the following assumptions for Lemma 1:

Assumption D1. *(i) The random variables $\{\mathbf{y}_i, \mathbf{x}_i, \xi_i\}$ are iid in i.*

(ii) For each $\epsilon > 0$, there exists δ independent of ξ_i such that

$$E_{p(\cdot|\xi_i,\theta_0)}\left[l_i(\omega_{i0})\right] > \sup_{\|\omega-\omega_{i0}\|>\epsilon} E_{p(\cdot|\xi_i,\theta_0)}\left[l_i(\omega)\right] + \delta.$$

(iii) $l_i(\gamma,\theta)$ is thrice continuously differentiable in both its arguments. For $C < \infty$, and all $a, b \in \{0,1,2,3\}$ such that $a+b \leq 3$, the derivatives satisfy

$$\sup_{\xi_i \in \Xi} E_{p(\cdot|\xi_i,\theta_0)}\left[\sup_{\|\gamma\|<M, \theta\in\Theta} \left\|\partial_\gamma^a \partial_\theta^b l_i(\gamma(\xi_i),\theta)\right\|^2\right] \leq C.$$

(iv) There exists $\underline{\lambda} > 0$ independent of ξ_i such that

$$\lambda_{min}\left(-E_{p(\cdot|\xi_i,\theta_0)}\left[\partial^2_{(\omega,\omega^\intercal)} l_i(\omega_{i0})\right]\right) \geq \underline{\lambda}.$$

(v) $\max_i \sup_{\gamma,\theta}\left\{l_i(\gamma,\theta) - E_{p(\cdot|\xi_i,\theta_0)}\left[l_i(\gamma,\theta)\right]\right\} = o_{\mathbb{P}}(1)$, and similarly for the first two derivatives of $l_i(\gamma,\theta)$.

(vi) $\sup_i E_{p(\cdot|\xi_i,\theta_0)}\left[\|\partial_\omega l_i(\omega_{i0})\|^2\right] = O(T^{-1})$.

Assumptions C2(ii)-(v) are essentially the same as that employed in Bonhomme et al. [2017], and we refer to that paper for a discussion. For Assumption C2(vi), note that Lemma E.1 below implies $E_{p(\cdot|\xi_i,\theta_0)}\left[\partial_\omega l_i(\omega_{i0})\right] = 0$. Hence, this assumption is really a statement on the time series dependence of $\{y_{it}, x_{it}\}_{t=1}^T$ conditional on ξ_i. In particular, note that the assumption is clearly satisfied if $\{y_{it}, x_{it}\}_{t=1}^T$ is iid across time. More generally, it allows for weak time series dependence, such as strong mixing with mixing coefficient $\gamma(t) \lesssim \exp(-at^b);\ a, b > 0$.

E.2. **Proof of Lemma 1.** We start with a useful lemma.

Lemma E.1. *Suppose that Assumption 1 in the main text holds and Assumption D1(iii) hold. Then,*

$$E_{p(\cdot|\xi_i,\theta_0)}\left[\partial_\gamma l_i(\gamma(\xi_i),\theta_0)\right] = 0 \text{ and } E_{p(\cdot|\xi_i,\theta_0)}\left[\partial_\theta l_i(\gamma(\xi_i),\theta_0)\right] = 0 \ \forall\ \xi_i.$$

Proof. Recall the definition of $l_i(\gamma,\theta) := T^{-1}\ln\int p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta)\pi(\alpha|\gamma)d\alpha$, and also that $p(\cdot|\xi_i,\theta_0) := p(\mathbf{y}_i|\mathbf{x}_i,\xi_i,\theta)p(\mathbf{x}_i|\xi_i)$. Now, $\exp(Tl_i(\gamma,\theta)\cdot p(\mathbf{x}_i|\xi_i) \equiv \int p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta)\pi(\alpha|\gamma)d\alpha \cdot p(\mathbf{x}_i|\xi_i)$ is a valid probability distribution over the support of $(\mathbf{y}_i,\mathbf{x}_i)$. Then, by the properties of KL divergence,

$$\begin{aligned} E_{p(\cdot|\xi_i,\theta_0)}\left[\ln p(\cdot|\xi_i,\theta_0)\right] &\geq E_{p(\cdot|\xi_i,\theta_0)}\left[\ln\{\exp(Tl_i(\gamma,\theta)\cdot p(\mathbf{x}_i|\xi_i)\}\right] \\ &= TE_{p(\cdot|\xi_i,\theta_0)}\left[l_i(\gamma,\theta)\right] + E_{p(\cdot|\xi_i,\theta_0)}\left[\ln p(\mathbf{x}_i|\xi_i)\right], \end{aligned}$$

for all γ. Now, under Assumption 1, $l_i(\gamma(\xi_i), \theta_0) := T^{-1} \ln p(\mathbf{y}_i|\mathbf{x}_i, \xi_i, \theta)$, and, as a consequence,

$$\begin{aligned} E_{p(\cdot|\xi_i,\theta_0)} \left[\ln p(\cdot|\xi_i, \theta_0)\right] &\equiv E_{p(\cdot|\xi_i,\theta_0)} \left[\ln\{p(\mathbf{y}_i|\mathbf{x}_i, \xi_i, \theta_0)p(\mathbf{x}_i|\xi_i)\}\right] \\ &= T E_{p(\cdot|\xi_i,\theta_0)} \left[l_i(\gamma(\xi_i), \theta_0)\right] + E_{p(\cdot|\xi_i,\theta_0)} \left[\ln p(\mathbf{x}_i|\xi_i)\right]. \end{aligned}$$

Combining the above, we find

$$E_{p(\cdot|\xi_i,\theta_0)} \left[l_i(\gamma(\xi_i), \theta_0)\right] \geq E_{p(\cdot|\xi_i,\theta_0)} \left[l_i(\gamma, \theta)\right] \ \forall \ (\gamma, \theta),$$

which implies, $E_{p(\cdot|\xi_i,\theta_0)} \left[\partial_\gamma l_i(\gamma(\xi_i), \theta_0)\right] = 0$ as long as the derivative exists. □

We restate Lemma 1 here for convenience:

Lemma. *Suppose Assumption 1 in the main text and Assumption D1 hold. Then*

$$\left\|\hat{\theta} - \theta_0\right\|^2 + \frac{1}{n}\sum_i \left\|\hat{\gamma}_{i,S}(\hat{\theta}) - \gamma(\xi_i)\right\|^2 = O_{\mathbb{P}}\left(\frac{1}{T} + S^{-2/d} + \frac{S^{-1/d}}{\sqrt{nT}}\right).$$

Proof. Let $\omega := (\gamma, \theta)$, and $\omega_{i0} := (\gamma(\xi_i), \theta_0)$. Also, set $\hat{\omega}_{i,S} := (\hat{\gamma}_{i,S}(\hat{\theta}), \hat{\theta})$.

Denote $d := \dim(\xi)$, and let $\bar{\Xi}_S \equiv \{\bar{\xi}_1, \ldots, \bar{\xi}_S\} \in \Xi$ be an $S^{-1/d}$ cover for Ξ i.e.,

$$\arg\min_{\bar{\xi}_s \in \bar{\Xi}_S} \left\|\xi - \bar{\xi}_s\right\| \leq \frac{1}{S^{1/d}} \ \forall \ \xi \in \Xi.$$

Such a set, $\bar{\Xi}_S$, exists because Ξ was assumed to be compact, see Assumption 1. Since $\gamma(\cdot)$ is Lipschitz continuous (Assumption 1), the above implies

$$\arg\min_{\bar{\xi}_s \in \bar{\Xi}_S} \left\|\gamma(\xi) - \gamma(\bar{\xi}_s)\right\| \leq \frac{1}{S^{1/d}} \ \forall \ \xi \in \Xi. \tag{E.1}$$

Now, by the definition of the GRE estimator,

$$\frac{1}{n}\sum_i l_i(\omega_{i0}) - \frac{1}{n}\sum_i l_i(\hat{\omega}_{i,S}) \leq \frac{1}{n}\sum_i l_i(\omega_{i0}) - \frac{1}{n}\sum_i \max_{\bar{\xi}_s \in \bar{\Gamma}_S} l_i(\gamma(\bar{\xi}_s), \theta_0). \tag{E.2}$$

We start with the right hand side of (E.2). For each i, define

$$\bar{\gamma}_{i,S} := \arg\min_{\bar{\xi}_s \in \bar{\Gamma}_S} \left\|\gamma(\xi_i) - \gamma(\bar{\xi}_s)\right\|,$$

and note that $\sup_i \|\gamma(\xi_i) - \bar{\gamma}_{i,S}\| = O(S^{-1/d})$ by (E.1). Denoting $\bar{\omega}_{i,S} := (\bar{\gamma}_{i,S}, \theta_0)$, we then have

$$\frac{1}{n}\sum_i l_i(\omega_{i0}) - \frac{1}{n}\sum_i \max_{\bar{\xi}_s \in \bar{\Gamma}_S} l_i(\gamma(\bar{\xi}_s), \theta_0) \leq \frac{1}{n}\sum_i l_i(\omega_{i0}) - \frac{1}{n}\sum_i l_i(\bar{\omega}_{i,S}).$$

We can further bound the above using a Taylor series expansion, which implies there exists $\tilde{\omega}_{i,S} := (\tilde{\gamma}_{i,S}, \theta_0)$ such that

$$\begin{aligned}
&\frac{1}{n}\sum_i l_i(\omega_{i0}) - \frac{1}{n}\sum_i l_i(\bar{\omega}_{i,S}) \\
&\leq \frac{1}{n}\sum_i \partial_\gamma l_i(\gamma(\xi_i),\theta_0)^\intercal(\gamma(\xi_i) - \bar{\gamma}_{i,S}) + \frac{1}{n}\sum_i \|\gamma(\xi_i) - \bar{\gamma}_{i,S}\|^2 \cdot \left\|\partial^2_{(\gamma,\gamma^\intercal)} l_i(\tilde{\gamma}_{i,S},\theta_0)\right\| \\
&\leq \left\|\frac{1}{n}\sum_i \partial_\gamma l_i(\gamma(\xi_i),\theta_0)\right\| O(S^{-1/d}) + O(S^{-2/d})\frac{1}{n}\sum_i \left\|\partial^2_{(\gamma,\gamma^\intercal)} l_i(\tilde{\gamma}_{i,S},\theta_0)\right\|.
\end{aligned}$$

Now, by Lemma E.1, the iid structure of $\{\mathbf{y}_i, \mathbf{x}_i, \xi_i\}$, and Assumption D1(vi),

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_i \partial_\gamma l_i(\gamma(\xi_i),\theta_0)\right\|^2\right] \leq \frac{1}{n^2}\sum_i E_{p(\cdot|\xi_i,\theta_0)}\left[\|\partial_\gamma l_i(\gamma(\xi_i),\theta_0)\|^2\right] = O\left(\frac{1}{nT}\right).$$

Furthermore, by Assumptions D1(iii)-(v),

$$\frac{1}{n}\sum_i \left\|\partial^2_\gamma l_i(\tilde{\gamma}_{i,S},\theta_0)\right\| = O_{\mathbb{P}}(1).$$

Combining the above, we thus have

$$\frac{1}{n}\sum_i l_i(\omega_{i0}) - \frac{1}{n}\sum_i \max_{\bar{\xi}_s \in \bar{\Gamma}_S} l_i(\gamma(\bar{\xi}_s),\theta_0) = O_{\mathbb{P}}\left(\frac{S^{-1/d}}{\sqrt{nT}} + S^{-2/d}\right). \tag{E.3}$$

Next, for the left hand side of (E.2), by a Taylor expansion, there exists some $\dot{\omega}_{i,S}$ satisfying $\|\omega_{i0} - \dot{\omega}_{i,S}\| \leq \|\omega_{i0} - \hat{\omega}_{i,S}\|$ such that

$$\begin{aligned}
&\frac{1}{n}\sum_i l_i(\omega_{i0}) - \frac{1}{n}\sum_i l_i(\hat{\omega}_{i,S}) \\
&= \frac{1}{n}\sum_i \partial_\omega l_i(\omega_{i0})^\intercal(\omega_{i0} - \hat{\omega}_{i,S}) + \frac{1}{n}\sum_i (\omega_{i0} - \hat{\omega}_{i,S})^\intercal \partial^2_{(\omega,\omega^\intercal)} l_i(\dot{\omega}_{i,S})(\omega_{i0} - \hat{\omega}_{i,S}).
\end{aligned} \tag{E.4}$$

By Assumption D1(vii),

$$\begin{aligned}
\frac{1}{n}\sum_i \partial_\omega l_i(\omega_{i0})^\intercal(\omega_{i0} - \hat{\omega}_{i,S}) &\leq \left(\frac{1}{n}\sum_i \|\partial_\omega l_i(\omega_{i0})\|^2\right)^{1/2}\left(\frac{1}{n}\sum_i \|\omega_{i0} - \hat{\omega}_{i,S}\|^2\right)^{1/2} \\
&= O_{\mathbb{P}}(T^{-1/2}) \cdot \left(\frac{1}{n}\sum_i \|\omega_{i0} - \hat{\omega}_{i,S}\|^2\right)^{1/2}.
\end{aligned} \tag{E.5}$$

Furthermore, using Assumption D1(ii)-(v), we can argue as in Bonhomme et al. [2017, Corollary E1], to show

$$\frac{1}{n}\sum_i (\omega_{i0} - \hat{\omega}_{i,S})^\intercal \partial^2_{(\omega,\omega^\intercal)} l_i(\dot{\omega}_{i,S})(\omega_{i0} - \hat{\omega}_{i,S}) \geq (\underline{\lambda} + o_{\mathbb{P}}(1))\frac{1}{n}\sum_i \|\omega_{i0} - \hat{\omega}_{i,S}\|^2. \tag{E.6}$$

In view of (E.2)-(E.6),

$$(\underline{\lambda}+o_{\mathbb{P}}(1))\frac{1}{n}\sum_i \|\omega_{i0}-\hat{\omega}_{i,S}\|^2 \leq O_{\mathbb{P}}(T^{-1/2})\cdot\left(\frac{1}{n}\sum_i \|\omega_{i0}-\hat{\omega}_{i,S}\|^2\right)^{1/2}+O_{\mathbb{P}}\left(\frac{S^{-1/d}}{\sqrt{nT}}+S^{-2/d}\right).$$

The above implies

$$\frac{1}{n}\sum_i \|\omega_{i0}-\hat{\omega}_{i,S}\|^2 = O_{\mathbb{P}}\left(\frac{1}{T}+S^{-2/d}+\frac{S^{-1/d}}{\sqrt{nT}}\right).$$

The claim then follows by the definition of ω_{i0} and $\hat{\omega}_{i,S}$. □

E.3. **Assumptions for Theorem 2.** We employ the following notation, due to Arellano and Bonhomme [2009]: Let α_i denote the unobserved heterogeneity corresponding to observation i, and define

$$l_i^{\mathrm{FE}}(\alpha,\theta) := \frac{1}{T}\ln p(\mathbf{y}_i|\mathbf{x}_i,\alpha,\theta).$$

The joint probability over $(\mathbf{y}_i,\mathbf{x}_i)$ for observation i, conditional on α_i is given by $p(\mathbf{y}_i,\mathbf{x}_i|\alpha_i,\theta) \equiv p(\mathbf{y}_i|\mathbf{x}_i,\alpha_i,\theta)p(\mathbf{x}_i|\alpha_i)$, shorthanded as $p(\cdot|\alpha_i,\theta)$. An important quantity is the target likelihood, defined as $\bar{l}_i^{\mathrm{FE}}(\theta) = l_i^{\mathrm{FE}}(\bar{\alpha}(\theta),\theta)$, where

$$\bar{\alpha}_i(\theta) := \arg\max_{\alpha} E_{p(\cdot|\alpha_i,\theta_0)}[l_i^{FE}(\alpha,\theta)].$$

Let $\bar{\theta} = \arg\max_\theta \bar{l}_i^{\mathrm{FE}}(\theta)$ denote the maximizer of target likelihood. Finally, we also define

$$\rho_i(\alpha_i,\theta_0) := -\left\{E_{p(\cdot|\alpha_i,\theta_0)}[\partial^2_{\alpha,\alpha'} l_i^{\mathrm{FE}}(\alpha,\theta)]\right\}^{-1} E_{p(\cdot|\alpha_i,\theta_0)}[\partial^2_{\theta,\alpha'} l_i^{\mathrm{FE}}(\alpha,\theta)].$$

Let $\Gamma_S := (\gamma_1,\ldots,\gamma_S)$ denote any S-dimensional set of prior parameters. We observe that the GRE estimator can be cast as a maximization problem of the form

$$(\hat{\Gamma}_S,\hat{\theta}) := \arg\max_{(\Gamma_S,\theta)} \frac{1}{n}\sum_i \max_{\gamma\in\Gamma_S} l_i(\gamma,\theta).$$

The following regularity conditions (in addition to Assumption D1) enable us to characterize the rate of estimation of θ_0 using $\hat{\theta}$:

Assumption D2. *(i)* $\{y_{it},x_{it}\}_{t=1}^T$ *is stationary in* t*, conditional on* α_i*, for each observation* i*.*

(ii) The domain, Θ*, of* θ *is a compact set. Furthermore, the domain* $\mathcal{C}$ *of* Γ_S *is a compact set satisfying* $\sup_{\gamma\in\Gamma_S}\|\gamma\| \leq M$ *for all* $\Gamma_S \in \mathcal{C}$*, where* $M > \sup_{\xi_i\in\Xi}\|\gamma(\xi_i)\|$*.*

(iii) $l_i^{FE}(\alpha,\theta)$ *is six times continuously differentiable in both its arguments. For* $C<\infty$*, and all* $a,b\in\{1,\ldots,6\}$ *such that* $a+b\leq 6$*, there exists* $M(\mathbf{y}_i,\mathbf{x}_i,\alpha_i)$ *such*

that

$$\sup_{\theta\in\Theta}\left\|\partial_\alpha^a\partial_\theta^b\ l_i^{FE}(\alpha_i,\theta)\right\| \leq M(\mathbf{y}_i,\mathbf{x}_i,\alpha_i)\ \textit{and}\ \mathbb{E}\left[|M(\mathbf{y}_i,\mathbf{x}_i,\alpha_i)|^2\right]<\infty.$$

(iv) $\sup_i \sup_{\alpha,\theta}\left\{l_i^{FE}(\alpha,\theta) - E_{p(\cdot|\alpha_i,\theta_0)}\left[l_i^{FE}(\alpha,\theta)\right]\right\} = O_{\mathbb{P}}(T^{-1/2})$*, and similarly for the first four derivatives of* $l_i^{FE}(\cdot,\cdot)$*.*

(v) $\sup_\theta \|\hat{\alpha}(\theta)-\bar{\alpha}(\theta)\| = O_{\mathbb{P}}(T^{-1/2})$ *. Additionally,* $\|\bar{\alpha}(\theta)-\bar{\alpha}(\theta_0)\| \leq C\|\theta-\theta_0\|$ *for all* θ *in some neighborhood of* θ_0*.*

(vi) There exists $\underline{\lambda}>0$ *such that*

$$\lambda_{min}\left(-\mathbb{E}\left[\partial^2_{(\theta,\theta^\intercal)}\bar{l}_i^{FE}(\theta_0)\right]\right)\geq\underline{\lambda}\ \textit{and}\ \inf_i \lambda_{\min}\left(E_{p(\cdot|\alpha_i,\theta_0)}\left[\partial_\alpha^2 l_i^{FE}(\alpha_i,\theta_0)\right]\right)\geq\underline{\lambda}.$$

(vii) Uniformly over all $\xi_i\in\Xi$*,* $\lim_{\alpha\to\pm\infty}\rho_i(\alpha,\theta_0)\pi(\alpha|\gamma(\xi_i))=0$*. Furthermore,*

$$\mathbb{E}\left[\left\|\frac{1}{\pi(\alpha_i|\gamma(\xi_i))}\frac{\partial}{\partial\alpha}\bigg|_{\alpha_i}\{\pi(\alpha|\gamma(\xi_i))\rho_i(\alpha,\theta_0)\}\right\|\right]<\infty.$$

(viii) $\pi(\cdot|\gamma)$ *is four times differentiable for each* γ *satisfying* $\|\gamma\|\leq M$*. Furthermore,* $\mathbb{E}\left[\left\|\partial_{\alpha,\theta}l_i^{FE}(\alpha_i,\theta_0)\partial_\alpha u(\alpha_i)\right\|^2\right]<\infty$*.*

Assumption D2(i) is made for convenience: our results also apply to non-stationary data at the expense of more involved notation.

Assumptions D2(ii) ensures that the parameter space, $\mathcal{C}\times\Theta$, of the GRE estimator is compact. We will henceforth also impose the requirement of a compact parameter space for the maximization problem defining the GRE estimator

$$(\hat{\Gamma}_S,\hat{\theta}) := \underset{(\Gamma_S,\theta)\in\mathcal{C}\times\Theta}{\arg\max}\frac{1}{n}\sum_i \max_{\gamma\in\Gamma_S} l_i(\gamma,\theta).$$

Admittedly, compactification of the parameter space is unsatisfactory as we do not impose this in computation. However, similarly to its common use in M-estimation, it simplifies the proofs considerably. Note that the imposition of a compact parameter space does not affect the proof of Lemma 1, which involves comparison of $(\hat{\Gamma}_S,\hat{\theta})$ with $(\bar{\Gamma}_S,\theta_0)$ (see the proof for the definition of $\bar{\Gamma}_S$) - the latter still lies within the compact space $\mathcal{C}\times\Theta$.

Assumptions D2(iii) ensures $l_i^{\mathrm{FE}}(\alpha,\theta)$ possesses sufficient regularity for a Laplace approximation to be applicable.

Assumptions D2(iv) and (v) are high level conditions. Assumption D2(iv) requires existence of suitable Donsker theorems (i.e., uniform central limit theorems) for $l_i^{\mathrm{FE}}(\alpha,\theta)$ and its derivatives. These Donsker theorems are further further required

to hold uniformly over all possible values of α_i indexing the underlying distribution $p(\cdot|\alpha_i, \theta_0)$. When $\{y_{it}, x_{it}\}_{t=1}^T$ is iid across t, this requirement can be proved using the techniques in Van Der Vaart and Wellner [1996, Chapter 2.8]. Assumption D2(v) ensures uniform convergence of $\hat{\alpha}(\theta)$ to $\bar{\alpha}(\theta)$ at the $\sqrt{T}$ rate, and also requires $\bar{\alpha}(\cdot)$ to be Lipschitz continuous in some neighborhood of θ_0. In fact, both these conditions can be shown to follow from the other assumptions in D2, but we let this remain as a further assumption for convenience.

Assumption D2(vii) is taken from Arellano and Bonhomme [2009]. It requires the tails of $\pi(\alpha_i|\gamma(\xi_i))$ to be thin enough, which should be easily satisfied for normal priors. Finally, Assumption D2(viii) imposes some regularity conditions on the prior $\pi(\alpha|\gamma)$.

Let $s_i^{(\theta)}$ denote the score for estimation of θ:

$$s_i^{(\theta)} := \left\{\mathbb{E}[\partial^2_{\theta,\theta'} \bar{l}_i^{\mathrm{FE}}(\theta)]\right\}^{-1} \partial_\theta \bar{l}_i^{\mathrm{FE}}(\theta). \tag{E.7}$$

The next set of assumptions enable us to characterize the estimation rates for the marginal effect $\tau_0 := E[\psi(\alpha_i, \theta_0)]$.

Assumption D3. *(i) $\psi(\cdot,\cdot)$ is twice continuously differentiable in both its arguments, and there exists $L(\alpha_i)$ such that for all $a, b \in \{0,1,2\}$ with $a+b \leq 2$,*

$$\sup_{\theta\in\Theta} \left\|\partial_\alpha^a \partial_\theta^b \ \psi(\alpha_i, \theta)\right\| \leq L(\alpha_i) \text{ and } \mathbb{E}\left[|L(\alpha_i)|^2\right] < \infty.$$

(ii) $\mathbb{E}[\left\|s_i^{(\theta)}\right\|^2] < \infty$. Furthermore, $n^{-1}\sum_i s_i^{(\theta)} = O_{\mathbb{P}}(1/\sqrt{nT})$.

(iii) Uniformly over all $\xi_i \in \Xi$, $\lim_{\alpha\to\pm\infty} \pi(\alpha|\gamma(\xi_i)) H_i(\alpha)^{-1} \partial_\alpha \psi(\alpha, \theta_0) = 0$, where $H_i(\alpha) := E_{p(\cdot|\alpha_i,\theta_0)}[-\partial_\alpha^2 l_i^{FE}(\alpha, \theta_0)]$. Furthermore,

$$\mathbb{E}\left[\left\|\frac{1}{\pi(\alpha_i|\gamma(\xi_i))} \frac{\partial}{\partial\alpha}\bigg|_{\alpha_i} \left\{\pi(\alpha|\gamma(\xi_i)) H_i(\alpha)^{-1} \partial_\alpha \psi(\alpha, \theta_0)\right\}\right\|\right] < \infty.$$

(iv) $E\left[\left\|\partial_\alpha \psi(\alpha_i, \theta_0) \partial_\alpha u(\alpha_i)\right\|^2\right] < \infty$.

Assumption D3(i) imposes some regularity conditions on $\psi(\cdot,\cdot)$. Assumption D3(ii) ensures the score function for θ has finite second moments. Furthermore it requires existence of a central limit theorem for $s_i^{(\theta)}$ at $\sqrt{nT}$ rates (note that $\mathbb{E}[s_i^{(\theta)}] = 0$). This is a mild requirement. Assumption D3(iii) is taken from Arellano and Bonhomme [2009].

E.4. **Proof of Theorem 2.** We first characterize the rate of convergence of $\hat{\theta}$ to θ_0. To this end, let us introduce

$$\tilde{\theta} := \arg\min_{\theta} \frac{1}{n} \sum_i l_i(\gamma(\xi_i), \theta).$$

We proceed in a couple of steps. First, we show

$$\tilde{\theta} - \theta_0 = \frac{1}{n} \sum_i s_i^{(\theta)} + O_{\mathbb{P}}\left(\frac{1}{T^2}\right) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{nT}}\right),$$

where $s_i^{(\theta)}$ is defined in (E.7). We then show

$$\hat{\theta} - \tilde{\theta} = O_{\mathbb{P}}\left(\frac{1}{T^{3/2}} + \frac{S^{-1/d}}{T}\right).$$

We start by recalling a few useful results from Arellano and Bonhomme [2009]. Let $q_i(\cdot|\gamma, \theta)$ denote the posterior distribution corresponding to the prior of $\pi(\alpha|\gamma)$, i.e.,

$$q_i(\alpha|\gamma, \theta) := \frac{p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma)}{\int p(\mathbf{y}_i|\mathbf{x}_i, \alpha, \theta)\pi(\alpha|\gamma)d\alpha}.$$

Lemma E.2. *Suppose that Assumptions D1, D2 hold. Then, for any function $m_i(\alpha_i, \theta) \equiv m(\mathbf{y}_i, \mathbf{x}_i, \alpha_i, \theta)$ such that $m_i(\cdot)$ is four times differentiable in its first argument,*

$$E_{q(\cdot|\gamma,\theta)}\left[m_i(\alpha, \theta)\right] = m_i(\hat{\alpha}_i(\theta), \theta) + \frac{\hat{H}_i(\theta)^{-1}}{T} \partial_\alpha \ln \pi\left(\hat{\alpha}_i(\theta)|\gamma\right) \partial_\alpha m_i\left(\hat{\alpha}_i(\theta), \theta\right)$$
$$+ \frac{\hat{H}_i(\theta)^{-1}}{2T} \partial_\alpha^2 m_i\left(\hat{\alpha}_i(\theta), \theta\right) - \frac{\hat{H}_i(\theta)^{-2}\hat{H}_{2i}(\theta)}{2T} \partial_\alpha m_i\left(\hat{\alpha}_i(\theta), \theta\right) + O_{\mathbb{P}}\left(\frac{1}{T^2}\right),$$

where

$$\hat{H}_i(\theta) := -\partial_\alpha^2 l_i^{FE}(\hat{\alpha}_i(\theta), \theta), \text{ and } \hat{H}_{2i}(\theta) := -\partial_\alpha^3 l_i^{FE}(\hat{\alpha}_i(\theta), \theta).$$

The above expression holds uniformly over all i, γ and $\theta \in \Theta$.

Proof. See, Arellano and Bonhomme [2009, Lemma S2] and Tierney et al. [1989, equation 2.6]. □

Lemma E.3. *Suppose that Assumptions D1, D2 hold. Then $\frac{\partial}{\partial\theta}\Big|_{\theta_0} \bar{\alpha}(\theta) = \rho_i(\alpha_i, \theta_0)$*

Proof. See, Arellano and Bonhomme [2009, Lemma A1]. □

Lemma E.4. *Suppose that Assumptions 1, 2 in the main text, and D1, D2 hold. Then, as $S, T, n \to \infty$,*

$$\tilde{\theta} - \theta_0 = \frac{1}{n} \sum_i s_i^{(\theta)} + O_{\mathbb{P}}\left(\frac{1}{T^2}\right) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{nT}}\right).$$

Proof. We start by noting that $\tilde{\theta} - \theta_0 = o_{\mathbb{P}}(1)$ under Assumption D2. This can be shown using standard arguments, e.g., Newey and McFadden [1994].

We have

$$\begin{aligned} 0 &= \frac{1}{n}\sum_i \partial_\theta l_i(\gamma(\xi_i), \tilde{\theta}) \\ &= \frac{1}{n}\sum_i \partial_\theta l_i(\gamma(\xi_i), \theta_0) + \left(\frac{1}{n}\sum_i \partial^2_{\theta,\theta^\intercal} l_i(\gamma(\xi_i), \dot{\theta})\right)^\intercal (\tilde{\theta} - \theta_0), \end{aligned} \tag{E.8}$$

for some $\dot{\theta}$ between $\tilde{\theta}$ and θ_0. Now, by a Laplace approximation argument as in Arellano and Bonhomme [2009, Lemma S1],

$$\frac{1}{n}\sum_i \partial_\theta l_i(\gamma(\xi_i), \theta_0) = \frac{1}{n}\sum_i \partial_\theta \bar{l}_i(\theta_0) + \frac{1}{n}\sum_i \frac{b_i}{T} + O_{\mathbb{P}}\left(\frac{1}{T^2}\right),$$

where

$$\begin{aligned} b_i &:= \partial_\theta \ln \pi(\bar{\alpha}_i(\theta_0)|\gamma(\xi_i)) + \partial_\alpha \rho_i(\alpha_i, \theta_0) \\ &= \frac{1}{\pi(\alpha_i|\gamma(\xi_i))} \partial_\alpha \left\{\pi(\alpha_i|\gamma(\xi_i))\rho_i(\alpha_i, \theta_0)\right\}, \end{aligned}$$

and the last equality follows by Lemma E.3. Hence, by Assumptions D2(vi)-(vii),

$$\begin{aligned} \frac{1}{n}\sum_i b_i &= \mathbb{E}\left[\frac{1}{\pi(\alpha_i|\gamma(\xi_i))} \left.\frac{\partial}{\partial \alpha}\right|_{\alpha_i} \left\{\pi(\alpha|\gamma(\xi_i))\rho_i(\alpha, \theta_0)\right\}\right] + O_{\mathbb{P}}(n^{-1/2}) \\ &= \int_{\xi_i \in \Xi} \int_{-\infty}^{\infty} \left.\frac{\partial}{\partial \alpha}\right|_{\alpha_i} \left\{\pi(\alpha_i|\gamma(\xi_i))\rho_i(\alpha, \theta_0)\right\} d\alpha_i d\xi_i + O_{\mathbb{P}}(n^{-1/2}) = O_{\mathbb{P}}(n^{-1/2}). \end{aligned}$$

The above implies

$$\frac{1}{n}\sum_i \partial_\theta l_i(\gamma(\xi_i), \theta_0) = \frac{1}{n}\sum_i \partial_\theta \bar{l}_i(\theta_0) + O_{\mathbb{P}}\left(\frac{1}{T^2}\right) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{nT}}\right). \tag{E.9}$$

Next, by consistency of $\tilde{\theta}$ (which also implies consistency of $\dot{\theta}$), and employing again a Laplace approximation argument as in Arellano and Bonhomme [2009], we obtain

$$\begin{aligned} \frac{1}{n}\sum_i \partial^2_{\theta,\theta^\intercal} l_i(\gamma(\xi_i), \dot{\theta}) &= \frac{1}{n}\sum_i E_{p(\cdot|\alpha_i,\theta_0)}\left[\partial^2_{\theta,\theta^\intercal} \bar{l}_i^{\mathrm{FE}}(\theta_0)\right] + o_{\mathbb{P}}(1) \\ &= \mathbb{E}[\partial^2_{\theta,\theta^\intercal} \bar{l}_i^{\mathrm{FE}}(\theta_0)] + o_{\mathbb{P}}(1), \end{aligned} \tag{E.10}$$

where the second equality follows by the law of large numbers and Assumption D2(ii).

The claim thus follows from (E.8)-(E.10). □

Lemma E.5. *Suppose that Assumptions 1, 2 in the main text, and D1, D2 hold. Then, as* $S, T, n \to \infty$,

$$\hat{\theta} - \tilde{\theta} = O_{\mathbb{P}}\left(\frac{1}{T^{3/2}} + \frac{S^{-1/d}}{T}\right).$$

Proof. Denote $\hat{\gamma}_{i,S} := \hat{\gamma}_{i,S}(\hat{\theta})$, $\hat{\alpha} := \hat{\alpha}(\hat{\theta})$, $\hat{q}_i(\cdot) := q(\cdot|\hat{\gamma}_{i,S}(\hat{\theta}), \hat{\theta})$, and $\tilde{q}_i(\cdot) := q(\cdot|\gamma(\xi_i), \tilde{\theta})$. We shall also assume that α_i is scalar to simplify the algebra.

By the M step in the EAMP algorithm, we note that $\hat{\theta}$ must solve the first order condition

$$\frac{1}{n}\sum_i E_{\hat{q}_i(\alpha)}\left[\partial_\theta l_i^{\mathrm{FE}}(\alpha, \hat{\theta})\right] = 0.$$

Then, taking $m_i(\alpha, \theta) = \partial_\theta l_i^{\mathrm{FE}}(\alpha, \hat{\theta})$, we obtain by a Laplace approximation as in Lemma E.2 that

$$\frac{1}{n}\sum_i \frac{\partial}{\partial\theta}\bigg|_{\hat{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\hat{\theta}), \theta) + \frac{1}{n}\sum_i \frac{\hat{\beta}_i(\hat{\gamma}_{i,S})}{T} + O_{\mathbb{P}}\left(\frac{1}{T^2}\right) = 0, \tag{E.11}$$

where

$$\begin{aligned}\hat{\beta}_i(\gamma) :=\ & \hat{H}_i^{-1}\partial_\alpha \ln \pi(\hat{\alpha}(\hat{\theta})|\gamma)\partial_{\alpha,\theta} l_i^{\mathrm{FE}}(\hat{\alpha}(\hat{\theta}), \hat{\theta}) + \frac{\hat{H}_i^{-1}}{2}\partial^3_{\alpha,\alpha,\theta} l_i^{\mathrm{FE}}(\hat{\alpha}(\hat{\theta}), \hat{\theta}) \\ & - \frac{\hat{H}_i^{-2}\hat{H}_{2i}}{2}\partial^2_{\alpha,\theta} l_i^{\mathrm{FE}}(\hat{\alpha}(\hat{\theta}), \hat{\theta}),\end{aligned}$$

and we have employed $\hat{H}_i := \hat{H}_i(\hat{\theta})$ and $\hat{H}_{2i} := \hat{H}_{2i}(\hat{\theta})$. Denote $H_i := E_{p(\cdot|\alpha_i,\theta_0)}[\partial^2_\alpha l_i^{\mathrm{FE}}(\alpha_i, \theta_0)]$ and $H_{2i} := E_{p(\cdot|\alpha_i,\theta_0)}[\partial^3_\alpha l_i^{\mathrm{FE}}(\alpha_i, \theta_0)]$. By Assumptions D2(ii)-(v) and $\sqrt{T}$ consistency of $\hat{\theta}$ (Lemma 1), we obtain after some tedious but straightforward algebra that

$$\hat{\beta}_i(\hat{\gamma}_{i,S}) = \beta_i(\hat{\gamma}_{i,S}) + O_{\mathbb{P}}(T^{-1/2}), \text{ uniformly over } i, \tag{E.12}$$

where

$$\begin{aligned}\beta_i(\gamma) :=\ & H_i^{-1}\partial_\alpha \ln \pi\left(\alpha_i|\gamma\right) E_{p(\cdot|\alpha_i,\theta_0)}\left[\partial_{\alpha,\theta} l_i^{\mathrm{FE}}(\alpha_i, \theta_0)\right] + \frac{H_i^{-1}}{2} E_{p(\cdot|\alpha_i,\theta_0)}\left[\partial^3_{\alpha,\alpha,\theta} l_i^{\mathrm{FE}}(\alpha_i, \theta_0)\right] \\ & - \frac{H_i^{-2}H_{2i}}{2} E_{p(\cdot|\alpha_i,\theta_0)}\left[\partial^2_{\alpha,\theta} l_i^{\mathrm{FE}}(\alpha_i, \theta_0)\right].\end{aligned}$$

The derivation of the above requires $|\partial_\alpha \ln \pi(\hat{\alpha}_i|\hat{\gamma}_{i,S}) - \partial_\alpha \ln \pi(\alpha_i|\hat{\gamma}_{i,S})| = O_{\mathbb{P}}(T^{-1/2})$, which necessitates $\hat{\gamma}_{i,S}$ to lie in a compact set (Assumption D2(ii)). In view of (E.11) and (E.12), we thus have

$$\frac{1}{n}\sum_i \frac{\partial}{\partial\theta}\bigg|_{\hat{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\hat{\theta}), \theta) + \frac{1}{n}\sum_i \frac{\beta_i(\hat{\gamma}_{i,S})}{T} + O_{\mathbb{P}}\left(\frac{1}{T^{3/2}}\right) = 0. \tag{E.13}$$

Now, note that $\tilde{\theta}$ can also be characterized as the solution to the first order condition

$$\frac{1}{n}\sum_i E_{\tilde{q}_i(\alpha)}\left[\partial_\theta l_i^{\mathrm{FE}}(\alpha,\tilde{\theta})\right]=0.$$

Hence, by a similar argument as that leading to (E.13), we obtain

$$\frac{1}{n}\sum_i \frac{\partial}{\partial\theta}\bigg|_{\tilde{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\tilde{\theta}),\theta)+\frac{1}{n}\sum_i\frac{\beta_i(\gamma(\xi_i))}{T}+O_{\mathbb{P}}\left(\frac{1}{T^{3/2}}\right)=0. \tag{E.14}$$

In view of (E.13) and (E.14),

$$\frac{1}{n}\sum_i \frac{\partial}{\partial\theta}\bigg|_{\hat{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\hat{\theta}),\theta)-\frac{1}{n}\sum_i \frac{\partial}{\partial\theta}\bigg|_{\tilde{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\tilde{\theta}),\theta)=\frac{1}{n}\sum_i\frac{\beta_i(\gamma(\xi_i))-\beta(\hat{\gamma}_{i,S})}{T}+O_{\mathbb{P}}\left(\frac{1}{T^{3/2}}\right).$$

Now,

$$\begin{aligned}
\frac{\partial}{\partial\theta}\bigg|_{\hat{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\hat{\theta}),\theta) &= \frac{\partial}{\partial\theta}\bigg|_{\hat{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\theta),\theta)\\
&= \frac{\partial}{\partial\theta}\bigg|_{\tilde{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\theta),\theta)+\left(\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\bigg|_{\breve{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\theta),\theta)\right)^\intercal(\hat{\theta}-\tilde{\theta})\\
&= \frac{\partial}{\partial\theta}\bigg|_{\tilde{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\tilde{\theta}),\theta)+\left(\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\bigg|_{\breve{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\theta),\theta)\right)^\intercal(\hat{\theta}-\tilde{\theta}),
\end{aligned}$$

for some $\breve{\theta}$ between $\hat{\theta}$ and $\tilde{\theta}$, where the first and last equalities follow by the envelope theorem. Furthermore, noting that $\hat{\alpha}(\theta)$ solves $\sum_i \partial_\alpha l_i^{\mathrm{FE}}(\hat{\alpha}(\theta),\theta)=0$ (using which $\partial_\theta\hat{\alpha}(\theta)$ and $\partial^2_{\theta,\theta^\intercal}\hat{\alpha}(\theta)$ can be determined by taking the derivatives with respect to ∂_θ and ∂^2_θ on both sides of this expression), we obtain under Assumptions D2(ii)-(vi), and some tedious but straightforward algebra that

$$\frac{1}{n}\sum_i\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\bigg|_{\breve{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\theta),\theta)=\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\bigg|_{\theta_0} l_i^{\mathrm{FE}}(\bar{\alpha}(\theta),\theta)\right]+o_{\mathbb{P}}(1).$$

By Assumption D2(iv), we thus have

$$\lambda_{\min}\left(\frac{1}{n}\sum_i\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\bigg|_{\breve{\theta}} l_i^{\mathrm{FE}}(\hat{\alpha}(\theta),\theta)\right)\geq\underline{\lambda}+o_{\mathbb{P}}(1).$$

The above implies

$$(\underline{\lambda}+o_{\mathbb{P}}(1))\left\|\hat{\theta}-\tilde{\theta}\right\|\leq\frac{1}{nT}\sum_i\left\|\beta_i(\gamma(\xi_i))-\beta_i(\hat{\gamma}_{i,S})\right\|+O_{\mathbb{P}}\left(\frac{1}{T^{3/2}}\right). \tag{E.15}$$

It thus remains to bound $R_i:=n^{-1}\sum_i\left\|\beta_i(\gamma(\xi_i))-\beta_i(\hat{\gamma}_{i,S})\right\|$. Observe that

$$\beta_i(\gamma(\xi_i))-\beta_i(\hat{\gamma}_{i,S})=H_i^{-1}E_{p(\cdot|\alpha_i,\theta_0)}\left[\partial_{\alpha,\theta}l_i^{\mathrm{FE}}(\alpha_i,\theta_0)\right]\frac{\partial}{\partial\alpha}\left\{\ln\pi(\alpha_i|\hat{\gamma}_{i,S})-\ln\pi(\alpha_i|\gamma(\xi_i))\right\}.$$

Now $\lambda_{\max}(H_i^{-1}) \leq \underline{\lambda}^{-1}$ by Assumption D2(v), and $\partial_\alpha \{\ln \pi(\alpha|\gamma_1) - \ln \pi(\alpha|\gamma_2)\} = \partial_\alpha u(\alpha)^\intercal(\gamma_1 - \gamma_2)$ for any γ_1, γ_2 by the definition of the exponential family. Hence,

$$R_i \leq \underline{\lambda}^{-1} \left(\frac{1}{n} \sum_i \left\| E_{p(\cdot|\alpha_i,\theta_0)} \left[\partial_{\alpha,\theta} l_i^{\mathrm{FE}}(\alpha_i, \theta_0) \right] \partial_\alpha u(\alpha_i) \right\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_i \|\hat{\gamma}_{i,S} - \gamma(\xi_i)\|^2 \right)^{1/2}$$
$$= O_{\mathbb{P}}(1) \cdot O_{\mathbb{P}} \left(\frac{1}{T^{1/2}} + S^{-1/d} \right),$$

where the second equality follows from Assumption D2(viii) and Lemma 1. We have thus shown

$$\frac{1}{n} \sum_i \|\beta_i(\gamma(\xi_i)) - \beta_i(\hat{\gamma}_{i,S})\| = O_{\mathbb{P}} \left(\frac{1}{T^{1/2}} + S^{-1/d} \right). \tag{E.16}$$

In view of (E.15) and (E.16),

$$\left\| \hat{\theta} - \tilde{\theta} \right\| = O_{\mathbb{P}} \left(\frac{1}{T^{3/2}} + \frac{S^{-1/d}}{T} \right).$$

This proves the desired claim. □

We now turn to the general case of estimation of the marginal effect $\tau_0 := E[\psi(\alpha_i, \theta_0)]$ using

$$\hat{\tau} = \frac{1}{n} \sum_i E_{\hat{q}_i(\alpha)}[\psi(\alpha, \hat{\theta})].$$

Let $s_i^{(\tau)}$ denote the score function for estimation of τ, defined as

$$s_i^{(\tau)} := \{\psi(\alpha_i, \theta_0) - \tau_0\} + \partial_\alpha \psi(\alpha_i, \theta_0) H_i^{-1} \partial_\alpha l_i^{\mathrm{FE}}(\alpha_i, \theta_0) + \partial_\theta \psi(\alpha_i, \theta_0) \left\{ \frac{1}{n} \sum_i s_i^{(\theta)} \right\}.$$

Here, H_i is defined in the proof of Lemma E.5 and $s_i^{(\theta)}$ is the score equation for estimation of θ, defined in (E.7). We are now ready to prove Theorem 2:

Theorem. *Suppose Assumptions 1, 2 in the main text and D1 - D3 hold. Then, as* $S, T, n \to \infty$,

$$\hat{\tau} - \tau_0 = \frac{1}{n} \sum_i s_i^{(\tau)} + O_{\mathbb{P}} \left(\frac{1}{T^{3/2}} + \frac{S^{-1/d}}{T} \right) + o_{\mathbb{P}} \left(\frac{1}{\sqrt{nT}} \right).$$

Proof. We shall suppose that α_i is a scalar to simplify the notation. Denote $\hat{\tau}_i(\theta) := E_{\hat{q}_i(\alpha)}[\psi(\alpha, \theta)]$, $\partial_\theta \hat{\tau}_i(\theta) := E_{\hat{q}_i(\alpha)}[\partial_\theta \psi(\alpha, \theta)]$ and $\partial^2_{\theta,\theta^\intercal} \hat{\tau}(\theta) := E_{\hat{q}_i(\alpha)}[\partial^2_{\theta,\theta^\intercal} \psi(\alpha, \theta)]$. Observe that $\hat{\tau} = n^{-1} \sum_i \hat{\tau}_i(\hat{\theta})$.

By a Taylor expansion, there exists some $\tilde{\theta}$ between $\hat{\theta}$ ad θ_0 such that

$$\hat{\tau} = \frac{1}{n} \sum_i \hat{\tau}_i(\theta_0) + \left(\frac{1}{n} \sum_i \partial_\theta \hat{\tau}_i(\theta_0) \right) (\hat{\theta} - \theta_0) + (\hat{\theta} - \theta_0)^\intercal \left(\frac{1}{n} \sum_i \partial^2_{\theta,\theta^\intercal} \hat{\tau}_i(\tilde{\theta}) \right) (\hat{\theta} - \theta_0)$$
$$:= A_1 + A_2 + A_3. \tag{E.17}$$

By Lemmas E.4 and E.5, $\left\|\hat{\theta}-\theta_0\right\|^2 = O_{\mathbb{P}}(T^{-3}+S^{-2/d}T^{-2})$. Furthermore, by Assumption D3(i) and some straightforward algebra, $\mathbb{E}\left[\left\|\partial^2_{\theta,\theta^\intercal}\hat{\tau}_i(\tilde{\theta})\right\|\right] \leq \mathbb{E}\left[|L(\alpha_i)|^2\right] < \infty$. Hence,

$$A_3 = O_{\mathbb{P}}\left(\frac{1}{T^3}+\frac{S^{-2/d}}{T^2}\right). \tag{E.18}$$

Next, consider A_2. By Lemmas E.4 and E.5, $\hat{\theta}-\theta_0 = n^{-1}\sum_i s_i^{(\theta)} + O_{\mathbb{P}}\left(\frac{1}{T^{3/2}}+\frac{S^{-1/d}}{T}\right)$. Furthermore under Assumptions D2 and D3, recalling that $\partial_\theta\hat{\tau}_i(\theta) := E_{\hat{q}_i(\alpha)}[\partial_\theta\psi(\alpha,\theta)]$, we can apply a Laplace approximation as in Lemma E.2, with $m_i(\alpha,\theta) = \partial_\theta\psi(\alpha,\theta)$, to obtain

$$\begin{aligned}\partial_\theta\hat{\tau}_i(\theta_0) &= \partial_\theta\psi(\hat{\alpha}_i,\theta_0) + O_{\mathbb{P}}\left(\frac{1}{T}\right)\\ &= \partial_\theta\psi(\alpha_i,\theta_0) + O_{\mathbb{P}}\left(\frac{1}{T}\right) \text{ uniformly over } i.\end{aligned}$$

In view of the above and Assumptions D3(i)-(ii),

$$A_2 = \left(\frac{1}{n}\sum_i \partial_\theta\psi(\alpha_i,\theta_0)\right)\left(\frac{1}{n}\sum_i s_i^{(\theta)}\right) + O_{\mathbb{P}}\left(\frac{1}{T^{3/2}}+\frac{S^{-1/d}}{T}\right). \tag{E.19}$$

Finally, consider A_1. By Arellano and Bonhomme [2009, Lemma S2],

$$\begin{aligned}\hat{\tau}_i(\theta_0) &:= E_{\hat{q}_i(\alpha)}[\psi(\alpha,\theta_0)]\\ &= \psi(\alpha_i,\theta_0) + \partial_\alpha\psi(\alpha_i,\theta_0)H_i^{-1}\partial_\alpha l_i^{\mathrm{FE}}(\alpha_i,\theta_0)\\ &\quad + \frac{1}{T\pi(\alpha_i|\hat{\gamma}_{i,S})}\left.\frac{\partial}{\partial\alpha}\right|_{\alpha_i}\pi(\alpha|\hat{\gamma}_{i,S})\left\{E_{p(\cdot|\alpha_i,\theta_0)}[-\partial^2_\alpha l_i^{\mathrm{FE}}(\alpha,\theta_0)]\right\}^{-1}\partial_\alpha\psi(\alpha,\theta_0).\end{aligned}$$

Recall $H_i(\alpha) := E_{p(\cdot|\alpha_i,\theta_0)}[-\partial^2_\alpha l_i^{\mathrm{FE}}(\alpha,\theta_0)]$. Then,

$$A_1 = \frac{1}{n}\sum_i \psi(\alpha_i,\theta_0) + \frac{1}{n}\sum_i \partial_\alpha\psi(\alpha_i,\theta_0)H_i^{-1}\partial_\alpha l_i^{\mathrm{FE}}(\alpha_i,\theta_0) + \frac{1}{T}\left(\frac{1}{n}\sum_i R_i\right),$$

where

$$\begin{aligned}R_i &:= \frac{1}{\pi(\alpha_i|\hat{\gamma}_{i,S})}\left.\frac{\partial}{\partial\alpha}\right|_{\alpha_i}\pi(\alpha|\hat{\gamma}_{i,S})H_i(\alpha)^{-1}\partial_\alpha\psi(\alpha,\theta_0)\\ &= \frac{1}{\pi(\alpha_i|\gamma(\xi_i))}\left.\frac{\partial}{\partial\alpha}\right|_{\alpha_i}\pi(\alpha|\gamma(\xi_i))H_i(\alpha)^{-1}\partial_\alpha\psi(\alpha,\theta_0)\\ &\quad + H_i(\alpha_i)^{-1}\partial_\alpha\psi(\alpha_i,\theta_0)\left.\frac{\partial}{\partial\alpha}\right|_{\alpha_i}\{\ln\pi(\alpha|\hat{\gamma}_{i,S}) - \ln\pi(\alpha|\gamma(\xi_i))\}\\ &:= R_{1i} + R_{2i}.\end{aligned}$$

Now, by a similar argument as in the proof of Lemma E.4, Assumption D3(iii) implies

$$\frac{1}{n}\sum_i R_{1i} = \mathbb{E}\left[\frac{1}{\pi(\alpha_i|\gamma(\xi_i))}\left.\frac{\partial}{\partial\alpha}\right|_{\alpha_i}\left\{\pi(\alpha|\gamma(\xi_i))H_i(\alpha)^{-1}\partial_\alpha\psi(\alpha,\theta_0)\right\}\right] + O_{\mathbb{P}}(n^{-1/2})$$
$$= \int_{\xi_i\in\Xi}\int_{-\infty}^{\infty}\left.\frac{\partial}{\partial\alpha}\right|_{\alpha_i}\left\{\pi(\alpha|\gamma(\xi_i))H_i(\alpha)^{-1}\partial_\alpha\psi(\alpha,\theta_0)\right\}d\alpha_i d\xi_i + O_{\mathbb{P}}(n^{-1/2}) = O_{\mathbb{P}}(n^{-1/2}).$$

Furthermore, by Assumption D3(iv) and Lemma 1,

$$\frac{1}{n}\sum_i R_{2i} \le \left(\frac{1}{n}\sum_i \left\|H_i(\alpha_i)^{-1}\partial_\alpha\psi(\alpha_i,\theta_0)\partial_\alpha u(\alpha_i)\right\|^2\right)^{1/2}\left(\frac{1}{n}\sum_i\|\hat{\gamma}_{i,S}-\gamma(\xi_i)\|^2\right)^{1/2}$$
$$= O_{\mathbb{P}}(1)\cdot O_{\mathbb{P}}\left(\frac{1}{T^{1/2}}+S^{-1/d}\right).$$

Combining the above, we thus have

$$A_1 = \frac{1}{n}\sum_i\psi(\alpha_i,\theta_0) + \frac{1}{n}\sum_i\partial_\alpha\psi(\alpha_i,\theta_0)H_i^{-1}\partial_\alpha l_i^{\mathrm{FE}}(\alpha_i,\theta_0) + O_{\mathbb{P}}\left(\frac{1}{T^{3/2}}+\frac{S^{-1/d}}{T}\right). \tag{E.20}$$

The claim thus follows from (E.17)-(E.20). □