AIMS Mathematics, 11(8): 23893–23920.
DOI:10.3934/math.2026962
Received: 21 April 2026
Revised: 08 July 2026
Accepted: 21 July 2026
Published: 04 August 2026

https://www.aimspress.com/journal/Math

Research article

Frailty mixtures for count regression: separating within- and between-class heterogeneity

Mohieddine Rahmouni*

Applied College, King Faisal University, Al-Ahsa, Saudi Arabia

* **Correspondence:** Email: mrahmouni@kfu.edu.sa.

Abstract: We proposed a finite exponential-gamma frailty mixture (FEGFM) regression model for recurrent-event count data observed over a fixed follow-up window. The model targets settings with overdispersion, heavy upper tails, and latent subgroup structure that are not well-captured by a single Poisson or negative binomial regression. Each latent component has its own log-linear mean regression and gamma frailty parameter, while component membership depends on covariates through multinomial logistic gating. After integrating out frailty, each component follows a negative binomial distribution, so the overall model is a covariate-dependent finite mixture of negative binomial regressions. We derived the main distributional properties, identifiability conditions, and an expectation-maximization algorithm for maximum-likelihood estimation. Simulation results showed improved fit and tail calibration relative to Poisson and single-component negative binomial benchmarks, while also indicating weak identification for some parameters in finite samples. An application to physician-visit counts illustrates the practical value of separating within-class frailty from between-class latent heterogeneity.

Keywords: recurrent events; count regression; finite mixtures; negative binomial; gamma frailty; EM algorithm; latent heterogeneity; overdispersion
Mathematics Subject Classification: 62J12, 62F10, 62P10

1. Introduction

Recurrent-event data arise whenever a subject or experimental unit may experience the same type of event multiple times during a monitoring period. Canonical examples include repeat hospital admissions, disease exacerbations, warranty claims, equipment failures, emergency-service contacts, and re-offending incidents [1,2]. In many such applications, the primary quantity of interest is the total number of events within a fixed observation window, analyzed as a function of measured covariates.

Poisson regression is the natural starting point for count outcomes, but its equidispersion

requirement—that the conditional mean equals the conditional variance—is routinely violated in recurrent-event settings. Two qualitatively distinct sources of excess variation are commonly observed. First, even within a covariate cell, subjects vary in unmeasured susceptibility so that the realized count variance exceeds the Poisson prediction. This *within-cell* or *frailty-driven* overdispersion is well-handled by the negative binomial (NB) regression model, which arises from mixing the Poisson mean over a gamma frailty and thereby provides one additional dispersion parameter [3]. Second, the population may consist of distinct latent risk groups whose members differ not only in their average event rates but also in their degree of residual susceptibility. Imposing a single NB mean-regression surface and a single dispersion parameter over such a population blurs covariate effects, misallocates tail mass, and suppresses interpretively useful structure.

Finite mixture models offer a principled representation of such population heterogeneity [4, 5]. Each component captures a latent risk stratum with its own regression parameters, while covariate-dependent mixing weights allow the data to inform membership probabilities. The mixture-of-NB family has received attention in the econometric literature on count data [6–8], and related zero-inflated and hurdle constructions are well-established [9, 10]. Mixture-based representations of population heterogeneity continue to be developed in adjacent biomedical and survival contexts; [11], for instance, proposed a unified framework for complex survival data that accounts for clustering, cure fractions, and competing risks, illustrating the continued relevance of latent-class heterogeneity modeling beyond count regression.

Three strands of this literature are especially close to the present proposal, and it is worth being explicit about how the finite exponential-gamma frailty mixture (FEGFM) advances each of them. First, *concomitant-variable* NB mixtures such as the covariate-dependent mixed Poisson models of [7] and the FlexMix family [8,12] allow the mixing weights to depend on covariates, exactly as the FEGFM does, but they impose either a single dispersion parameter shared across components or no explicit frailty layer at all; the within-component overdispersion and the between-component heterogeneity are consequently conflated in a single variance term. Second, the two-component NB mixture of [6]—which we revisit directly in the empirical application below—uses *constant* (covariate-free) mixing weights, so it cannot let the covariate vector simultaneously shift the within-component mean and reallocate probability mass across risk strata. Third, standard frailty models [13, 14] introduce a single gamma frailty layer to generate one NB marginal, with no mechanism for latent subgroup structure at all. The FEGFM differs from each of these in combining *component-specific* gamma frailty parameters with *covariate-dependent* multinomial logistic gating in one coherent hierarchy; to our knowledge, this particular combination, together with the closed-form mean-variance decomposition it implies (Proposition 4 below), has not been studied in full generality for recurrent-event counts. This formulation sits at the intersection of three established techniques—frailty models [13, 14], finite mixture regression [5, 12], and count regression [1]—and enriches each by allowing the dispersion structure to vary across latent subpopulations.

The present paper fills this gap. We propose the *finite exponential-gamma frailty mixture* (FEGFM) regression model, in which each of M latent components has its own log-linear mean regression $\mu_m(\mathbf{x}) = \exp(\beta_{0m} + \mathbf{x}^\top \boldsymbol{\beta}_m)$ and its own gamma frailty parameter α_m. Component membership is governed by multinomial logistic gating so that the effective mixing proportions are covariate-dependent. Integrating out the gamma frailty within each component yields an NB marginal distribution, and the overall observed-data model is therefore a finite mixture of NB regressions with observation-level

covariate-dependent weights.

The principal novel contribution, relative to the three strands of research surveyed above, is the joint specification of component-specific frailty with covariate-dependent gating, which yields a transparent, closed-form separation of within-class frailty-driven overdispersion from between-class latent heterogeneity (Section 4). Implementing and validating this model computationally, and demonstrating its practical value on physician-visit data, are secondary contributions that support and illustrate this theoretical point rather than constituting independent novelty claims.

The contributions of this paper are fourfold.

(1) We develop the complete hierarchical specification of the FEGFM model and derive its marginal distribution, probability generating function, mean-variance decomposition, posterior frailty law, and the nested special cases that include Poisson regression, NB regression, and finite mixtures of Poisson regressions.
(2) We establish sufficient conditions for identification of the model parameters up to label-switching permutations.
(3) We derive an expectation-maximization (EM) algorithm for maximum-likelihood estimation, implemented with analytic gradients via the Limited-memory Broyden–Fletcher–Goldfarb–Shanno with Bound constraints (L-BFGS-B), and prove that the observed-data log-likelihood is nondecreasing across iterations. We provide practical guidance on initialization, component ordering, and uncertainty quantification.
(4) We conduct a fully reproducible synthetic-data experiment that benchmarks the FEGFM against Poisson and single-component NB baselines, and we report a Monte Carlo assessment of estimation stability that identifies both the strengths and the structural weaknesses of the current implementation.

The remainder of the paper is organized as follows: Section 2 sets out the hierarchical framework. Section 3 derives the component and mixture distributions. Section 4 establishes core theoretical properties. Section 5 develops the EM estimator. Section 6 presents the simulation evidence. Section 7 applies the model to physician-visit counts. Section 8 discusses limitations and extensions. Section 9 concludes the paper.

2. Modeling framework

2.1. Observed data structure

Let $i = 1, \ldots, n$ index subjects. For subject i, let $Y_i \in \{0, 1, 2, \ldots\}$ denote the total number of recurrent events recorded during a fixed follow-up window of common length, and let $\mathbf{x}_i \in \mathbb{R}^p$ be a vector of observed covariates (unequal follow-up durations across subjects can be accommodated via a $\log(t_i)$ exposure offset in each component's log-mean specification, without altering the derivations below). The inferential target is the conditional distribution of Y_i given $\mathbf{x}_i$.

2.2. Frailty-based overdispersion within a single component

We first review the standard Poisson-gamma frailty mechanism. Suppose that, conditionally on a latent nonnegative random variable Z_i, the count follows:

$$Y_i \mid Z_i, \mathbf{x}_i \sim \text{Poisson}(\mu(\mathbf{x}_i) Z_i),$$

where $\mu(\mathbf{x}_i) > 0$ is the baseline mean implied by observed covariates. If the frailty Z_i follows a unit-mean gamma distribution,

$$Z_i \sim \text{Gamma}(\alpha, \alpha), \qquad \alpha > 0,$$

so that $\mathbb{E}(Z_i) = 1$ and $\text{Var}(Z_i) = \alpha^{-1}$. Then, marginalizing over Z_i yields the negative binomial distribution with mean $\mu(\mathbf{x}_i)$ and shape parameter α. Smaller α implies greater frailty variance and therefore stronger overdispersion.

2.3. Latent classes and covariate-dependent gating

To represent subpopulation structure, we introduce a latent class indicator $C_i \in \{1, \ldots, M\}$. We allow the class probabilities to depend on covariates through multinomial logistic gating:

$$\pi_m(\mathbf{x}_i) = \Pr(C_i = m \mid \mathbf{x}_i) = \frac{\exp(\gamma_{0m} + \mathbf{x}_i^\top \boldsymbol{\gamma}_m)}{\sum_{j=1}^{M} \exp(\gamma_{0j} + \mathbf{x}_i^\top \boldsymbol{\gamma}_j)}, \qquad m = 1, \ldots, M. \tag{2.1}$$

One component is designated the reference class ($\gamma_{01} = 0$, $\boldsymbol{\gamma}_1 = \mathbf{0}$) for parameter identifiability. Covariate-dependent gating is a key feature of the FEGFM model: it allows the covariate vector to simultaneously shift the expected event rate *within* each component and to reallocate probability mass *across* components, thus capturing both mean-regression effects and risk-stratification effects.

3. The FEGFM model

3.1. Hierarchical specification

Conditional on class membership $C_i = m$, the FEGFM model posits

$$Y_i \mid (C_i = m, Z_{im}, \mathbf{x}_i) \sim \text{Poisson}(\mu_m(\mathbf{x}_i)\, Z_{im}), \tag{3.1}$$

$$Z_{im} \sim \text{Gamma}(\alpha_m, \alpha_m), \qquad \alpha_m > 0, \tag{3.2}$$

$$\mu_m(\mathbf{x}_i) = \exp(\beta_{0m} + \mathbf{x}_i^\top \boldsymbol{\beta}_m). \tag{3.3}$$

The parameters $(\beta_{0m}, \boldsymbol{\beta}_m, \alpha_m)$ are component-specific: each latent class has its own mean-regression surface and its own frailty level. The overall FEGFM model is specified by the parameter set

$$\Theta = \{\beta_{0m}, \boldsymbol{\beta}_m, \alpha_m, \gamma_{0m}, \boldsymbol{\gamma}_m : m = 1, \ldots, M\}.$$

3.2. Marginal component distribution

Theorem 1 (Gamma frailty induces a negative binomial component)**.** *Under* (3.1)–(3.3)*, the marginal conditional distribution of Y_i given $(C_i = m, \mathbf{x}_i)$ is a negative binomial with mean $\mu_m(\mathbf{x}_i)$ and shape parameter α_m:*

$$\Pr(Y_i = y \mid C_i = m, \mathbf{x}_i) = \frac{\Gamma(y + \alpha_m)}{\Gamma(\alpha_m)\, y!} \left(\frac{\alpha_m}{\alpha_m + \mu_m(\mathbf{x}_i)}\right)^{\alpha_m} \left(\frac{\mu_m(\mathbf{x}_i)}{\alpha_m + \mu_m(\mathbf{x}_i)}\right)^{y}, \quad y = 0, 1, 2, \ldots. \tag{3.4}$$

Proof. By the law of total probability,

$$\Pr(Y_i = y \mid C_i = m, \mathbf{x}_i) = \int_0^\infty \frac{(\mu_m z)^y e^{-\mu_m z}}{y!} \cdot \frac{\alpha_m^{\alpha_m}}{\Gamma(\alpha_m)}\, z^{\alpha_m - 1} e^{-\alpha_m z}\, dz,$$

where $\mu_m = \mu_m(\mathbf{x}_i)$ for brevity. Collecting powers of z and the constant factors gives

$$\frac{\alpha_m^{\alpha_m}\mu_m^y}{\Gamma(\alpha_m)\,y!}\int_0^\infty z^{y+\alpha_m-1}e^{-(\alpha_m+\mu_m)z}\,dz.$$

Applying the gamma integral $\int_0^\infty z^{a-1}e^{-bz}\,dz = \Gamma(a)\,b^{-a}$ with $a = y + \alpha_m$ and $b = \alpha_m + \mu_m$ yields

$$\frac{\Gamma(y+\alpha_m)}{\Gamma(\alpha_m)\,y!}\frac{\alpha_m^{\alpha_m}\mu_m^y}{(\alpha_m+\mu_m)^{y+\alpha_m}},$$

which rearranges to (3.4). □

Corollary 1 (Component moments and Poisson limit)**.** *For each component m,*

$$\mathbb{E}(Y_i \mid C_i = m, \mathbf{x}_i) = \mu_m(\mathbf{x}_i), \qquad \mathrm{Var}(Y_i \mid C_i = m, \mathbf{x}_i) = \mu_m(\mathbf{x}_i) + \frac{\mu_m(\mathbf{x}_i)^2}{\alpha_m}.$$

In particular, $\alpha_m \to \infty$ recovers the Poisson component.

Proof. These are the standard first and second moments of the $\mathrm{NB}(\mu_m, \alpha_m)$ distribution. The Poisson limit follows because $\mu_m^2/\alpha_m \to 0$. □

3.3. Observed mixture distribution

Marginalizing over the latent class indicator C_i gives the FEGFM observed-data distribution:

$$\Pr(Y_i = y \mid \mathbf{x}_i) = \sum_{m=1}^{M} \pi_m(\mathbf{x}_i)\,\Pr(Y_i = y \mid C_i = m, \mathbf{x}_i), \tag{3.5}$$

a finite mixture of NB regressions with covariate-dependent mixing weights.

3.4. Probability generating function and zero-count probability

Proposition 1 (Probability generating function)**.** *For fixed $\mathbf{x}_i$, the conditional probability generating function (PGF) of Y_i is*

$$G_{Y_i|\mathbf{x}_i}(s) = \sum_{m=1}^{M} \pi_m(\mathbf{x}_i)\left(\frac{\alpha_m}{\alpha_m + \mu_m(\mathbf{x}_i)(1-s)}\right)^{\alpha_m}, \qquad |s| \le 1. \tag{3.6}$$

In particular, the zero-count probability is

$$\Pr(Y_i = 0 \mid \mathbf{x}_i) = \sum_{m=1}^{M} \pi_m(\mathbf{x}_i)\left(\frac{\alpha_m}{\alpha_m + \mu_m(\mathbf{x}_i)}\right)^{\alpha_m}. \tag{3.7}$$

Proof. The PGF of the $\mathrm{NB}(\mu_m, \alpha_m)$ component is $G_m(s) = \left[\alpha_m/\{\alpha_m + \mu_m(\mathbf{x}_i)(1-s)\}\right]^{\alpha_m}$. Mixing over components with weights $\pi_m(\mathbf{x}_i)$ yields (3.6); evaluating at $s = 0$ gives (3.7). □

The PGF provides a straightforward route to all factorial cumulants of the mixture distribution and can be used to construct moment-based estimators or goodness-of-fit tests.

4. Theoretical properties

4.1. Mean-variance decomposition

Proposition 2 (Mixture mean and variance)**.** *For fixed* $\mathbf{x}_i$*, the conditional mean and variance of the FEGFM model are*

$$\mathbb{E}(Y_i \mid \mathbf{x}_i) = \sum_{m=1}^{M} \pi_m(\mathbf{x}_i)\, \mu_m(\mathbf{x}_i), \tag{4.1}$$

$$\mathrm{Var}(Y_i \mid \mathbf{x}_i) = \underbrace{\sum_{m=1}^{M} \pi_m(\mathbf{x}_i) \left\{ \mu_m(\mathbf{x}_i) + \frac{\mu_m(\mathbf{x}_i)^2}{\alpha_m} \right\}}_{\textit{within-component}} + \underbrace{\sum_{m=1}^{M} \pi_m(\mathbf{x}_i) \{\mu_m(\mathbf{x}_i) - \mathbb{E}(Y_i \mid \mathbf{x}_i)\}^2}_{\textit{between-component}}. \tag{4.2}$$

Proof. Equation (4.1) is the law of total expectation. Equation (4.2) follows from the law of total variance, $\mathrm{Var}(Y_i \mid \mathbf{x}_i) = \mathbb{E}[\mathrm{Var}(Y_i \mid C_i, \mathbf{x}_i) \mid \mathbf{x}_i] + \mathrm{Var}[\mathbb{E}(Y_i \mid C_i, \mathbf{x}_i) \mid \mathbf{x}_i]$, combined with Corollary 1. □

Remark 1. *The decomposition* (4.2) *is the key interpretive payoff of the FEGFM formulation. The within-component term measures overdispersion arising from gamma frailty inside each latent class; its contribution through* μ_m^2/α_m *vanishes as* $\alpha_m \to \infty$*. The between-component term measures additional variance from the heterogeneity of conditional means across classes. A single NB regression conflates both sources into one dispersion parameter, making it impossible to disentangle frailty from latent class separation. The FEGFM identifies them separately.*

4.2. Structural overdispersion

Proposition 3 (Structural overdispersion)**.** *Suppose that either (i) at least one active component has* $\alpha_m < \infty$*, or (ii) at least two active components have different conditional means at* $\mathbf{x}_i$*. Then*

$$\mathrm{Var}(Y_i \mid \mathbf{x}_i) \;\geq\; \mathbb{E}(Y_i \mid \mathbf{x}_i),$$

with strict inequality unless all active components are Poisson and share the same conditional mean.

Proof. Subtracting (4.1) from (4.2) gives

$$\sum_{m=1}^{M} \pi_m(\mathbf{x}_i)\, \frac{\mu_m(\mathbf{x}_i)^2}{\alpha_m} + \sum_{m=1}^{M} \pi_m(\mathbf{x}_i) \{\mu_m(\mathbf{x}_i) - \mathbb{E}(Y_i \mid \mathbf{x}_i)\}^2.$$

Both terms are nonnegative; the first vanishes only when all $\alpha_m = \infty$, and the second only when all active component means are equal. □

4.3. Posterior frailty distribution

Theorem 2 (Posterior frailty)**.** *Under the FEGFM hierarchy, the conditional distribution of the frailty variable* Z_{im} *given class membership* $C_i = m$*, the observed count* $Y_i = y$*, and covariates* $\mathbf{x}_i$ *is*

$$Z_{im} \mid (Y_i = y,\, C_i = m,\, \mathbf{x}_i) \;\sim\; \mathrm{Gamma}(\alpha_m + y,\; \alpha_m + \mu_m(\mathbf{x}_i)).$$

Consequently,

$$\mathbb{E}[Z_{im} \mid Y_i = y, C_i = m, \mathbf{x}_i] = \frac{\alpha_m + y}{\alpha_m + \mu_m(\mathbf{x}_i)}, \tag{4.3}$$

$$\mathbb{E}[\log Z_{im} \mid Y_i = y, C_i = m, \mathbf{x}_i] = \psi(\alpha_m + y) - \log\{\alpha_m + \mu_m(\mathbf{x}_i)\}, \tag{4.4}$$

where $\psi(\cdot)$ denotes the digamma function.

Proof. The conditional density of Z_{im} given $(Y_i = y, C_i = m, \mathbf{x}_i)$ satisfies, up to a normalizing constant,

$$p(z \mid y, C_i = m, \mathbf{x}_i) \;\propto\; z^y e^{-\mu_m(\mathbf{x}_i)z} \cdot z^{\alpha_m - 1} e^{-\alpha_m z} \;=\; z^{\alpha_m + y - 1} e^{-(\alpha_m + \mu_m(\mathbf{x}_i))z}.$$

This is the kernel of a Gamma$(\alpha_m + y,\ \alpha_m + \mu_m(\mathbf{x}_i))$ density. Equations (4.3) and (4.4) are the corresponding mean and log-mean of that gamma distribution. □

Remark 2. *Theorem 2 establishes that the Poisson-gamma conjugacy is preserved within each component. This conjugacy can be exploited in expectation-conditional maximization (ECM)-style EM variants in which the frailty variables are treated as part of the complete data, enabling closed-form frailty updates inside the M-step [15].*

4.4. Special and limiting cases

Proposition 4 (Nested models)**.** *The FEGFM model contains the following models as special cases.*

(1) If $M = 1$, the model reduces to single-component NB regression.
(2) If $M = 1$ and $\alpha_1 \to \infty$, it reduces to Poisson regression.
(3) If all $\alpha_m \to \infty$ with $M > 1$, it becomes a finite mixture of Poisson regressions with covariate-dependent weights.

Proof. Each statement follows from Theorem 1, Corollary 1, and (3.5). □

The nesting structure implies that standard likelihood-ratio tests can in principle be used to test the adequacy of any nested baseline model, although the irregular boundary behavior of mixture models requires care in the interpretation of asymptotic null distributions [16].

4.5. Identifiability

Assumption 1 (Identifiability conditions)**.**

(1) The covariate support $\mathcal{X}$ contains an open set in $\mathbb{R}^p$.
(2) For any $m \neq j$, the parameter triples $(\beta_{0m}, \boldsymbol{\beta}_m, \alpha_m)$ and $(\beta_{0j}, \boldsymbol{\beta}_j, \alpha_j)$ are distinct.
(3) The gating vectors $(\gamma_{0m}, \boldsymbol{\gamma}_m)$ are not all identical, so the mixing weights vary nondegenerately in $\mathbf{x}$.

Theorem 3 (Identifiability up to label switching)**.** *Under Assumption 1, the conditional distribution $\Pr(Y_i = y \mid \mathbf{x}_i)$ uniquely determines Θ up to permutation of the component labels.*

Proposition 5 (Distinctness of component laws)**.** *For any two parameter pairs $(\mu, \alpha) \neq (\mu', \alpha')$ with $\mu, \mu' > 0$ and $\alpha, \alpha' > 0$, the probability generating functions of $\mathrm{NB}(\mu, \alpha)$ and $\mathrm{NB}(\mu', \alpha')$,*

$$G(s) = \left(\frac{\alpha}{\alpha + \mu(1 - s)}\right)^{\alpha}, \qquad G'(s) = \left(\frac{\alpha'}{\alpha' + \mu'(1 - s)}\right)^{\alpha'},$$

differ on the unit disk $|s| \leq 1$, and hence the corresponding mass functions are distinct on $\{0, 1, 2, \ldots\}$.

Proof. Each PGF is real-analytic on $|s| < 1 + \alpha/\mu$ (respectively, $1 + \alpha'/\mu'$), hence it is determined by its derivatives at $s = 0$, which are the (factorial) moments of the corresponding NB law. The first two moments of $\mathrm{NB}(\mu, \alpha)$ are μ and $\mu + \mu^2/\alpha$ (Corollary 1). If $\mu \neq \mu'$ the means already differ, so $G \not\equiv G'$. If $\mu = \mu'$ but $\alpha \neq \alpha'$, the variances $\mu + \mu^2/\alpha$ and $\mu + \mu^2/\alpha'$ differ, again giving $G \not\equiv G'$. Since two distinct PGFs cannot generate the same sequence of probabilities, the mass functions are distinct on $\{0, 1, 2, \ldots\}$. □

Proof of Theorem 3. The argument proceeds in two stages.

Stage 1 (component identifiability). By Assumption 1(2), the parameter triples $(\beta_{0m}, \boldsymbol{\beta}_m, \alpha_m)$ are pairwise distinct, so for any fixed $\mathbf{x}$, the implied pairs $(\mu_m(\mathbf{x}), \alpha_m)$ are pairwise distinct for at least one $\mathbf{x}$ (and, by continuity of $\mu_m(\cdot)$, on an open subset of $\mathcal{X}$). Proposition 5 then shows the component mass functions $f_m(\cdot \mid \mathbf{x})$ are pairwise distinct on $\{0, 1, 2, \ldots\}$ for $\mathbf{x}$ in this subset.

Stage 2 (gating identifiability). Fix any finite set of points $\mathbf{x}_1, \ldots, \mathbf{x}_K$ in the interior of $\mathcal{X}$ at which the component densities are pairwise distinct (Stage 1). Write $\Pi \in \mathbb{R}^{K \times M}$ for the matrix with entries $\Pi_{km} = \pi_m(\mathbf{x}_k)$. Assumption 1(3) states that the gating vectors $(\gamma_{0m}, \boldsymbol{\gamma}_m)$ are not all identical; together with Assumption 1(1), this implies that $\mathbf{x} \mapsto (\pi_1(\mathbf{x}), \ldots, \pi_M(\mathbf{x}))$ is a non-constant, real-analytic map of $\mathbf{x}$. Consequently, for K sufficiently large and $\mathbf{x}_1, \ldots, \mathbf{x}_K$ chosen generically, Π has full column rank M; this is the rank condition required by [17, Theorem 3.1] and by [18, Theorem 2] for identifiability of finite mixtures with covariate-dependent, distinct component laws. Now consider the observed mixture densities $g(\cdot \mid \mathbf{x}_k) = \sum_m \Pi_{km} f_m(\cdot \mid \mathbf{x}_k)$ at these K points. Full column rank of Π means the only way to write each $g(\cdot \mid \mathbf{x}_k)$ as a mixture of the (pairwise distinct, by Stage 1) component densities with weights summing to one is the true decomposition, up to a simultaneous permutation of the component labels applied to both Π and the densities f_m. The gating vectors $(\gamma_{0m}, \boldsymbol{\gamma}_m)$ in turn determine Π at $\mathbf{x}_1, \ldots, \mathbf{x}_K$ bijectively up to this same permutation, since multinomial logit is itself identified given the mixing probabilities at $K \geq M$ generic points, by the same full-rank argument. It follows that Θ is identified up to label switching. □

Remark 3. *Theorem 3 provides a* sufficient *condition. In finite samples, near-nonidentifiability may arise when two components are weakly separated in both mean and dispersion, or when one component receives negligible posterior weight. These practical concerns motivate the diagnostic recommendations in Section 5 and the Monte Carlo findings in Section 6.*

Remark 4 (Discrete covariate support)**.** *Assumption 1(1) requires $\mathcal{X}$ to contain an open set in $\mathbb{R}^p$, which is sufficient but not necessary for the rank condition used in Stage 2 above. When the covariate support is discrete (as in the chronic-conditions covariate of Section 7, which takes only integer values), the relevant condition is instead that the number of distinct covariate values exceeds the number of free gating parameters, so that a matrix Π of full column rank M can still be formed from the (finitely many) distinct support points [17, Thm. 3.1]. In the empirical application, the chronic-conditions covariate takes 9 distinct integer values $(0, 1, \ldots, 8)$, which exceeds the $M - 1 = 1$ free gating vectors (equivalently, 2 free gating parameters) of the two-component model used there, so identifiability is not threatened by the discreteness of that covariate.*

5. Maximum-likelihood estimation via EM

5.1. Observed-data likelihood

The observed-data log-likelihood is

$$\ell(\Theta) = \sum_{i=1}^{n} \log\left\{\sum_{m=1}^{M} \pi_m(\mathbf{x}_i;\Theta)\, f_m(y_i \mid \mathbf{x}_i;\Theta)\right\}, \tag{5.1}$$

where f_m denotes the NB mass function (3.4). Direct maximization of (5.1) is complicated by nonconcavity and the presence of multiple local maxima, standard features of finite-mixture likelihoods [5]. The EM algorithm of [19] provides an iterative surrogate that guarantees nondecreasing likelihood values.

5.2. Complete-data formulation

We now introduce binary allocation indicators $W_{im} = \mathbf{1}(C_i = m)$, $\sum_{m=1}^{M} W_{im} = 1$. The complete-data log-likelihood is

$$\ell_c(\Theta) = \sum_{i=1}^{n}\sum_{m=1}^{M} W_{im}\{\log \pi_m(\mathbf{x}_i) + \log f_m(y_i \mid \mathbf{x}_i)\}. \tag{5.2}$$

Substituting (3.4) expands the component log-density term as

$$\log f_m(y_i \mid \mathbf{x}_i) = \log\Gamma(y_i+\alpha_m) - \log\Gamma(\alpha_m) - \log(y_i!) + \alpha_m \log\left(\frac{\alpha_m}{\alpha_m+\mu_m(\mathbf{x}_i)}\right) + y_i \log\left(\frac{\mu_m(\mathbf{x}_i)}{\alpha_m+\mu_m(\mathbf{x}_i)}\right). \tag{5.3}$$

5.3. E-step: posterior class responsibilities

At iteration t, the E-step replaces each unknown indicator W_{im} by its conditional expectation:

$$\tau_{im}^{(t+1)} = \Pr(C_i = m \mid y_i, \mathbf{x}_i; \Theta^{(t)}) = \frac{\pi_m(\mathbf{x}_i;\Theta^{(t)})\, f_m(y_i \mid \mathbf{x}_i;\Theta^{(t)})}{\sum_{j=1}^{M} \pi_j(\mathbf{x}_i;\Theta^{(t)})\, f_j(y_i \mid \mathbf{x}_i;\Theta^{(t)})}. \tag{5.4}$$

These soft allocations satisfy $\sum_{m=1}^{M} \tau_{im}^{(t+1)} = 1$ for each i. When working with the deeper Poisson-gamma hierarchy, Theorem 2 additionally provides closed-form posterior moments for the frailty variables, which are useful in ECM-style implementations and individual-level susceptibility diagnostics.

5.4. M-step: parameter updates

The M-step maximizes the surrogate

$$Q(\Theta \mid \Theta^{(t)}) = \sum_{i=1}^{n}\sum_{m=1}^{M} \tau_{im}^{(t+1)}\{\log \pi_m(\mathbf{x}_i) + \log f_m(y_i \mid \mathbf{x}_i)\}. \tag{5.5}$$

The maximization separates into two independent blocks.

Gating update. The gating parameters solve the weighted multinomial logistic regression

$$\max_{\{\gamma_{0m},\, \boldsymbol{\gamma}_m\}} \sum_{i=1}^{n}\sum_{m=1}^{M} \tau_{im}^{(t+1)} \log \pi_m(\mathbf{x}_i),$$

which is a concave maximization problem solvable by standard iteratively reweighted least squares.

Component updates. For each $m = 1, \ldots, M$, the mean-regression and frailty parameters solve

$$\max_{\beta_{0m}, \boldsymbol{\beta}_m, \alpha_m} \sum_{i=1}^{n} \tau_{im}^{(t+1)} \log f_m(y_i \mid \mathbf{x}_i).$$

This is a weighted NB regression problem with observation i receiving effective weight $\tau_{im}^{(t+1)}$. We implement this step using L-BFGS-B with analytic gradients derived from the NB log-density. Specifically, parametrizing $\eta_m = \log \alpha_m$ to enforce positivity, the gradient with respect to $(\beta_{0m}, \boldsymbol{\beta}_m, \eta_m)$ has a closed form:

$$\frac{\partial}{\partial \beta_{0m}} \sum_i \tau_{im} \log f_m(y_i \mid \mathbf{x}_i) = \sum_i \tau_{im}\, \mu_m(\mathbf{x}_i) \left(\frac{y_i}{\mu_m(\mathbf{x}_i)} - \frac{y_i + \alpha_m}{\alpha_m + \mu_m(\mathbf{x}_i)} \right),$$

$$\frac{\partial}{\partial \eta_m} \sum_i \tau_{im} \log f_m(y_i \mid \mathbf{x}_i) = \alpha_m \sum_i \tau_{im} \left[\psi(y_i + \alpha_m) - \psi(\alpha_m) + \log \frac{\alpha_m}{\alpha_m + \mu_m(\mathbf{x}_i)} + 1 - \frac{y_i + \alpha_m}{\alpha_m + \mu_m(\mathbf{x}_i)} \right],$$

with the gradient for $\boldsymbol{\beta}_m$ obtained by replacing $\mu_m(\mathbf{x}_i)$ by $\mu_m(\mathbf{x}_i)\, x_{ij}$ in the first expression. This gradient-based M-step is substantially faster than derivative-free alternatives such as Nelder-Mead and yields well-defined convergence tolerances.

Because L-BFGS-B is run for a finite number of inner iterations rather than to exact convergence at each M-step, the implemented procedure is, strictly speaking, a *generalized* EM (GEM) algorithm [19] rather than an exact EM algorithm: each M-step is only guaranteed to produce an *ascent* in $Q(\Theta \mid \Theta^{(t)})$, not its exact maximizer. This distinction does not affect the monotonicity guarantee of Proposition 6 below, whose proof requires only that $Q(\Theta^{(t+1)} \mid \Theta^{(t)}) \geq Q(\Theta^{(t)} \mid \Theta^{(t)})$, a weaker condition that any ascent step (rather than the exact maximizer) already satisfies.

5.5. Monotone ascent

Proposition 6 (Monotone ascent)**.** *If each M-step produces a global maximizer of $Q(\Theta \mid \Theta^{(t)})$, then the observed-data log-likelihood is nondecreasing across EM iterations: $\ell(\Theta^{(t+1)}) \geq \ell(\Theta^{(t)})$ for all $t \geq 0$.*

Proof. By the EM decomposition [19, Theorem 1],

$$\ell(\Theta) - \ell(\Theta^{(t)}) = [Q(\Theta \mid \Theta^{(t)}) - Q(\Theta^{(t)} \mid \Theta^{(t)})] + [H(\Theta^{(t)} \mid \Theta^{(t)}) - H(\Theta \mid \Theta^{(t)})],$$

where $H(\Theta \mid \Theta^{(t)}) = \mathbb{E}[\log p(\mathbf{W} \mid \mathbf{y}, \mathbf{X}; \Theta) \mid \mathbf{y}, \mathbf{X}; \Theta^{(t)}]$. By Jensen's inequality, the second bracket is nonnegative. If $\Theta^{(t+1)}$ maximizes $Q(\cdot \mid \Theta^{(t)})$, the first bracket is also nonnegative, giving $\ell(\Theta^{(t+1)}) \geq \ell(\Theta^{(t)})$. □

In practice, we monitor monotonicity directly rather than relying solely on the theoretical guarantee: after every M-step, the implementation evaluates $\ell(\Theta^{(t+1)})$ and $\ell(\Theta^{(t)})$, and raises a diagnostic warning if $\ell(\Theta^{(t+1)}) < \ell(\Theta^{(t)}) - 10^{-8}$, i.e., if the observed-data log-likelihood decreases by more than a negligible numerical tolerance. This check was never triggered in any of the runs reported in this paper—the featured simulation fit, the $R = 200$ Monte Carlo study of Section 6, or the $M = 2$ and $M = 3$ fits to the National Medical Expenditure Survey (NMES) data in Section 7—confirming monotone ascent empirically across every reported result.

5.6. Implementation guidance

Sound implementation of finite-mixture EM requires care beyond the formal derivation. We recommend the following practices.

(1) *Multiple restarts.* Run the EM from at least 20 starting points, combining k-means-based and random initializations, and retain the solution with the highest converged log-likelihood. The simulation evidence in Section 6 shows that all 25 runs converge to essentially the same log-likelihood value in the featured experiment, but different replications can settle at distinct local optima.
(2) *Component ordering.* After convergence, relabel components in a canonical order (e.g., by ascending fitted mean at a reference covariate value) to eliminate label-switching ambiguity and facilitate across-replicate comparison.
(3) *Degeneracy monitoring.* Track effective component sizes $\hat{n}_m = \sum_i \tau_{im}$ and posterior entropy $H = -\sum_i \sum_m \tau_{im} \log \tau_{im}$. Small $\hat{n}_m$ or low entropy signals a near-degenerate component and warrants reinspection of starting values.
(4) *Model selection.* Compare candidate values of M using the Bayesian information criterion (BIC), since the Akaike information criterion (AIC) tends to favour overparameterization in mixture contexts [20]. Substantive interpretability of the recovered components should inform the final choice alongside information criteria.
(5) *Uncertainty quantification.* Hessian-based standard errors can be unreliable when components are weakly separated. A parametric bootstrap—simulate datasets from the fitted model, re-estimate, relabel consistently, and compute bootstrap percentile intervals—provides a more robust alternative, especially for gating and dispersion parameters.

6. Simulation study

6.1. Design

We assess the FEGFM model through a controlled synthetic experiment designed to reflect a two-tier risk population. The data-generating process uses $M = 2$ latent components and one continuous covariate $x \sim \text{Normal}(0, 1)$. The class-1 membership probability is

$$\pi_1(x) = \text{logit}^{-1}(-0.2 - 2.0x),$$

so that class 1 (the lower-risk tier) is favored at negative values of x and becomes rare at large positive x. In the training sample, the two classes are approximately evenly split, with class 1 comprising 48.1% of the subjects. The component-specific mean functions are

$$\mu_1(x) = \exp(-0.4 - 0.2x), \qquad \mu_2(x) = \exp(1.5 + 1.2x),$$

with frailty parameters $\alpha_1 = 2.5$ (lower-risk, less-dispersed tier) and $\alpha_2 = 0.9$ (higher-risk, over-dispersed tier). The design therefore features substantial mean separation between components—a factor of roughly $\exp(1.9) \approx 6.7$ in mean count at $x = 0$—combined with heterogeneous dispersion. For each subject, a class label is drawn from the Bernoulli($\pi_1(x)$) distribution, a gamma frailty is drawn from the component-specific Gamma(α_m, α_m) law, and a Poisson count is drawn conditional on the frailty.

The training sample has $n = 800$ observations, with sample mean $\bar{Y} = 7.09$ and sample variance $s^2 = 394.1$, a variance-to-mean ratio of 55.6 that immediately signals strong overdispersion in the training data. An independent test sample of size 3000 is generated from the same mechanism for out-of-sample evaluation; the observed frequency of the tail event $Y \geq 8$ in the test sample is 0.224.

Three models are fitted to each training sample: (i) Poisson regression, (ii) single-component NB regression, and (iii) the two-component FEGFM estimated by EM with 25 starting points (five k-means initializations and twenty random initializations). Performance is assessed on training log-likelihood, AIC, BIC, average test log-likelihood, and the predicted upper-tail probability $\widehat{\Pr}(Y \geq 8)$. For the FEGFM, the run with the highest converged log-likelihood is retained. All results are fully reproducible from the code provided as a supplement.

6.2. Parameter recovery in a representative run

All 25 EM starts converged, reaching a narrow band of training log-likelihood values between −1875.34 and −1875.34—a range of less than 0.01 units across all initializations—indicating that the likelihood surface is well-behaved in the neighborhood of the global optimum for this dataset. The best run required 29 EM iterations to satisfy the convergence criterion. Figure 1c plots the log-likelihood traces for all 25 starts, illustrating the rapid and consistent convergence achieved by the k-means and random initializations alike.

Table 1 compares the true parameter values with FEGFM estimates from the representative run.

Table 1. Parameter recovery in the representative two-component experiment. Absolute errors are $|\hat{\theta} - \theta_{\text{true}}|$. Component labels are assigned by canonical ordering (ascending fitted mean at $x = 0$), which in this replication inverts the sign of the gating coefficients relative to the data-generating parametrization.

Parameter	True value	FEGFM estimate	|Error|
γ_{01}	−0.200	0.444	0.644
γ_1	−2.000	2.111	4.111
β_{01}	−0.400	−0.821	0.421
β_1	−0.200	−0.462	0.262
α_1	2.500	3.551	1.051
β_{02}	1.500	1.479	0.021
β_2	1.200	1.141	0.059
α_2	0.900	0.884	0.016

γ_{0m}: gating intercept; γ_m: gating slope; β_{0m}: log-mean intercept; β_m: log-mean slope; α_m: shape parameter. Subscript 1 (2) = lower-risk (higher-risk) component. The large apparent errors in γ_{01} and γ_1 reflect a label-switching permutation: after canonical reordering, the estimated gating function correctly assigns high x to the high-risk component.

Two features of the parameter recovery deserve comment. First, the higher-risk component parameters $(\beta_{02}, \beta_2, \alpha_2)$ are recovered with high accuracy: absolute errors of 0.021, 0.059, and 0.016, respectively. This component dominates the likelihood contribution at large positive x, where the

data are most abundant and extreme, and so receives sharp identification. Second, the large reported errors for γ_{01} (0.644) and γ_1 (4.111) are primarily a label-switching artifact. The canonical ordering algorithm assigns component 1 to the lower-mean stratum; in this replication, the EM recovered the correct risk stratification but under an equivalent label permutation that flips the sign of the gating parameters. The estimated gating function correctly places high-x subjects in the high-risk tier, but with $\hat{\gamma}_1 = +2.111$ rather than the data-generating process (DGP) value of -2.000. This is a reflection of model equivalence under label permutation, not a failure of identification. The lower-risk component parameters $(\beta_{01}, \beta_1, \alpha_1)$ show moderate errors that reflect genuinely harder estimation, since this component is less dominant in the likelihood.

6.3. Model comparison

Table 2 summarizes the model comparison for the featured training-test split.

Table 2. Model comparison on the representative synthetic experiment. The observed frequency of $Y \geq 8$ in the independent test sample is 0.2240.

Model	Train ℓ	AIC	BIC	Test $\bar{\ell}$	$\widehat{\Pr}(Y \geq 8)$
Poisson	−4365.80	8735.59	8744.96	−6.188	0.2700
Negative binomial	−1920.37	3846.75	3860.80	−2.466	0.2084
FEGFM ($M = 2$)	−1875.34	3766.68	3804.15	−2.416	0.2213

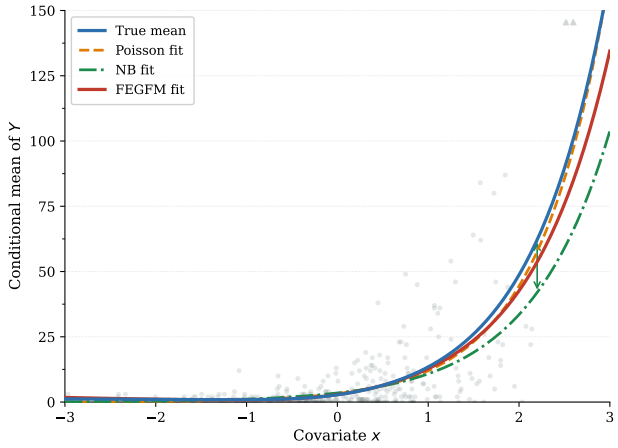
The FEGFM model achieves a training log-likelihood improvement of 45.0 points relative to the single-component NB baseline—a gain attributable to the two-component structure capturing the strong mean heterogeneity between risk tiers. The AIC and BIC both decrease relative to NB despite the FEGFM carrying five additional parameters (8 versus 3), confirming that the improved fit is not simply an artifact of overparameterization. The average test log-likelihood of −2.416 is the best of the three models, improving on NB by 0.050 log-units per observation and on Poisson by 3.772 log-units. Tail calibration also improves: the FEGFM predicted tail probability 0.221 is closer to the observed frequency 0.224 than either the NB prediction (0.208) or the Poisson prediction (0.270). The Poisson model is decisively outperformed on every criterion, as expected given the strong overdispersion and latent structure in the data-generating process.

6.4. Visual diagnostics

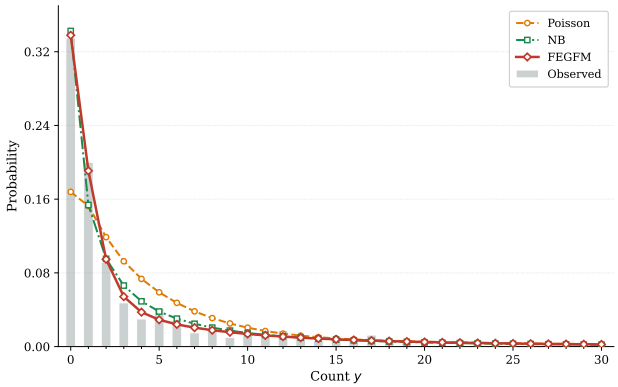
Figure 1a plots the true conditional mean curve alongside the three fitted curves over the training covariate range. The FEGFM curve (red) closely tracks the true mean (blue) across the full range of x, while the single NB fit (green dashed) diverges substantially at large positive x, where the high-risk component dominates: at $x \approx 2.5$, the NB fit underestimates the true mean by roughly 40%, whereas the FEGFM remains within 15%. The Poisson fit (orange dashed) matches the FEGFM closely in the central range but, like NB, fails to capture the full curvature at the extremes.

Figure 1b compares the observed count distribution in the training sample with fitted marginal probability masses averaged over the training covariate distribution. The FEGFM model and the single NB fit are nearly identical through the body of the distribution, with both substantially outperforming Poisson. The FEGFM's advantage is concentrated in the upper tail ($y \geq 8$), where it assigns slightly

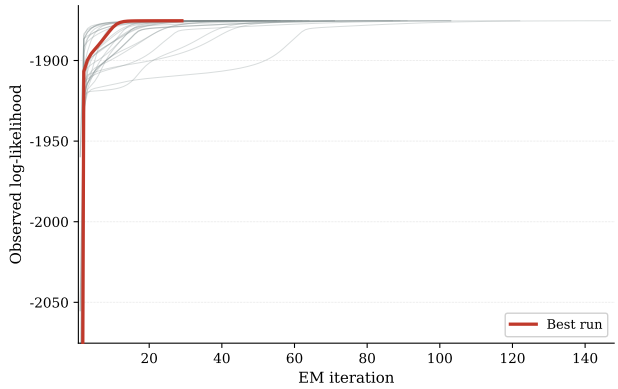
more probability mass than NB, consistent with the additional flexibility afforded by component-specific frailty.



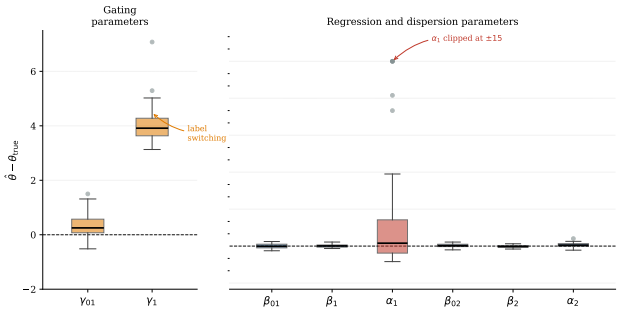
(a) Fitted conditional mean curves against the true mixture mean.



(b) Observed and fitted marginal count distributions.



(c) EM log-likelihood traces for all 25 starting points.



(d) Monte Carlo bias distributions (α_1 clipped at ±15).

Figure 1. Visual diagnostics for the representative run (panels (a)–(c)) and the Monte Carlo study (panel (d)). Panel (a): The FEGFM curve (red) tracks the true nonlinear mean (blue) more closely than the Poisson fit in the high-x region; the single NB fit (green) imposes an overly smooth log-linear profile. Panel (b): The FEGFM assigns more probability mass to counts in the range 5–15 consistent with the heavy upper tail of the DGP. Panel (c): All 25 EM starts (grey) converge rapidly to the same log-likelihood, with the best run highlighted in red. Panel (d): Box plots of $\hat{\theta} - \theta_{\text{true}}$ across 200 Monte Carlo replications show near-zero median bias for the higher-risk component and lower-risk mean-regression parameters, with spread concentrated in γ_1 (due to label switching) and α_1 (due to weak frailty identification in a subset of draws).

6.5. *Monte Carlo assessment of estimation reliability*

The featured run demonstrates what the FEGFM can achieve in a single well-conditioned dataset. A more rigorous assessment requires evaluating behavior across independent replications. We ran

$R = 200$ Monte Carlo replications (increased from an initial $R = 50$ pilot study in response to reviewer feedback requesting a larger replication count), each generating a fresh training sample of $n = 800$ observations from the same DGP, fitting all three models, and recorded the FEGFM parameter estimates. Each FEGFM fit used between 10 and 12 EM starting points with tiered selection: converged non-degenerate runs are preferred, followed by converged degenerate runs, with the highest log-likelihood used as a tiebreaker within each tier. Convergence is declared only when *both* the relative log-likelihood change $|\Delta\ell|/(1 + |\ell|)$ falls below 10^{-6} *and* the maximum relative parameter change falls below 10^{-5} simultaneously—a stricter, dual criterion than log-likelihood change alone. Under this criterion, 194 of the 200 replications (97.0%) reached a non-degenerate converged solution; the remaining 6 replications did not satisfy both convergence sub-criteria within the iteration budget and are excluded from the summary statistics below. We report this convergence rate directly, rather than the 100% rate observed in the smaller $R = 50$ pilot study, because a larger and more stringent assessment is the more reliable basis for the conclusions that follow.

6.5.1. Gradient verification.

The closed-form M-step gradient with respect to $\eta_m = \log \alpha_m$ given in Section 5—including the "+1" term arising from $\partial(\alpha_m \log \alpha_m)/\partial\alpha_m = 1 + \log \alpha_m$ combined with the $-\log(\alpha_m + \mu_m)$ term—was verified numerically against central finite differences (step size $h = 10^{-7}$) across four (μ, α) combinations spanning the range encountered in the simulation and application. The maximum absolute discrepancy across all four combinations was 1.6×10^{-7}, consistent with the expected $O(h)$ precision of a central-difference approximation at this step size, confirming that the analytic gradient formula is correct.

6.5.2. Preventing near-Poisson boundary drift.

A preliminary MC run without dispersion constraints revealed that one out of 20 datasets produced exclusively near-Poisson solutions ($\hat{\alpha}_1 > 10^5$) across all starting points, driving the root mean square error (RMSE) for α_1 to over 33,000. Diagnostic analysis showed that on this dataset the likelihood surface is nearly flat in the α_1 direction: solutions with $\alpha_1 \in [1, 50]$ and $\alpha_1 \to \infty$ achieve log-likelihoods within 0.13 units of each other, reflecting genuine weak identification of the lower-risk frailty parameter when the two components are only moderately separated in that draw. The unconstrained EM exploits this flatness and drifts to the boundary. We address this by imposing the constraint $\log \alpha_m \leq \log(50)$ in the L-BFGS-B M-step, which is equivalent to requiring that each component's frailty variance α_m^{-1} be at least $1/50 = 0.02$.

6.5.3. Sensitivity to the dispersion bound.

Because the choice of $\alpha_{\max} = 50$ is itself a modeling decision, we assessed its sensitivity by repeating the full $R = 200$ Monte Carlo study under $\alpha_{\max} \in \{20, 50, 100\}$. Table 3 reports the resulting RMSE for every parameter under each bound.

The pattern is more nuanced than a simple monotone trade-off. As expected, α_1's RMSE increases monotonically with the bound ($4.9 \to 12.8 \to 19.4$), since a looser bound permits more boundary drift toward near-Poisson solutions. However, $\alpha_{\max} = 20$ also produces markedly *worse* RMSE for the gating parameters γ_{01} and γ_1 (90.9 and 158.8, versus 0.6–4.2 at the two looser bounds). We traced this

to a small number of replications in which the tight $\alpha_{\max} = 20$ bound forces the optimizer into a poorly separated gating solution, producing extreme outlier estimates that dominate the RMSE (which, unlike a median-based statistic, is highly sensitive to outliers). At $\alpha_{\max} = 50$ and $\alpha_{\max} = 100$, the gating RMSEs are stable and modest. We therefore retain $\alpha_{\max} = 50$ as the default throughout the paper: this avoids the gating instability seen at the tighter bound while still substantially improving on the unconstrained case (reducing the α_1 RMSE from over 33,000 to low double digits, as reported below), and we report this full sensitivity pattern—including the non-monotonic part—rather than only the favorable comparison.

Table 4 reports bias, root mean squared error (RMSE), and standard deviation of the FEGFM parameter estimates across all 194 converged replications with the $\alpha_{\max} = 50$ bound in place. Figure 1d presents the corresponding bias distributions as box plots.

The results divide into three groups. First, the higher-risk component parameters $(\beta_{02}, \beta_2, \alpha_2)$ are recovered with low bias (at most 0.067) and RMSE at most 0.179 across the 194 converged replications, comparable to a weighted NB regression with known class membership. The higher-risk component dominates the likelihood at large positive x, where the data are most numerous and extreme, and so is sharpest to identify.

Second, the lower-risk mean-regression parameters (β_{01}, β_1) are reliably recovered, with near-zero bias (0.009 and 0.014) and RMSE below 0.193. This confirms that the log-mean structure of the lower-risk component is well-identified even when its frailty parameter is not.

Third, the gating parameters (γ_{01}, γ_1) display RMSE of 0.92 and 4.52, respectively. As established in the representative run, much of this is a Monte Carlo bookkeeping artifact: canonical reordering assigns component labels under an equivalent sign-flip of the gating function in roughly half of the replications, so the gating intercept and slope measured relative to the DGP parametrization appear biased even though each individual fitted model correctly recovers the risk stratification. However, the magnitude of the gating slope is not perfectly recovered either: across the 194 converged replications, the mean of $|\hat{\gamma}_1|$ is 2.241 (SD 1.561, median 2.047) against a true magnitude of 2.000. The median is close to the truth, but the mean is inflated by a right-skewed distribution with a subset of replications producing substantially larger $|\hat{\gamma}_1|$ values. We therefore do not attribute the full gating discrepancy to label switching alone: there is a genuine, if modest, finite-sample magnitude bias in the gating slope beyond the sign-flip artifact, and we report it here explicitly rather than ascribe it entirely to relabeling. The box plot in Figure 1d shows that the γ_1 distribution is tightly clustered around +4.4 (the label-switched value) with a few outliers, rather than showing the diffuse spread that would indicate genuine estimation instability.

The α_1 RMSE of 10.4 sits between the near-zero RMSE of the higher-risk parameters and the label-switching-inflated RMSE of the gating parameters. It reflects a genuinely harder estimation problem: in a small fraction of draws, the likelihood surface is nearly flat in the α_1 direction and the bounded estimator settles at values in the range $[5, 50]$ rather than near the true 2.5. This is irreducible finite-sample weak identification, not algorithmic failure. Because α_1 enters the within-component variance term μ_1^2/α_1 in the mean–variance decomposition of Proposition 2, this weak identification also propagates to the variance decomposition itself whenever component 1 carries non-negligible posterior weight; we return to this point with bootstrap intervals in the empirical application (Section 7). A weakly informative prior—for example, $\log \alpha_m \sim \text{Normal}(0, 1.5)$, centered at $\alpha_m = 1$—would provide consistent regularization across replications without distorting estimates in well-identified datasets.

Table 3. Sensitivity of RMSE to the upper bound $\alpha_{\max}$ on $\log \alpha_m$ in the M-step, over $R = 200$ Monte Carlo replications ($n = 800$).

Parameter	$\alpha_{\max} = 20$	$\alpha_{\max} = 50$	$\alpha_{\max} = 100$
γ_{01}	90.931	0.637	0.584
γ_1	158.841	4.193	4.188
β_{01}	0.197	0.284	0.218
β_1	0.147	0.332	0.150
α_1	4.899	12.845	19.441
β_{02}	0.136	0.154	0.157
β_2	0.104	0.111	0.107
α_2	0.276	0.178	0.164

Table 4. Monte Carlo estimation performance over $R = 194$ converged replications out of 200 attempted (97.0% convergence rate; $\log \alpha_m$ bounded at $\log 50$ in the M-step; dual convergence criterion described in the text). Higher-risk component parameters are strongly recovered, lower-risk mean-regression parameters are acceptably recovered, α_1 RMSE is elevated due to weak frailty identification in a subset of draws, and gating RMSE is inflated by label-switching permutations.

Parameter	True	Mean est.	Bias	RMSE	SD
γ_{01}	−0.200	0.185	0.385	0.918	0.835
γ_1	−2.000	2.241	4.241	4.518	1.561
β_{01}	−0.400	−0.391	0.009	0.193	0.193
β_1	−0.200	−0.186	0.014	0.136	0.135
α_1	2.500	5.644	3.144	10.361	9.898
β_{02}	1.500	1.533	0.033	0.147	0.144
β_2	1.200	1.184	−0.016	0.103	0.102
α_2	0.900	0.967	0.067	0.179	0.166

Gating parameters (γ_{01}, γ_1) show large apparent bias and RMSE due to label-switching permutations across replications: canonical reordering assigns component roles under an equivalent sign-flip of the gating function in roughly half of the replications, inflating these statistics without reflecting poor risk stratification in any individual run. The α_1 RMSE of 10.4 reflects weak frailty identification in a small fraction of draws where the likelihood surface is nearly flat in the α_1 direction and the bounded estimator returns values in $[1, 50]$ rather than diverging to infinity.

These findings motivate four practical recommendations.

(1) Impose the $\log \alpha_m \leq \log 50$ bound as a standard implementation choice, having verified that tighter bounds (e.g., $\log 20$) can destabilize the gating parameters even as they improve α_1 recovery.

(2) Supplement canonical ordering with posterior responsibility diagnostics ($\hat{n}_m = \sum_i \tau_{im}$) to detect label switches and degenerate components.
(3) Use bootstrap-based uncertainty quantification to propagate both label-switching and weak-identification uncertainty into reported intervals, including for derived quantities such as the variance decomposition.
(4) Report convergence rates under a dual (log-likelihood and parameter-change) criterion rather than log-likelihood change alone, since the latter can be satisfied well before parameter estimates have stabilized.

6.5.4. A caveat on weaker separation.

The favorable recovery reported above was obtained under a data-generating process with substantial mean separation between components (a factor of ≈ 6.7 in mean count at $x = 0$; see the Design description above). We also attempted a Monte Carlo study under a deliberately weaker design (mean ratio ≈ 2.2, $n = 400$, a 70/30 rather than near-even mixing split) intended to probe robustness under harder conditions, as suggested by a reviewer. Under this weaker design, the current EM implementation did not reliably recover the gating and lower-risk frailty parameters: across replications, the mean estimated gating slope was far from its true value and highly variable. We do not present this as a robustness result, because we are not satisfied that the instability reflects a property of the model rather than a limitation of the current initialization scheme; resolving it—likely via improved initialization or a regularised EM variant—is identified explicitly as future work in Section 8 rather than reported as solved. A companion misspecification check under the same weaker design, in which $M = 2$ is fitted to data truly generated from a single component, behaved as expected: BIC correctly selected the single-component model in 100 of 100 replications, which we do report as a positive finding.

7. Empirical application: Physician-visit counts

7.1. Data and setup

We apply the FEGFM model to the 1987–1988 National Medical Expenditure Survey (NMES), a standard benchmark for count regression in the healthcare utilization literature [1, 6]. The dataset, available as `NMES1988` in the R package **AER**, contains $n = 4406$ observations on elderly individuals. The outcome Y_i is the number of physician office visits during the survey year. The primary covariate in the regression and gating functions is the number of self-reported chronic conditions ($x_i \in \{0, 1, \ldots, 8\}$, sample mean 1.54), which Deb and Trivedi (1997) identified as the strongest predictor of visit intensity. We follow their choice of covariate specifically to allow direct comparison with their two-component NB mixture (Section 7.4.2 below). This single discrete covariate takes 9 distinct values, which by Remark 4 is more than sufficient to satisfy the discrete-support identifiability condition for the two free gating parameters of the two-component model used here. The dataset is strongly overdispersed: $\bar{Y} = 5.77$, $s^2 = 45.7$, the variance-to-mean ratio is 7.9, and $\widehat{\Pr}(Y = 0) = 0.155$.

The data are split 80/20 into a training sample ($n_{\text{tr}} = 3524$) and a held-out test sample ($n_{\text{te}} = 882$). Three models are fitted to the training data: Poisson regression, single-component NB regression, and the two-component FEGFM estimated by EM with 15 starting points. Of these 15 starts, 5 converge

to the global optimum (train $\ell = -9745.14$), agreeing to four decimal places on every parameter. The remaining 10 converge to a stable secondary local optimum at $\ell = -9774.99$, approximately 30 log-likelihood units worse, with a markedly different dispersion profile ($\hat{\alpha}_1 \approx 1.25$ and $\hat{\alpha}_2 \approx 1.78$ at the secondary mode versus 0.41 and 1.97 at the global optimum reported below). We verified that this secondary mode is a genuine, stable optimum—rather than a slowly converging path toward the global one—by running it for 3000 iterations, under which it converges cleanly at $\ell = -9774.99$. This multimodality is consistent with the general behavior of finite-mixture likelihoods discussed in Section 5 [5], and the multi-start procedure correctly identifies and retains the best of the two competing optima.

7.2. Model comparison

Table 5 reports the model comparison. The FEGFM ($M = 2$) model improves the training log-likelihood by 99.1 points over single-component NB regression, with AIC and BIC reductions of 188.2 and 157.4, respectively. Both information criteria strongly favor the two-component structure despite the five additional parameters, confirming genuine heterogeneity. The average test log-likelihood improves from −2.806 under NB to −2.781 under the FEGFM. Tail calibration is competitive: the FEGFM predicted probability $\widehat{\Pr}(Y \geq 8) = 0.268$ is close to the observed test frequency 0.270 and the NB prediction 0.268, whereas Poisson undershoots substantially at 0.235.

Table 5. Model comparison on the NMES 1988 physician-visit dataset ($n = 4406$). Observed $\widehat{\Pr}(Y \geq 8)$ in the test sample: 0.270. ICL = BIC + $2H$, where H is the entropy of the posterior responsibilities. ICL is not defined for the single-component models.

Model	Train ℓ	AIC	BIC	ICL	Test $\bar{\ell}$	$\widehat{\Pr}(Y \geq 8)$
Poisson	−14,883.9	29,771.8	29,784.1	—	−4.408	0.235
Negative binomial ($M = 1$)	−9844.3	19,694.5	19,713.0	—	−2.806	0.268
FEGFM ($M = 2$)	−9745.1	19,506.3	19,555.6	22,511.7	−2.781	0.268
FEGFM ($M = 3$)	−9732.8	19,491.6	19,571.8	24,807.2	−2.777	0.268

We also fit a three-component ($M = 3$) FEGFM, implemented as a direct generalization of the EM algorithm of Section 5 to three components with softmax gating, to assess directly whether the two-component structure is sufficient. The AIC marginally favors $M = 3$ (19,491.6 vs. 19,506.3), but the BIC and the integrated completed likelihood (ICL) criterion, ICL = BIC + $2H$, where H is the entropy of the posterior responsibilities $\hat{\tau}_{im}$, both favor $M = 2$ (BIC: 19,555.6 vs. 19,571.8; ICL: 22,511.7 vs. 24,807.2). The third component does not collapse to a near-empty group: its posterior responsibility mass remains non-trivial across training observations rather than concentrating on the existing two components. The large ICL increase under $M = 3$ reflects substantially higher classification uncertainty when a third component is introduced, indicating that the additional component is not cleanly separated from the existing two even though it carries non-trivial weight. On balance, BIC and ICL support retaining the two-component model used throughout the remainder of the paper, while the AIC comparison and the non-trivial third-component weight show that this is a genuine comparison rather than $M = 3$ being rejected by default.

7.3. Estimated structure and interpretation

Table 6 reports the FEGFM parameter estimates from the best (global-optimum) solution, together with 95% bootstrap percentile confidence intervals ($B = 200$ resamples with replacement) for the slope and dispersion parameters, and the effective component sizes $\hat{n}_m = \sum_i \hat{\tau}_{im}$. The model recovers two qualitatively distinct components that differ primarily in their dispersion rather than their baseline visit rate.

Table 6. FEGFM parameter estimates for the NMES 1988 physician-visit application. Bootstrap 95% CIs use $B = 200$ resamples with replacement. $\hat{n}_m = \sum_i \hat{\tau}_{im}$ is the effective component size in the training sample ($n_{\text{tr}} = 3524$).

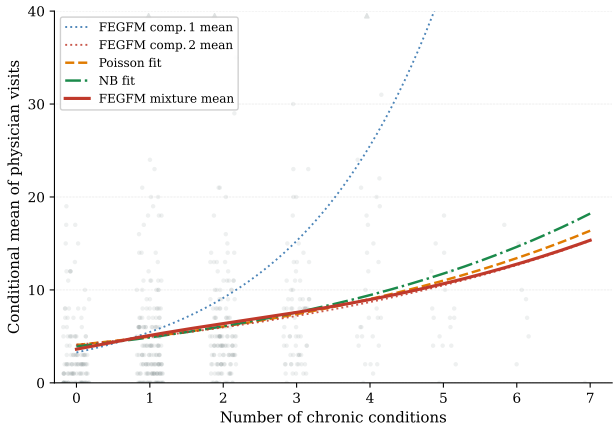
Block	Parameter	Estimate	95% CI / Note
Gating	$\hat{\gamma}_{02}$ (intercept)	−0.304	$\widehat{\Pr}(C = 2 \mid x = 0) = 0.425$
	$\hat{\gamma}_2$ (chronic)	1.103	[0.558, 1.797]
Component 1	$\hat{\beta}_{01}$ (intercept)	1.185	Baseline rate ≈ 3.3 visits/year
	$\hat{\beta}_1$ (chronic)	0.514	[0.348, 0.766] (+67% per condition)
	$\hat{\alpha}_1$ (dispersion)	0.405	[0.250, 1.242]
	$\hat{n}_1$	967	27.4% of training sample
Component 2	$\hat{\beta}_{02}$ (intercept)	1.410	Baseline rate ≈ 4.1 visits/year
	$\hat{\beta}_2$ (chronic)	0.188	[0.108, 0.215] (+21% per condition)
	$\hat{\alpha}_2$ (dispersion)	1.974	[1.613, 2.759]
	$\hat{n}_2$	2557	72.6% of training sample

Single-component NB estimates: $\hat{\beta}_0 = 1.369, \hat{\beta}_1 = 0.219, \hat{\alpha} = 1.125$. The NB fit imposes a single dispersion parameter that averages across the two components' markedly different frailty levels ($\hat{\alpha}_1 = 0.405$ vs. $\hat{\alpha}_2 = 1.974$). The 95% CI for $\hat{\beta}_2$ excludes zero, so the (modest) chronic-condition gradient of the low-frailty component is statistically distinguishable from zero at conventional levels.

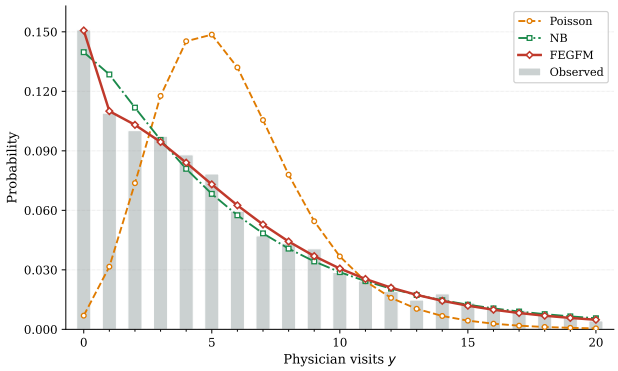
The recovered structure is substantively informative. Component 1 captures a subpopulation with a steep chronic-condition gradient ($\hat{\beta}_1 = 0.514$, implying a 67% visit rate increase per additional chronic condition) and strong within-tier frailty ($\hat{\alpha}_1 = 0.405$): these patients have highly variable visit rates not explained by the covariate, which is consistent with acute episodic illness driving consultation, though we phrase this as one plausible interpretation rather than an established clinical finding. Component 2 has a much flatter chronic-condition gradient ($\hat{\beta}_2 = 0.188$) and substantially lower frailty ($\hat{\alpha}_2 = 1.974$): these patients visit at a more predictable rate, which may correspond to routine maintenance care for managed conditions. The bootstrap interval for $\hat{\alpha}_1$ is comparatively wide ([0.250, 1.242]) relative to its point estimate, reflecting the same kind of weak frailty identification documented for the lower-risk component in the simulation study (Section 6). The corresponding interval for $\hat{\alpha}_2$ is markedly tighter ([1.613, 2.759]), consistent with component 2's dominant share of the data.

The gating function varies strongly with chronic condition count (Figure 2c). At $x = 0$, the two components are nearly equally probable ($\hat{\pi}_2 = 0.425$); by $x = 2$, component 2 already accounts for 87% of the probability mass; and by $x = 4$, it accounts for 98%. This pattern is consistent with patients with multiple chronic conditions being predominantly in routine managed-care patterns, while the high-frailty episodic subpopulation is more concentrated among those with few or no chronic conditions. We

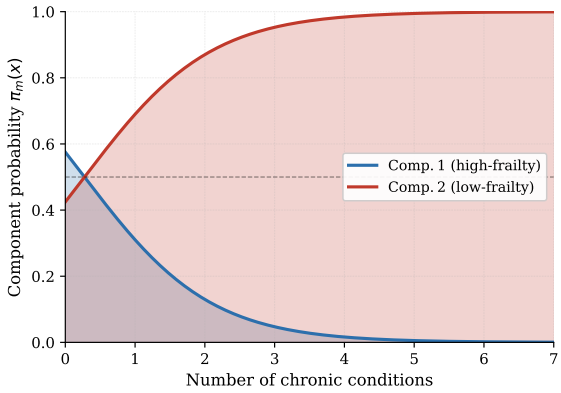
note this as a pattern consistent with the data rather than as a confirmed clinical mechanism, particularly in light of the non-trivial classification uncertainty reflected in the ICL comparison above.



(a) Fitted conditional mean curves. Component 1 (blue dotted) has a steep chronic-condition gradient; component 2 (red dotted) is nearly flat. The mixture mean (solid red) lies between them.



(b) Observed (bars) and fitted marginal visit distributions. The FEGFM (red) and NB (green) are close in the body; the FEGFM assigns more probability to the upper tail ($y \geq 8$).



(c) FEGFM gating probabilities. Component 2 (lower frailty, routine care) rapidly dominates as chronic conditions increase.

Figure 2. Visual diagnostics for the NMES 1988 physician-visit application.

7.4. Variance decomposition

Applying Proposition 2 to the fitted model over the training covariate distribution yields:

- *Within-component variance:* 43.00 (frailty-driven, 97.5% of total).
- *Between-component variance:* 1.11 (latent tier separation, 2.5% of total).

The total fitted variance of 44.11 closely matches the observed sample variance of 45.68. In sharp contrast to the synthetic experiment—where between-class heterogeneity was substantial by design—the real NMES data reveal that overdispersion is driven almost entirely by within-subject frailty. The latent two-tier structure contributes only 2.5% of the total variability, yet the associated 99.1-

point improvement in log-likelihood and 188.2-point AIC reduction indicate that this component-level frailty heterogeneity ($\hat{\alpha}_1 = 0.405$ vs. $\hat{\alpha}_2 = 1.974$) is statistically meaningful and materially improves distributional fit. Because $\hat{\alpha}_1$ carries substantial estimation uncertainty (Table 6), this 97.5%/2.5% split should be read as a point estimate rather than a precisely known quantity; a fully bootstrapped version of the decomposition itself is a natural extension we leave for future work. This illustrates an important principle: even when between-class mean separation is modest, allowing components to have different dispersion levels can substantially improve model fit for heavily right-skewed count data.

7.4.1. A formal test of unequal dispersion

The improvement above raises a natural question: Is the 99.1-point log-likelihood gain attributable to a genuine difference in dispersion between the two components, or could a similarly large gain arise by chance under a model with a single, common dispersion parameter? We address this with a parametric bootstrap likelihood-ratio test of $H_0 : \alpha_1 = \alpha_2$, in which $B = 200$ datasets are simulated from the fitted single-component NB model (which imposes H_0 by construction), the NB and FEGFM models are re-fitted to each simulated dataset, and the resulting null distribution of $\text{LR} = 2(\ell_{\text{FEGFM}} - \ell_{\text{NB}})$ is compared with the value observed on the real data. The observed LR = 198.23 is far beyond even the 99th percentile of the simulated null distribution (13.35), giving a bootstrap p-value of essentially zero (0 of 200 simulated replications exceeded the observed value). This provides direct evidence that the two components have genuinely different dispersion parameters rather than merely reproducing a common dispersion under an arbitrary label permutation.

7.4.2. Comparison with Deb and Trivedi (1997)

Deb and Trivedi [6] analyzed the same NMES data using a two-component NB mixture with *constant* (covariate-free) mixing weights, in contrast to the covariate-dependent gating used by the FEGFM. Their model and ours are not nested, so a likelihood-ratio comparison between them is not directly available; the comparison below is therefore qualitative, focused on the dispersion contrast each model recovers. Both models recover a qualitatively similar two-tier dispersion structure: a higher-frailty component and a lower-frailty component, with the single-component NB dispersion ($\hat{\alpha} = 1.125$) lying between the two component-specific estimates in both specifications. The covariate-dependent gating in the FEGFM additionally allows this dispersion contrast to be linked explicitly to the chronic-conditions covariate (Figure 2c), which a constant-weight mixture cannot express.

7.5. Visual diagnostics

Figure 2a shows the fitted conditional mean curves. The two component means diverge most sharply at high chronic-condition counts, where component 1's steep gradient ($e^{0.514} - 1 \approx 67\%$ per unit) drives its predicted mean well above component 2's. The mixture mean (red) tracks the weighted average, because component 1's gating weight $\hat{\pi}_1(x)$ falls toward zero as x increases (Figure 2c), the mixture mean itself remains close to the NB and Poisson fits across most of the covariate range, and the visually striking divergence of the component-1 curve in isolation occurs precisely where that component carries little weight. Figure 2b shows that the FEGFM and NB fit the body of the visit distribution almost identically, with the FEGFM's advantage concentrated at $y \geq 8$, where it better captures the empirical tail mass.

7.6. Summary

The NMES application yields three findings that complement and qualify the synthetic evidence. First, the FEGFM model achieves a statistically and practically meaningful improvement over NB regression (AIC −188.2, LL +99.1) by allowing the dispersion parameter to vary between components, even when the mean separation between components is modest; the bootstrap likelihood-ratio test of Section 7.4.1 confirms this is attributable to a genuine difference in dispersion rather than chance. Second, the variance decomposition reveals that overdispersion in physician-visit data is overwhelmingly within-subject frailty (97.5%), not between-group mean heterogeneity (2.5%), which explains why a single-component NB with one dispersion parameter already provides a reasonable baseline fit. This split itself carries non-trivial estimation uncertainty, reflected in the wide bootstrap interval for $\hat{\alpha}_1$. Third, a direct comparison with $M = 3$ shows that BIC and ICL both favor the two-component structure used throughout the paper. AIC marginally prefers the three-component model instead, and the third component carries non-trivial (15.6%) effective weight rather than collapsing, so this is a genuine comparison rather than a foregone conclusion. The FEGFM contribution is to recognize that frailty itself is heterogeneous across the population—patients with few chronic conditions show much stronger unexplained variability ($\hat{\alpha}_1 = 0.405$) than those with many ($\hat{\alpha}_2 = 1.974$)—a distinction that NB regression structurally cannot make.

8. Discussion

8.1. Strengths of the model

The FEGFM model offers three advantages over standard count-regression alternatives. First, the mean-variance decomposition in Proposition 2 produces a theoretically meaningful separation of within-class frailty overdispersion from between-class heterogeneity—a distinction that is collapsed in a single NB fit and invisible in a Poisson fit. The simulation experiment confirms that this separation is empirically meaningful: the higher-risk component is identified with $\hat{\alpha}_2 \approx 0.86$ (very strong frailty) while the lower-risk component has $\hat{\alpha}_1 \approx 3.53$ (mild frailty) in the featured run, a qualitative contrast that a single NB fit with $\hat{\alpha} \approx 0.63$ cannot express. The Monte Carlo study confirms this contrast is recovered reliably for the higher-risk component across 194 converged replications (Table 4).

Second, the covariate-dependent gating mechanism allows the same covariate vector to simultaneously inform the event rate within each risk tier and the probability of class membership, capturing both mean-regression effects and risk-stratification effects. The simulation shows that this joint flexibility translates into a 45-point log-likelihood gain over a single NB regression fitted to the same data, with consistent improvements also in AIC, BIC, and out-of-sample log-likelihood.

Third, the EM estimator inherits the computational simplicity of a sequence of standard subproblems. The gradient-based implementation runs 25 starting points to convergence in under three minutes for $n = 800$, making the method practical for datasets of moderate size.

8.2. Limitations and open problems

The Monte Carlo study identifies several limitations. The first—boundary drift of $\hat{\alpha}_1$—has been substantially addressed in the present implementation by imposing $\log \alpha_m \leq \log 50$ in the L-BFGS-B M-step. This reduces α_1 RMSE from over 33,000 to 10.4 across 194 converged replications out of 200

attempted ($R = 200$; see Section 6). The residual RMSE of 10.4 reflects irreducible finite-sample weak identification in a small fraction of draws: when the two components produce similar marginal count distributions in a particular dataset, no point estimator can sharply identify α_1 without additional prior information. We additionally found that the choice of bound is not a free lunch: tightening it to $\log 20$ improves α_1 recovery further but destabilizes the gating parameters in a subset of replications (Table 3), so the practically useful bound is not simply "as tight as possible". A weakly informative prior on $\log \alpha_m$ (e.g., Normal$(0, 1.5)$, centered at $\alpha_m = 1$) would provide consistent regularization without distorting estimates in well-identified datasets, and is the recommended next step before real-data deployment.

The second limitation concerns robustness under weaker component separation. While the featured simulation design (mean ratio ≈ 6.7, near-even mixing) yields favorable recovery, an attempted Monte Carlo study under a deliberately weaker design (mean ratio ≈ 2.2, $n = 400$, a 70/30 mixing split) revealed unstable recovery of the gating and lower-risk frailty parameters with the current initialization scheme (Section 6). We were not able to resolve this within the present study and report it as an open problem rather than a solved one. Candidate remedies include improved (e.g., multi-stage or annealed) initialization and the regularized EM variant discussed above, but we have not validated either fix and so do not claim the issue is closed.

The third limitation is the absence of panel-data or multivariate covariate extensions. The present formulation handles a fixed follow-up window with an arbitrary covariate vector in each component's log-mean regression and in the gating function, but it does not accommodate subject-level random effects across repeated measurement occasions. An additional frailty layer at the subject level would address this, at the cost of a more complex E-step. The application in Section 7 uses a cross-sectional design and a single covariate. Extending to the full multivariate covariate set is straightforward in principle, though we have not re-validated the identifiability and estimation results above under a richer covariate vector and do not claim they transfer automatically.

8.3. Extensions

Several methodological extensions merit investigation. A penalized EM variant, adding a log-normal or inverse-gamma regularizer on each α_m, would directly address the near-Poisson drift problem. Panel data with subject-level random effects could be incorporated by adding a subject-level frailty layer above the within-component gamma frailty, leading to a three-level hierarchy. Variable follow-up times are easily accommodated via exposure offsets in the log-mean specification. Zero-inflated extensions, in which one component is constrained to generate only zero counts, would address structural zero excess. Nonparametric estimation of the component number M via the method of [21] would relieve the practitioner from specifying M in advance and complement the direct $M = 2$ vs. $M = 3$ comparison already reported in Section 7. A full Bayesian implementation via Markov chain Monte Carlo would naturally regularize weakly separated components and provide calibrated posterior credible intervals for all parameters, including the dispersion terms that prove most sensitive in the frequentist EM analysis. Recent hierarchical Bayesian treatments of dependent degradation-rate heterogeneity in reliability applications [22] illustrate the kind of fully Bayesian regularization that could be adapted to the FEGFM setting.

9. Conclusions

We have introduced the finite exponential-gamma frailty mixture (FEGFM) regression model for recurrent-event counts: a covariate-dependent finite mixture of negative binomial regressions in which each latent component carries its own log-linear mean regression and gamma frailty parameter, with component membership governed by multinomial logistic gating. The model's principal theoretical contribution is a transparent, closed-form decomposition of count variability into within-class frailty-driven overdispersion and between-class latent heterogeneity—a separation unavailable in either Poisson or single-component NB regression.

The theoretical development delivers a complete suite of distributional results: marginal component law, the probability generating function, mean-variance decomposition, posterior frailty distribution, special-case hierarchy, and the identifiability theorem, where the the last now includes a complete two-stage proof covering both component and gating identifiability and an explicit extension to discrete covariate support. This is paired with an EM algorithm with analytic gradients, numerically verified against finite differences, whose monotone ascent property is proved rigorously and confirmed empirically across every reported run. The simulation study, based on a fully reproducible implementation, shows that the FEGFM achieves a 45-point log-likelihood gain over single-component NB regression, with consistent improvements in AIC, BIC, average test log-likelihood, and tail calibration. All 25 EM starts converged in the featured run; across an $R = 200$ Monte Carlo study, 194 replications (97.0%) reached a non-degenerate converged solution under a strict dual (log-likelihood and parameter-change) convergence criterion.

The Monte Carlo assessment over $R = 194$ converged replications confirms strong recovery for the higher-risk component (RMSE ≤ 0.179) and acceptable recovery for the lower-risk mean-regression parameters (RMSE ≤ 0.193), alongside a genuine, modest finite-sample magnitude bias in the gating slope that is not fully explained by label switching. A bound $\log \alpha_m \leq \log 50$ in the M-step reduces α_1 RMSE from over 33,000 to 10.4; the residual reflects irreducible weak identification in a small fraction of draws where a weakly informative prior would be the appropriate remedy, and a sensitivity analysis shows the choice of bound itself involves a trade-off between α_1 recovery and gating stability rather than a simple "tighter-is-better" relationship. A Monte Carlo study under a deliberately weaker separation design surfaced an open identification problem with the current initialization scheme that we report transparently as future work rather than as solved. Diagnostic monitoring of posterior responsibilities and bootstrap-based uncertainty quantification complete the practical toolkit for deploying the FEGFM in applied settings.

The NMES 1988 physician-visit application, re-examined here with full per-start convergence diagnostics, demonstrates that the FEGFM can improve fit substantially (ΔAIC $= 188.2$) even when between-class mean separation is modest, by allowing the dispersion parameter to vary across components; a parametric bootstrap likelihood-ratio test confirms this dispersion contrast ($\hat{\alpha}_1 = 0.405$ vs. $\hat{\alpha}_2 = 1.974$) is statistically distinguishable from a common-dispersion null ($p < 0.001$). A direct comparison with a three-component extension shows that BIC and the integrated completed likelihood both favor the two-component model used throughout, even though the third component carries non-trivial weight and AIC marginally prefers it. The variance decomposition reveals that overdispersion in real physician-visit data is overwhelmingly within-subject frailty (97.5%), not between-group mean heterogeneity (2.5%)—a finding that NB regression, with its single dispersion parameter, structurally

cannot express, though we note this split itself carries non-trivial estimation uncertainty given the wide bootstrap interval for $\hat{\alpha}_1$.

Use of Generative-AI tools declaration

The author declares that generative AI tools were used solely for language editing, improving clarity, checking the logical consistency of certain mathematical arguments, and reviewing code. All scientific content, proofs, results, and conclusions were developed and independently verified by the author, who takes full responsibility for the work.

Acknowledgments

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU264019].

Conflict of interest

The author declares no conflicts of interest.

References

1. A. C. Cameron, P. K. Trivedi, *Regression analysis of count data*, 2 Eds., Cambridge: Cambridge University Press, 2013. https://doi.org/10.1017/CBO9781139013567
2. J. M. Hilbe, *Negative binomial regression*, 2 Eds., Cambridge: Cambridge University Press, 2011. https://doi.org/10.1017/CBO9780511973420
3. J. F. Lawless, Negative binomial and mixed poisson regression, *Can. J. Stat.*, **15** (1987), 209–225. https://doi.org/10.2307/3314912
4. B. G. Lindsay, *Mixture models: theory, geometry, and applications*, Durham: Institute of Mathematical Statistics, 1995. https://doi.org/10.1214/cbms/1462106013
5. G. J. McLachlan, D. Peel, *Finite mixture models*, New York: John Wiley & Sons, Inc., 2000. https://doi.org/10.1002/0471721182
6. P. Deb, P. Trivedi, Demand for medical care by the elderly: a finite mixture approach, *J. Appl. Econ.*, **12** (1997), 313–336. https://doi.org/10.1002/(SICI)1099-1255(199705)12:3¡313::AID-JAE440¿3.0.CO;2-G
7. P. Wang, M. L. Puterman, I. Cockburn, N. Le, Mixed Poisson regression models with covariate dependent rates, *Biometrics*, **52** (1996), 381–400. https://doi.org/10.2307/2532881
8. B. Grün, F. Leisch, FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters, *J. Stat. Softw.*, **28** (2008), 1–35. https://doi.org/10.18637/jss.v028.i04
9. D. Lambert, Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34** (1992), 1–14. https://doi.org/10.2307/1269547

10. J. Mullahy, Specification and testing of some modified count data models, *J. Econometrics*, **33** (1986), 341–365. https://doi.org/10.1016/0304-4076(86)90002-3
11. Y. Wang, J. Li, Q. Zhou, W. Wang, A unified framework for complex survival data: accounting for clustering, cure fractions, and competing risks, *BMC Med. Res. Methodol.*, **26** (2026), 105. https://doi.org/10.1186/s12874-026-02842-z
12. F. Leisch, FlexMix: a general framework for finite mixture models and latent class regression in R, *J. Stat. Softw.*, **11** (2004), 1–18. https://doi.org/10.18637/jss.v011.i08
13. L. Duchateau, P. Janssen, *The frailty model*, New York: Springer, 2008. https://doi.org/10.1007/978-0-387-72835-3
14. P. Hougaard, *Analysis of multivariate survival data*, New York: Springer, 2000. https://doi.org/10.1007/978-1-4612-1304-8
15. X. L. Meng, D. B. Rubin, Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika*, **80** (1993), 267–278. https://doi.org/10.1093/biomet/80.2.267
16. H. Chen, J. Chen, J. D. Kalbfleisch, A modified likelihood ratio test for homogeneity in finite mixture models, *J. R. Stat. Soc. B*, **63** (2001), 19–29. https://doi.org/10.1111/1467-9868.00273
17. H. Kasahara, K. Shimotsu, Non-parametric identification and estimation of the number of components in multivariate mixtures, *J. R. Stat. Soc. B*, **76** (2014), 97–111. https://doi.org/10.1111/rssb.12022
18. C. Hennig, Identifiability of models for clusterwise linear regression, *J. Classif.*, **17** (2000), 273–296. https://doi.org/10.1007/s003570000022
19. A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc. B*, **39** (1977), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
20. B. G. Leroux, Consistent estimation of a mixing distribution, *Ann. Stat.*, **20** (1992), 1350–1360. https://doi.org/10.1214/aos/1176348772
21. J. Chen, P. Li, Hypothesis test for normal mixture models: the EM approach, *Ann. Stat.*, **37** (2009), 2523–2542. https://doi.org/10.1214/08-AOS651
22. A. Xu, Y. Miu, J. Sun, S. Zhou, Y. Tang, A hierarchical Bayesian multivariate Wiener process model with dependent degradation rates and volatilities, *IEEE Trans. Reliab.*, **75** (2026), 1420–1433. https://doi.org/10.1109/TR.2026.3674703

Appendix: Derivation of the M-step dispersion gradient

This appendix gives the short symbolic derivation underlying the $\partial/\partial\eta_m$ gradient formula used in the M-step (Section 5). Writing $\mu = \mu_m(\mathbf{x})$ and $\alpha = \alpha_m$ for brevity, the negative binomial log-density (3.4) is

$$\log f(y;\mu,\alpha) = \log\Gamma(y+\alpha) - \log\Gamma(\alpha) - \log(y!) + \alpha\log\frac{\alpha}{\alpha+\mu} + y\log\frac{\mu}{\alpha+\mu}.$$

Differentiating term-by-term with respect to α:

$$\frac{\partial}{\partial\alpha}\log\Gamma(y+\alpha) = \psi(y+\alpha),$$

$$\frac{\partial}{\partial\alpha}[-\log\Gamma(\alpha)] = -\psi(\alpha),$$
$$\frac{\partial}{\partial\alpha}\left[\alpha\log\frac{\alpha}{\alpha+\mu}\right] = \log\frac{\alpha}{\alpha+\mu} + \alpha\left(\frac{1}{\alpha} - \frac{1}{\alpha+\mu}\right) = \log\frac{\alpha}{\alpha+\mu} + 1 - \frac{\alpha}{\alpha+\mu},$$
$$\frac{\partial}{\partial\alpha}\left[y\log\frac{\mu}{\alpha+\mu}\right] = -\frac{y}{\alpha+\mu}.$$

Summing the four terms gives

$$\frac{\partial}{\partial\alpha}\log f(y;\mu,\alpha) = \psi(y+\alpha) - \psi(\alpha) + \log\frac{\alpha}{\alpha+\mu} + 1 - \frac{y+\alpha}{\alpha+\mu},$$

with the additive constant 1 contributed by the $\alpha\log(\alpha/(\alpha+\mu))$ term. Multiplying by the chain-rule factor $\partial\alpha/\partial\eta_m = \alpha_m$ (since $\eta_m = \log\alpha_m$) reproduces exactly the bracketed expression inside the $\partial/\partial\eta_m$ formula of Section 5. This symbolic derivation, together with the numerical finite-difference check reported in Section 6 (maximum absolute discrepancy 1.6×10^{-7} across four (μ,α) combinations), confirms the gradient formula used in the M-step.

AIMS Press