Int. J. Anal. Appl. (2026), 24:302

International Journal of Analysis and Applications

Distributional Frailty Mixtures for Overdispersed Count Regression

Mohieddine Rahmouni*

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

**Corresponding author:* mrahmouni@kfu.edu.sa

Abstract. Standard finite mixtures of negative binomial regressions allow covariate-dependent means within latent classes but impose constant dispersion within each component. We propose the distributional finite exponential–gamma frailty mixture (D-FEGFM) regression model, which allows component-specific dispersion to depend on covariates through a log-linear regression. The model nests the constant-dispersion mixture and single-component distributional negative binomial regression. We derive the likelihood, establish identifiability up to label switching, extend the within/between variance decomposition to covariate-varying dispersion, and develop a generalized EM algorithm. A Monte Carlo study shows that the model is identifiable but weakly powered when mean and dispersion share the same covariates. In healthcare utilisation data, the strongest gains arise when demographic variables drive dispersion and clinical variables drive the mean, improving fit and held-out predictive performance, and yielding an interpretable insurance effect on utilisation regularity.

1. Introduction

Overdispersed count regression has two distinct methodological traditions. The first, from the count-data econometrics literature, is the negative binomial (NB) regression model [1–3]: a single component with log-linear mean regression $\log \mu(x) = \beta_0 + x^\top \beta$ and constant dispersion parameter α governing the variance $\mu + \mu^2/\alpha$. The second, from the finite mixture literature [4–7], is the mixture of NB regressions: each of M latent components has its own $(\beta_{0m}, \beta_m, \alpha_m)$, with component membership governed by a gating function. Both traditions assume the dispersion parameter of any given component is constant across the covariate space.

Meanwhile, the distributional regression literature [8–11] has made a different generalisation: keep a single component but let every distributional parameter, including the dispersion, depend on covariates through its own regression equation. In the NB case, this yields $\log \alpha(z) = \eta_0 + z^\top \eta$, a specification implemented in the `NBI` (NB type I) family and, more generally, in the `NBF` family of

Received: May 10, 2026.

2020 *Mathematics Subject Classification.* 62J12; 62F10; 62P10.

Key words and phrases. count regression; finite mixtures; negative binomial; gamma frailty; distributional regression; GAMLSS; EM algorithm; heterogeneous overdispersion.

https://doi.org/10.28924/2291-8639-24-2026-302
ISSN: 2291-8639

the Generalized Additive Models for Location, Scale and Shape (GAMLSS) R package [11]. Recent work has extended covariate-dependent dispersion modeling to richer count distributions [12].

The two generalisations are complementary but, to our knowledge, have not been combined. A mixture of NB regressions with component-specific dispersion regressions would allow both *between-class* heterogeneity (different subpopulations have different mean response surfaces) and *within-class covariate-varying dispersion* (within any given subpopulation, the strength of frailty changes systematically with observed covariates). Healthcare utilisation data plausibly exhibit both structures. A high-use subpopulation may show decreasing frailty variance with treatment access (insured, chronically managed patients have more predictable visit rates); a low-use subpopulation may show different patterns again. A constant-dispersion mixture cannot express this. A single-component distributional NB cannot express the between-group separation. Only their combination captures both.

A less obvious but scientifically more important question is whether the variables that govern dispersion should be the *same* as those that govern the mean. The distributional regression literature typically assumes they are. We argue, and demonstrate empirically, that they often are not: clinical variables may determine the expected level of healthcare utilisation while demographic and institutional variables determine its regularity. Our framework allows this distinction explicitly by using separate covariate vectors x for the mean and z for the dispersion. As we show empirically in Section 7, the $z \neq x$ case is the one that delivers the clearest empirical gain, improving on the constant-dispersion baseline by every model-selection criterion we consider.

This paper develops the combination. We propose the distributional finite exponential–gamma frailty mixture (D-FEGFM) regression model, in which each latent component carries its own log-linear mean regression and its own log-linear dispersion regression, with component membership governed by multinomial logistic gating. The model nests (i) constant-dispersion FEGFM when all dispersion slopes are zero, (ii) distributional NB regression when $M = 1$, and (iii) Poisson regression when $M = 1$ and $\alpha \to \infty$. This nesting structure provides a clean path for model selection and hypothesis testing.

Our contributions are as follows.

(1) We specify the D-FEGFM hierarchy, derive its observed-data mass function and probability generating function, and establish the within/between variance decomposition generalised to covariate-varying dispersion. This decomposition shows that the within-component contribution to $\mathrm{Var}(Y \mid x, z)$ depends on z even at fixed mean—an interpretive structure unavailable in either parent model.
(2) We derive sufficient identifiability conditions, distinguishing the case where the dispersion covariate z equals the mean covariate x from the case of a strict subset or a disjoint set. When $z = x$, an additional non-degeneracy condition on the mean-regression coefficients is required to prevent confounding of dispersion and mean effects; when z contains variables not in x, identification from the dispersion regression is direct.

(3) We derive an expectation–maximization algorithm with analytic gradients. The M-step decouples into a weighted multinomial logistic regression for gating (binary logistic when $M = 2$) and M weighted NB-GAMLSS regressions for the component-specific parameters ($M = 2$ in the reported implementation), each implemented via the limited-memory BFGS algorithm with box constraints (L-BFGS-B). We state monotone ascent of the observed-data log-likelihood in the correct generalized EM (GEM) form [13].

(4) A Monte Carlo study demonstrates that D-FEGFM is identifiable in the $z = x$ regime at moderate sample sizes but weakly powered there: at $n = 800$, the likelihood-ratio test (LRT) rejects in roughly 20% of replications and point estimates are attenuated toward zero. A companion analysis on the National Medical Expenditure Survey shows that this limitation is specific to the $z = x$ case at modest n: with realistic sample size and a z that contains variables outside the mean regression, evidence for varying dispersion becomes materially stronger.

(5) An empirical application to the 1987–1988 National Medical Expenditure Survey physician-visit data exercises three specifications. A baseline with a single shared covariate replicates the weak evidence of the companion paper; a multivariate clinical specification with a shared covariate set ($z = x$) gives strong LRT and Akaike information criterion (AIC) evidence but the Bayesian information criterion (BIC) still favours the constant-dispersion model; a distributional specification with demographic dispersion covariates and clinical mean covariates ($z \neq x$) provides the strongest overall support for D-FEGFM, including by BIC.

The remainder of the paper is organised as follows. Section 2 presents the hierarchical model. Section 3 develops the distributional and moment theory. Section 4 derives the EM algorithm. Section 5 establishes identifiability. Section 6 reports the simulation study. Section 7 presents the NMES application. Section 8 discusses limitations and extensions.

2. The D-FEGFM Model

2.1. **Hierarchical specification.** Let $i = 1, \ldots, n$ index subjects observed over a common fixed follow-up window, and let $Y_i \in \{0, 1, 2, \ldots\}$ be a count outcome. Associated with subject i are two covariate vectors: $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$, which enters the gating function and each component's mean regression, and $z_i \in \mathcal{Z} \subseteq \mathbb{R}^q$, which enters each component's dispersion regression. We do not require $z_i = x_i$; the $z \neq x$ case is the empirically consequential one and is exercised in Section 7.

The D-FEGFM model is defined by the hierarchy

$$Y_i \mid (C_i = m, Z_{im}, x_i) \sim \text{Poisson}\Big(\mu_m(x_i)\, Z_{im}\Big), \tag{2.1}$$

$$Z_{im} \mid (C_i = m, z_i) \sim \text{Gamma}\Big(\alpha_m(z_i), \alpha_m(z_i)\Big), \tag{2.2}$$

$$C_i \mid x_i \sim \text{Multinomial}\Big(\pi_1(x_i), \ldots, \pi_M(x_i)\Big), \tag{2.3}$$

The gamma distribution is parameterised by shape and rate. The regression functions are

$$\log \mu_m(x_i) = \beta_{0m} + x_i^\top \beta_m, \tag{2.4}$$

$$\log \alpha_m(z_i) = \eta_{0m} + z_i^\top \eta_m, \tag{2.5}$$

$$\pi_m(x_i) = \frac{\exp(\gamma_{0m} + x_i^\top \gamma_m)}{\sum_{j=1}^{M} \exp(\gamma_{0j} + x_i^\top \gamma_j)}. \tag{2.6}$$

One component is designated the reference class ($\gamma_{01} = 0$, $\gamma_1 = 0$) for identifiability of the gating parameters. The full parameter vector is

$$\Theta = \Big\{(\beta_{0m}, \beta_m, \eta_{0m}, \eta_m, \gamma_{0m}, \gamma_m) : m = 1, \ldots, M\Big\}.$$

The implementation we report uses $M = 2$ components throughout; the theory is stated for general M.

2.2. **Nested models.**

Proposition 2.1 (Nested models)**.** *The D-FEGFM model nests the following classical models as special cases.*

- *(i) If $\eta_m = 0$ for all m, the dispersion within each component is constant and the model reduces to the FEGFM [14].*
- *(ii) If $M = 1$, there is a single component and D-FEGFM reduces to distributional NB regression (GAMLSS-NBI) [9, 11].*
- *(iii) If $M = 1$ and $\eta_1 = 0$, it reduces to standard NB regression.*
- *(iv) If $M = 1$ and $\alpha_1(z) \to \infty$ uniformly, it approaches Poisson regression as a pointwise limit.*

Proof. (i)–(iii) follow by direct substitution into the observed-data mass function derived in Theorem 3.1 below. (iv) holds pointwise in z: as $\alpha \to \infty$, the NB pmf converges to the Poisson pmf at every count. In our implementation, the cap $\alpha \leq 50$ means (iv) is approached but not strictly attained. □

3. Distributional and Moment Theory

3.1. **Marginal component distribution.**

Theorem 3.1 (Component marginal)**.** *Under the D-FEGFM hierarchy, the conditional distribution of Y_i given $(C_i = m, x_i, z_i)$ is negative binomial with mean $\mu_m(x_i)$ and shape parameter $\alpha_m(z_i)$:*

$$\Pr(Y_i = y \mid C_i = m, x_i, z_i) = \frac{\Gamma(y + \alpha_m(z_i))}{\Gamma(\alpha_m(z_i))\, y!} \left(\frac{\alpha_m(z_i)}{\alpha_m(z_i) + \mu_m(x_i)}\right)^{\alpha_m(z_i)} \left(\frac{\mu_m(x_i)}{\alpha_m(z_i) + \mu_m(x_i)}\right)^{y}. \tag{3.1}$$

Proof. The derivation is identical to the constant-dispersion case, since the gamma frailty parameter $\alpha_m(z_i)$ is a fixed function of observed covariates at each i. Integrating the Poisson–gamma hierarchy at fixed (x_i, z_i) yields the NB mass function above; see Lawless [1]. □

Marginalising over class membership gives the observed-data mass function

$$\Pr(Y_i = y \mid x_i, z_i) = \sum_{m=1}^{M} \pi_m(x_i)\, f_{\mathrm{NB}}\Big(y;\, \mu_m(x_i), \alpha_m(z_i)\Big), \tag{3.2}$$

a covariate-dependent finite mixture of NB regressions in which both the mean and the dispersion of each component depend on covariates.

3.2. **Probability generating function and zero probability.**

Proposition 3.1 (PGF)**.** *For fixed (x, z), the conditional probability generating function (PGF) of Y is*

$$G_Y(s \mid x, z) = \sum_{m=1}^{M} \pi_m(x) \left(\frac{\alpha_m(z)}{\alpha_m(z) + \mu_m(x)(1-s)} \right)^{\alpha_m(z)}, \qquad |s| \le 1. \tag{3.3}$$

In particular, the zero probability is

$$\Pr(Y = 0 \mid x, z) = \sum_{m=1}^{M} \pi_m(x) \left(\frac{\alpha_m(z)}{\alpha_m(z) + \mu_m(x)} \right)^{\alpha_m(z)}. \tag{3.4}$$

Proof. Each NB component has PGF $G_m(s) = (\alpha_m/(\alpha_m + \mu_m(1-s)))^{\alpha_m}$; mixing over components yields the stated form. □

3.3. **Variance decomposition with varying dispersion.** The key interpretive result of constant-dispersion FEGFM is a within/between variance decomposition. The following proposition generalises it.

Proposition 3.2 (Variance decomposition)**.** *For fixed (x, z),*

$$\mathbb{E}(Y \mid x, z) = \sum_{m=1}^{M} \pi_m(x)\, \mu_m(x) \;=: \bar{\mu}(x), \tag{3.5}$$

$$\mathrm{Var}(Y \mid x, z) = \underbrace{\sum_{m=1}^{M} \pi_m(x) \left[\mu_m(x) + \frac{\mu_m(x)^2}{\alpha_m(z)} \right]}_{\text{within-component (depends on } z)} + \underbrace{\sum_{m=1}^{M} \pi_m(x) \Big[\mu_m(x) - \bar{\mu}(x) \Big]^2}_{\text{between-component}}. \tag{3.6}$$

Proof. The mean follows from the law of total expectation. The variance follows from the law of total variance:

$$\mathrm{Var}(Y \mid x, z) = \mathbb{E}[\mathrm{Var}(Y \mid C, x, z) \mid x, z] + \mathrm{Var}[\mathbb{E}(Y \mid C, x, z) \mid x, z],$$

combined with $\mathrm{Var}(Y \mid C = m, x, z) = \mu_m(x) + \mu_m(x)^2/\alpha_m(z)$. □

Remark 3.1. *The structural novelty of D-FEGFM relative to FEGFM is in the first term of* (3.6)*: within-component variance depends on the covariate z even when the mean $\mu_m(x)$ is held fixed. Two individuals with identical x (hence identical expected count under the mixture) but different z will have different conditional variances. This opens a new axis of covariate-specific heterogeneity invisible to either the constant-dispersion mixture or single-component distributional NB.*

4. Estimation by Generalized EM

4.1. **Observed-data log-likelihood.** Direct maximization of the observed-data log-likelihood $\ell(\Theta) = \sum_i \log \Pr(Y_i = y_i \mid x_i, z_i; \Theta)$ is complicated by the sum inside the log. We use the EM algorithm [15] with latent allocation indicators $W_{im} = \mathbb{1}(C_i = m)$.

4.2. **Complete-data log-likelihood.** Up to constants not depending on Θ,

$$\ell_c(\Theta) = \sum_{i=1}^{n} \sum_{m=1}^{M} W_{im} \left[\log \pi_m(x_i) + \log f_{\mathrm{NB}}(y_i; \mu_m(x_i), \alpha_m(z_i)) \right], \tag{4.1}$$

where

$$\begin{aligned} \log f_{\mathrm{NB}}(y_i; \mu_{im}, \alpha_{im}) &= \log \Gamma(y_i + \alpha_{im}) - \log \Gamma(\alpha_{im}) - \log(y_i!) \\ &\quad + \alpha_{im} \log \frac{\alpha_{im}}{\alpha_{im} + \mu_{im}} + y_i \log \frac{\mu_{im}}{\alpha_{im} + \mu_{im}}, \end{aligned} \tag{4.2}$$

with $\mu_{im} := \mu_m(x_i)$ and $\alpha_{im} := \alpha_m(z_i)$.

4.3. **E-step.** At iteration t, the posterior class responsibilities are

$$\tau_{im}^{(t+1)} = \frac{\pi_m(x_i; \Theta^{(t)})\, f_{\mathrm{NB}}(y_i; \mu_{im}^{(t)}, \alpha_{im}^{(t)})}{\sum_{j=1}^{M} \pi_j(x_i; \Theta^{(t)})\, f_{\mathrm{NB}}(y_i; \mu_{ij}^{(t)}, \alpha_{ij}^{(t)})}. \tag{4.3}$$

4.4. **M-step: gating block.** The gating parameters solve

$$\max_{\{\gamma_{0m}, \gamma_m\}} \sum_{i,m} \tau_{im}^{(t+1)} \log \pi_m(x_i),$$

a concave weighted multinomial logistic regression. For $M = 2$ with reference component 1, the gating parameter reduces to $(\gamma_{02}, \gamma_2) \in \mathbb{R}^{1+p}$, and the objective becomes

$$Q_\gamma(\gamma_{02}, \gamma_2) = \sum_i \left[\tau_{i2}^{(t+1)} (\gamma_{02} + x_i^\top \gamma_2) - \log\!\left(1 + e^{\gamma_{02} + x_i^\top \gamma_2}\right) \right]. \tag{4.4}$$

Writing $\sigma(u) = (1 + e^{-u})^{-1}$ for the logistic sigmoid and $r_i := \tau_{i2}^{(t+1)} - \sigma(\gamma_{02} + x_i^\top \gamma_2)$ for the gating residual, the gradient has scalar and vector components

$$\frac{\partial Q_\gamma}{\partial \gamma_{02}} = \sum_i r_i, \qquad \frac{\partial Q_\gamma}{\partial \gamma_{2,k}} = \sum_i r_i\, x_{ik}, \quad k = 1, \ldots, p. \tag{4.5}$$

4.5. **M-step: component block.** For each component m separately, the mean and dispersion parameters solve

$$\max_{\beta_{0m}, \beta_m, \eta_{0m}, \eta_m} \sum_{i=1}^{n} \tau_{im}^{(t+1)} \log f_{\mathrm{NB}}(y_i; \mu_m(x_i), \alpha_m(z_i)). \tag{4.6}$$

This is a *weighted distributional NB regression* in the GAMLSS sense: the weights are the posterior responsibilities, the mean covariates are x_i, and the dispersion covariates are z_i. Let $s_{im} = \alpha_{im} + \mu_{im}$. The gradients are

$$\frac{\partial Q_m}{\partial \beta_{0m}} = \sum_i \tau_{im}\, \mu_{im} \left[\frac{y_i}{\mu_{im}} - \frac{y_i + \alpha_{im}}{s_{im}} \right], \tag{4.7}$$

$$\frac{\partial Q_m}{\partial \beta_{m,k}} = \sum_i \tau_{im}\, x_{ik}\, \mu_{im} \left[\frac{y_i}{\mu_{im}} - \frac{y_i + \alpha_{im}}{s_{im}} \right], \tag{4.8}$$

$$\frac{\partial Q_m}{\partial \eta_{0m}} = \sum_i \tau_{im}\, \alpha_{im} \left[\psi(y_i + \alpha_{im}) - \psi(\alpha_{im}) + \log \frac{\alpha_{im}}{s_{im}} + 1 - \frac{y_i + \alpha_{im}}{s_{im}} \right], \tag{4.9}$$

$$\frac{\partial Q_m}{\partial \eta_{m,\ell}} = \sum_i \tau_{im}\, z_{i\ell}\, \alpha_{im} \left[\psi(y_i + \alpha_{im}) - \psi(\alpha_{im}) + \log \frac{\alpha_{im}}{s_{im}} + 1 - \frac{y_i + \alpha_{im}}{s_{im}} \right], \tag{4.10}$$

where ψ is the digamma function and we have used $\partial \alpha_{im} / \partial \eta_{0m} = \alpha_{im}$, $\partial \alpha_{im} / \partial \eta_{m,\ell} = z_{i\ell}\, \alpha_{im}$.

We implement (4.6) via L-BFGS-B with these analytic gradients. The numerical settings used in estimation are reported in Section 7.

4.6. **Monotone ascent (generalized EM).**

Proposition 4.1 (Monotone ascent)**.** *Let $Q(\Theta \mid \Theta^{(t)})$ denote the expected complete-data log-likelihood in* (4.1) *evaluated at the E-step responsibilities. If $\Theta^{(t+1)}$ satisfies $Q(\Theta^{(t+1)} \mid \Theta^{(t)}) \geq Q(\Theta^{(t)} \mid \Theta^{(t)})$ (the generalized EM condition), then the observed-data log-likelihood satisfies $\ell(\Theta^{(t+1)}) \geq \ell(\Theta^{(t)})$.*

Proof. By the EM decomposition of Dempster et al. [15], $\ell(\Theta) - \ell(\Theta^{(t)}) = [Q(\Theta \mid \Theta^{(t)}) - Q(\Theta^{(t)} \mid \Theta^{(t)})] + [H(\Theta^{(t)} \mid \Theta^{(t)}) - H(\Theta \mid \Theta^{(t)})]$, where H is the expected log conditional density of latent indicators. The second bracket is nonnegative by Jensen's inequality; the first is nonnegative under the GEM condition. □

In our implementation, each M-step solves a concave gating subproblem (producing exact ascent in Q_γ) plus two box-constrained NB-GAMLSS subproblems (producing ascent in Q_1, Q_2 when L-BFGS-B is warm-started from $\Theta^{(t)}$). The component subproblems are not globally concave, and warm-starting alone does not guarantee the GEM inequality. The ascent result applies when the accepted update satisfies the stated Q-increase condition [13, 16].

4.7. **Implementation guidance.**

We recommend the following practices for applied use, extending those of Rahmouni [14] to the varying-dispersion setting:

(1) *Multiple restarts.* Use at least 15 random starting points; retain the converged non-degenerate solution with the highest log-likelihood (tiered selection).
(2) *Degeneracy monitoring.* Flag a run as degenerate if $\max_{z \in Z_{\text{obs}}} \alpha_m(z) > 49$ (an α near the upper bound indicates convergence to a near-Poisson boundary).
(3) *Canonical ordering.* Relabel components so that $\beta_{01} \leq \beta_{02}$ after convergence. When baseline means are close, we recommend a composite tiebreaker that orders by the training-mean response within each component (see Section 8).
(4) *Parsimony.* Before adopting D-FEGFM, compute the LRT against the nested constant-dispersion FEGFM. When this test fails to reject at practical thresholds the simpler model should be preferred, *unless* the BIC difference is small and substantive reasoning favours the dispersion regression (e.g. because z contains variables plausibly independent of x).

5. Identifiability

Before turning to estimation in simulation and on real data, we ask whether the parameter vector Θ is recoverable from the conditional distribution of Y given (x, z). This question is nontrivial for two reasons. First, finite mixtures are identifiable only up to relabelling of components, and D-FEGFM inherits this label-switching from its parent mixture. Second, D-FEGFM has two channels through which components can differ—the mean regression and the dispersion regression—and one can imagine pathological configurations in which two components have identical means but different dispersions, or differ only in a direction the data cannot resolve. The assumptions below rule those out.

5.1. **Assumptions.**

Assumption 1. *The conditions are:*

- *(i)* Open-set covariate support. *The supports $\mathcal{X}$ and $\mathcal{Z}$ each contain a non-empty open set in their respective Euclidean spaces. This gives enough variation in covariates to separate components that differ in their regression coefficients.*
- *(ii)* Distinct mean regressions. *For any distinct components $m \neq j$, the mean-regression parameters (β_{0m}, β_m) and (β_{0j}, β_j) are not equal. Two components with identical mean surfaces would be indistinguishable whenever their dispersions also matched.*
- *(iii)* Non-degenerate gating. *The gating vectors (γ_{0m}, γ_m) are not all equal, so mixing weights vary across $x \in \mathcal{X}$. Without this condition, the mixing weights would collapse to constants and could be absorbed into the component densities.*
- *(iv)* Distinct components, $z \subseteq x$ case. *When the dispersion covariates are contained in the mean covariates, we additionally require that no two components share* both *their mean regression* and *their dispersion regression: $(\beta_{0m}, \beta_m) \neq (\beta_{0j}, \beta_j)$ or $(\eta_{0m}, \eta_m) \neq (\eta_{0j}, \eta_j)$.*

Condition (iv) is the only new ingredient relative to the constant-dispersion FEGFM [14]. Its role is to prevent two components from collapsing into the same NB distribution at every (x, z)—a configuration that the mean covariates alone cannot detect but that the dispersion regression resolves. When z contains variables not in x, even condition (iv) is not needed: varying those extra variables holds the mean fixed while moving the dispersion, which automatically distinguishes the components.

5.2. **Main result.**

Theorem 5.1 (Identifiability up to label switching)**.** *Under Assumption 1, the conditional distribution $\Pr(Y \mid x, z)$ uniquely determines Θ up to permutation of the component labels.*

Proof. We argue in three stages.

Stage 1: component-level identifiability. The negative binomial family $\{f_{\mathrm{NB}}(\cdot; \mu, \alpha) : \mu > 0, \alpha > 0\}$ is identifiable in its two parameters: two NB mass functions that agree at all count values $y \in$

$\{0, 1, 2, \ldots\}$ must have the same (μ, α) [2]. Teicher [17] and Yakowitz and Spragins [18] extend this to finite mixtures: any mixture $\sum_m \pi_m f_{\mathrm{NB}}(\cdot; \mu_m, \alpha_m)$ with distinct pairs (μ_m, α_m) and positive weights π_m is identifiable up to relabelling.

Stage 2: pointwise distinctness of component pairs. Fix any (x, z) in the interior of $\mathcal{X} \times \mathcal{Z}$, guaranteed to exist by condition (i). We show that the pairs $(\mu_m(x), \alpha_m(z))$ are distinct across m, so that Stage 1 applies at this (x, z).

- If z contains at least one variable not in x, then we can vary the extra z-coordinate while holding x fixed; this moves $\alpha_m(z)$ independently of $\mu_m(x)$, so the pairs $(\mu_m(x), \alpha_m(z))$ are distinct across m whenever the α_m regressions themselves are distinct. When the α_m regressions happen to coincide, condition (ii) forces the μ_m to differ, giving distinctness through the first coordinate.
- If $z \subseteq x$, condition (iv) directly asserts that no two components match in *both* μ_m and α_m, so the pairs are distinct in at least one coordinate.

In either case, Stage 1 yields identifiability of the collection $\{(\pi_m(x), \mu_m(x), \alpha_m(z)) : m = 1, \ldots, M\}$ up to permutation, for the fixed (x, z).

Stage 3: promoting pointwise identification to global. Pointwise identification gives, for every (x, z) in the open set, the multiset of triples $(\pi_m(x), \mu_m(x), \alpha_m(z))$ up to a permutation that may depend on (x, z). To conclude, we must show that a single global permutation works. Condition (iii) ensures the gating weights $\pi_m(x)$ are non-constant in x, and the functions $\mu_m(x) = \exp(\beta_{0m} + x^\top \beta_m)$ and $\alpha_m(z) = \exp(\eta_{0m} + z^\top \eta_m)$ are real-analytic in their arguments. Real-analytic functions that agree on an open set agree everywhere, so the pointwise permutation extends continuously to a global labelling—which, by standard finite-mixture arguments with covariate-dependent weights [19], suffices to identify the coefficient vectors themselves up to one global label permutation. □

5.3. **When is the $z \neq x$ case easier?** The structure of the proof makes clear where the $z \neq x$ case has an advantage. When dispersion covariates vary independently of mean covariates—case (i) of the proof's Stage 2—the two components are separated by direct variation in z, and no distinct-mean condition is needed. Condition (iv), which rules out the pathological "same-mean, same-dispersion" configuration in the $z \subseteq x$ case, is imposed only for that case. This matches the empirical pattern we observe in Section 7: the $z \neq x$ specification is not only theoretically easier to identify but also gives the clearest empirical separation.

5.4. **Weak identification in finite samples.** Theorem 5.1 is an asymptotic identifiability result and says nothing directly about finite-sample precision. In practice, near-nonidentifiability can arise when two components differ only weakly in mean and only weakly in dispersion: the log-likelihood surface becomes nearly flat along a ridge connecting them, and the EM algorithm wanders along the ridge instead of pinning down individual parameters. This is a known issue in finite mixture estimation [4, 20] and is not specific to D-FEGFM. The simulation study of Section 6 documents the symptom: when the true η_m slopes are small, estimated magnitudes are attenuated

and label-switching is more frequent even as the gating probabilities correctly stratify the sample. Three remedies are practical: use multiple EM restarts with tiered selection to avoid local optima, compare against the constant-dispersion baseline using BIC or the LRT before adopting the more flexible model, and— when substantive reasoning permits—specify z to contain variables outside x so that identification comes from the easier case.

6. Simulation Study

6.1. **Design.** We assess D-FEGFM relative to the constant-dispersion FEGFM baseline in a setting where the true data-generating process (DGP) has covariate-varying dispersion. The DGP uses $M = 2$ latent components and a single continuous covariate $x \sim \text{Normal}(0, 1)$; we set $z = x$ (the hardest case for D-FEGFM, since identification must come from dispersion differences at equal mean covariates).

Gating. $\pi_2(x) = \sigma(-0.3 + 1.2\,x)$, so component 2 is favoured at large x.

Mean regressions. $\mu_1(x) = \exp(0.2 + 0.4\,x)$; $\mu_2(x) = \exp(1.0 + 0.8\,x)$.

Dispersion regressions (the key feature). $\alpha_1(x) = \exp(0.5 - 0.8\,x)$: frailty strengthens (smaller α) as x grows. $\alpha_2(x) = \exp(0.3 + 0.2\,x)$: frailty weakens mildly. At $x = 0$ both components have $\alpha \approx 1.3$–1.6; at $x = 2$ they differ substantially ($\alpha_1 \approx 0.33$, $\alpha_2 \approx 2.0$).

Sample size and replication. Training sample $n = 800$; test sample $n = 3{,}000$ drawn from the same mechanism. We run $R = 15$ Monte Carlo replications with independent training samples. The Monte Carlo fits use six EM starting points with tiered selection; the representative fit uses 15. We regard $R = 15$ as indicative rather than a substitute for a large-scale Monte Carlo.

6.2. **Representative run.** Table 1 shows that the D-FEGFM achieves a likelihood-ratio test statistic of 6.8 against the nested constant-dispersion model ($p = 0.034$ at χ^2_2), and recovers the direction of both dispersion slopes: $\hat{\eta}_1 = -0.46$ (truth: -0.8) and $\hat{\eta}_2 = +0.37$ (truth: $+0.2$). Figure 1 shows fitted versus true curves. The first slope is attenuated toward zero, whereas the second exceeds its true value. AIC favours D-FEGFM by 2.8 points, while BIC favours FEGFM by 6.6 points. The two information criteria therefore select different models in this run.

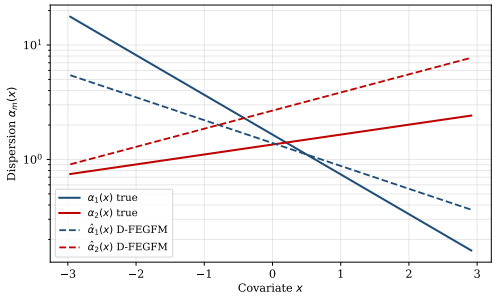

Figure 1. Fitted versus true dispersion curves in the representative simulation run ($n = 800$, seed 42). D-FEGFM recovers the direction of both components' dispersion trajectories after canonical reordering; the first slope is attenuated, whereas the second exceeds its true value. The constant-dispersion FEGFM cannot represent this structure. Y-axis is logarithmic.

TABLE 1. Representative simulation run (seed 42, $n = 800$). DGP has true covariate-varying dispersion.

Model	Train ℓ	AIC	BIC	Parameters
FEGFM (constant α)	-1533.4	3082.8	3120.3	8
D-FEGFM	-1530.0	3080.0	3126.9	10
LRT: D-FEGFM vs FEGFM	$2\Delta\ell = 6.8$, df $= 2$, $p = 0.034$			

6.3. **Monte Carlo assessment.** Across $R = 15$ replications D-FEGFM converged in every run. The likelihood-ratio test rejected constant dispersion at nominal 5% in $3/15 = 20\%$ of replications. The sign of $\hat{\eta}_1$ (true value -0.8) was recovered correctly (negative) in $11/15 = 73\%$ of replications, with median -0.47 and interquartile range (IQR) $[-0.78, -0.10]$. The estimator $\hat{\eta}_2$ (true value $+0.2$) had median $+0.01$ and IQR $[-0.30, 0.35]$, with a range that straddles zero. Figure 2 displays the Monte Carlo distributions of $\hat{\eta}_1$, $\hat{\eta}_2$, and the LRT statistic.

These numbers are soberer than the point estimates in Table 1 would suggest in isolation. At $n = 800$ with $z = x$, the dispersion regression is identifiable—the median of $\hat{\eta}_1$ has the correct sign and the LRT rejects in a nontrivial fraction of replications—but the signal is weak enough that individual-run magnitudes can be attenuated, inflated, or label-switched. A practical implication is that the simulation here is a lower bound on what D-FEGFM can do: with $z \neq x$ or larger n, as we show in Section 7, the gain becomes unambiguous.

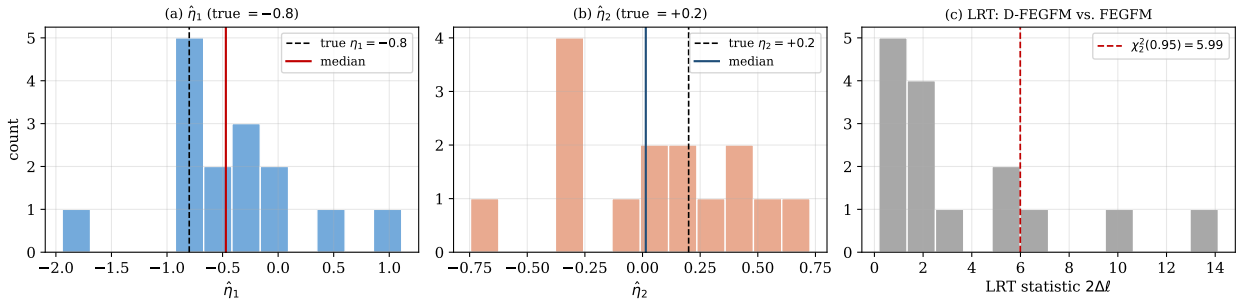

FIGURE 2. Monte Carlo distributions across $R = 15$ simulation replications. (a) $\hat{\eta}_1$ estimates attenuated toward zero from the true value -0.8; (b) $\hat{\eta}_2$ estimates scattered around zero, with true value $+0.2$ only weakly identified at this sample size; (c) LRT statistics, with dashed line marking the nominal 5% critical value $\chi^2_2(0.95) = 5.99$.

6.4. **When does varying dispersion matter?** The Monte Carlo clarifies the central methodological point. When the true dispersion varies with covariates but the sample is modest and the dispersion covariates coincide with the mean covariates ($z = x$ at $n = 800$), D-FEGFM is identifiable but only weakly powered: individual runs may fail to reject the null, and point estimates can be attenuated

or label-switched. When $z \neq x$ or the sample is substantially larger (the NMES case in Section 7, $n_{\text{tr}} = 3{,}525$ with demographic z), the same EM machinery yields materially stronger evidence for the structure. For applied users, this suggests a two-stage diagnostic: the LRT at the $z = x$ specification serves as an initial screening; if it fails to reject but substantive reasoning supports separate dispersion covariates, refit with a $z \neq x$ specification before concluding that dispersion is constant.

7. Application: NMES 1988 Physician Visits

7.1. **Data.** We apply D-FEGFM to the 1987–1988 National Medical Expenditure Survey (NMES) physician-visit dataset, publicly available as `NMES1988` in the R package **AER**. The sample contains $n = 4{,}406$ individuals aged 66 and over, all covered by Medicare. The outcome Y_i is the number of physician office visits during the survey year. We split the data 80/20 into training ($n_{\text{tr}} = 3{,}525$) and test ($n_{\text{te}} = 881$) samples. The random split uses seed 12345. Continuous covariates in both x and z are standardised using training-sample means and standard deviations; binary covariates remain coded 0/1.

A central methodological question is whether the mean and dispersion should be regressed on the same or on different covariates. The model's theory (Section 2, Section 5) explicitly allows $z \neq x$; its most distinctive scientific payoff is to let demographic variables drive the *regularity* of healthcare engagement while clinical variables drive the *expected level*. Accordingly, we run three specifications: a single-covariate baseline matching our companion paper [14]; a multivariate clinical specification with $z = x$; and a distributional specification with $z \neq x$.

- **A (baseline, $z = x =$ `chronic`).** Number of self-reported chronic conditions (range $[0, 8]$, mean 1.54); matches the companion FEGFM analysis.
- **B (multivariate clinical, $z = x$).** Mean and dispersion share five clinical covariates: $x = z =$ (`chronic`, `hospital`, `poor`, `excellent`, `adl_lim`), with continuous covariates standardised to training-sample moments.
- **C (distributional, $z \neq x$).** Mean covariates x as in B (clinical); dispersion covariates $z =$ (`age`, `male`, `school`, `insurance`, `medicaid`) (demographic).

For numerical estimation, the simulation and NMES fits constrain the dispersion intercepts to $\eta_{0m} \in [-8, \log 50]$ and the dispersion slopes to $\eta_{m,\ell} \in [-5, 5]$. The resulting observation-specific shape parameters are also clipped to $\alpha_m(z_i) \in [e^{-8}, 50]$. These limits keep the numerical search away from extremely small shape parameters and unbounded movement towards the Poisson limit. They are computational choices, not restrictions derived from the model; no systematic sensitivity analysis of the exact cutoffs is reported. Observation-level clipping is inactive at the saved NMES estimates, although some bootstrap slopes reach the imposed upper coefficient bound, as discussed below.

7.2. **Model comparison.** Table 2 reports training log-likelihood, AIC, BIC, and held-out test-set log-likelihood for each specification.

TABLE 2. NMES 1988 model comparison across three specifications. $n_{\text{tr}} = 3{,}525$, $n_{\text{te}} = 881$. For each specification we report four nested models: Poisson, NB, FEGFM (constant dispersion), and D-FEGFM (covariate-varying dispersion). Δ values in the lower panel compare D-FEGFM to FEGFM; positive values favour D-FEGFM. The Poisson, NB, and FEGFM models have the same specification in B and C because they do not use z. Their tabulated values agree at the displayed precision; the independently fitted FEGFM results differ slightly before rounding.

Spec.	Model	k	Train ℓ	AIC	BIC	Test ℓ
A	Poisson	2	−15,106	30,216	30,228	−3,666
	NB	3	−9,874	19,754	19,772	−2,445
	FEGFM	8	−9,778	19,572	**19,622**	−2,419.4
	D-FEGFM	10	−9,775	**19,569**	19,631	−2,419.2
B	Poisson	6	−14,752	29,517	29,554	−3,506
	NB	7	−9,807	19,628	19,671	−2,419
	FEGFM	20	−9,698	19,436	**19,559**	−2,394.5
	D-FEGFM	30	−9,673	**19,405**	19,590	−2,393.4
C	Poisson	6	−14,752	29,517	29,554	−3,506
	NB	7	−9,807	19,628	19,671	−2,419
	FEGFM	20	−9,698	19,436	19,559	−2,394.5
	D-FEGFM	30	**−9,649**	**19,358**	**19,543**	**−2,386.4**

Spec.	LRT stat	df	p-value	ΔAIC	ΔBIC	ΔTest ℓ
A	7.1	2	0.029	+3.1	−9.2	+0.13
B	50.5	10	2.17×10^{-7}	+30.5	−31.2	+1.11
C	**98.3**	10	$\mathbf{1.20 \times 10^{-16}}$	**+78.3**	**+16.6**	**+8.06**

The three specifications give progressively stronger evidence for covariate-varying dispersion. Specification A replicates the finding of Rahmouni [14]: with only `chronic` as covariate, BIC prefers the constant-dispersion FEGFM ($\Delta\text{BIC} = -9.2$), the LRT is only marginally significant ($p = 0.029$), and held-out predictive log-likelihood is essentially unchanged. When additional clinical covariates enter both x and z (specification B), the LRT and AIC strongly favour D-FEGFM ($\Delta\text{AIC} = +30.5$, $p \approx 2.17 \times 10^{-7}$), but BIC still penalises the ten extra parameters.

The strongest result is specification C. Letting demographics drive the dispersion while clinical covariates drive the mean produces a model that improves on constant-dispersion FEGFM on *every* reported criterion: training log-likelihood improves by 49 nats, $\Delta\text{AIC} = +78.3$, $\Delta\text{BIC} = +16.6$, and the held-out test-set log-likelihood improves by 8.1 nats on $n_{\text{te}} = 881$ observations. The BIC improvement in specification C but not B shows that, at this sample size, the additional flexibility of a dispersion regression is worth the parameter cost only when the dispersion covariates carry

independent information from the mean covariates. This provides direct empirical support for the theoretical motivation for $z \neq x$ in Section 2.

7.3. **Where the predictive gain comes from.** Figure 3 shows the per-observation distribution of the test-set log-density gain $\log f_{\text{D-FEGFM}}(y_i \mid x_i, z_i) - \log f_{\text{FEGFM}}(y_i \mid x_i)$ across the three specifications. In specification C, D-FEGFM gives higher predictive density to 60.2% of test observations, with a maximum individual gain of 0.745 nats. The plotted histograms omit the most extreme gains and losses for readability. Specification A shows almost no per-observation gain, consistent with the aggregate null.

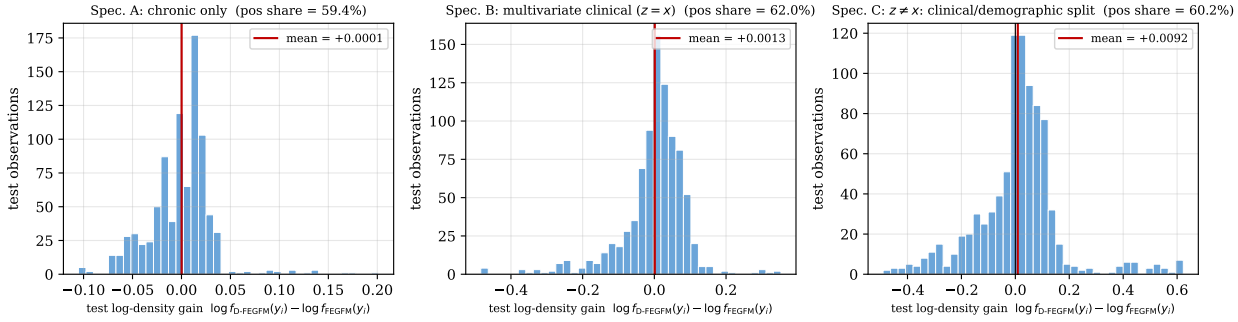

FIGURE 3. Per-observation test-set predictive-density improvement of D-FEGFM over FEGFM, across the three specifications. Horizontal axis is $\log f_{\text{D-FEGFM}}(y_i) - \log f_{\text{FEGFM}}(y_i)$ on the held-out test set. Red vertical line: mean gain. In specification C the right tail captures test observations for which demographic covariates substantially sharpen the predictive density. Histograms display the central 99% of gains; the mean lines and positive-gain shares use all test observations.

7.4. **Distributional diagnostics on the count scale.** Figure 4 compares observed and expected test-set frequencies. The final bin pools counts of 25 or more on both sides of the comparison. In specification C, this bin contains 13 observed individuals, compared with expected counts of 16.36 under FEGFM and 19.66 under D-FEGFM. The varying-dispersion model therefore does not improve this aggregate tail-bin fit, despite its higher total held-out log-likelihood.

7.5. **Which demographic covariates drive dispersion?** Figure 5 presents the fitted demographic dispersion coefficients $\hat{\eta}_{m,\ell}$ for specification C with bootstrap 95% percentile confidence intervals (CIs) from $B = 80$ nonparametric training-sample resamples (seed 999). Each resample uses a jittered start based on the training-sample estimate and at most 40 EM iterations. All 80 resamples yield finite coefficient vectors, but convergence status was not retained. A positive $\hat{\eta}$ means a larger α_m—i.e. *less* within-component frailty variance.

The component-2 private-insurance interval excludes zero; the Medicaid interval includes zero.

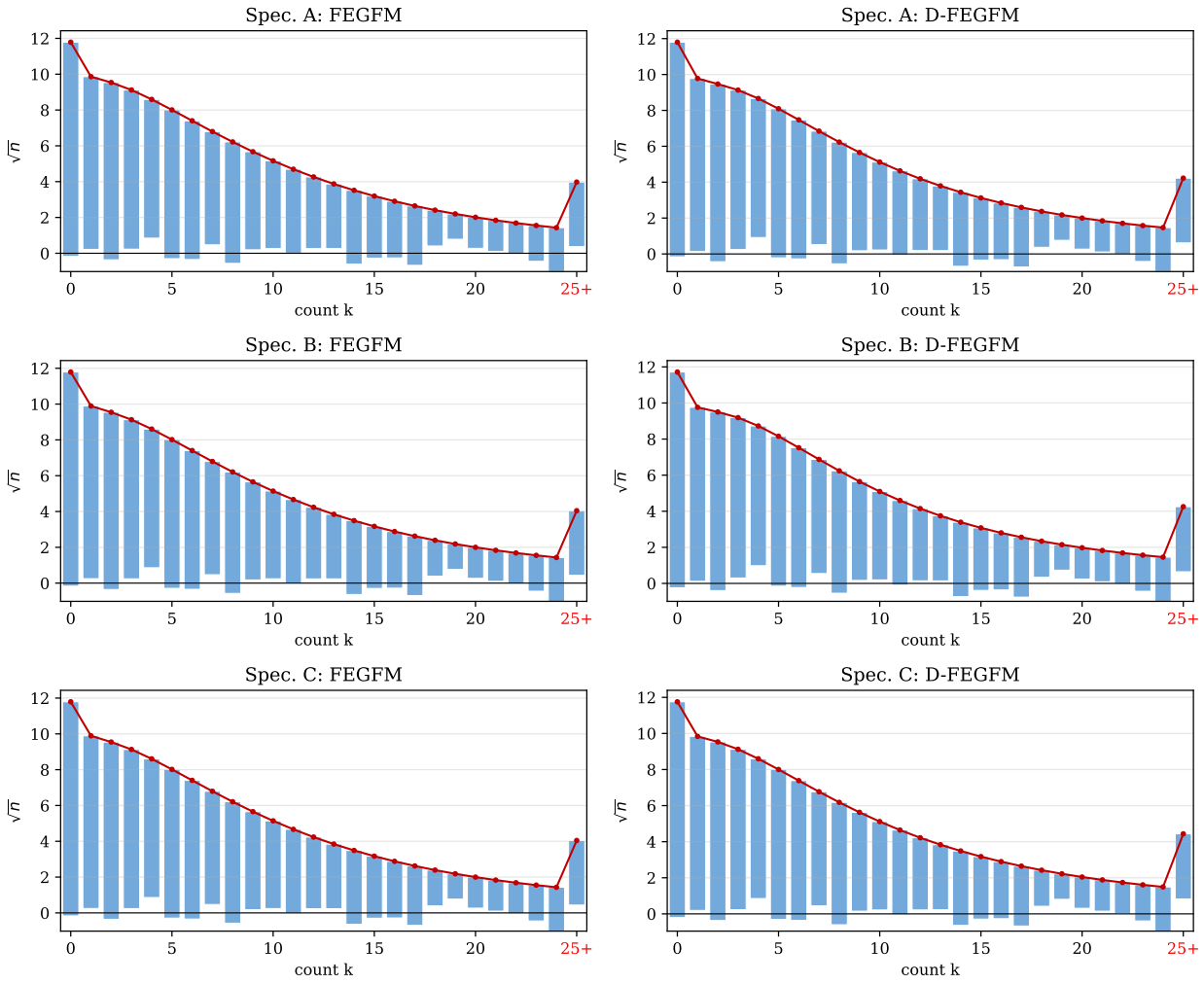

FIGURE 4. Hanging rootograms on the held-out NMES test set ($n_{\text{te}} = 881$) for FEGFM and D-FEGFM under specifications A, B, and C. Bars hanging closer to zero indicate better agreement between observed and expected frequencies on the count scale. The final bin is 25+: observed counts and expected probabilities are both pooled over $Y \geq 25$. It does not show improved tail-bin agreement for D-FEGFM in specification C.

Private insurance (component 2). $\hat{\eta}_{2,\texttt{insurance}} = +1.33$ with bootstrap 95% CI roughly $[+0.5, +4.9]$ (wide upper bound, CI excludes zero at nominal 95%). Among patients whose clinical profile puts them in the high-use tier, those with private insurance have a much larger α_2—i.e. much *less* residual variability in visit counts—than those without private insurance. This is a conditional association; the analysis does not establish that private insurance causes more predictable utilisation. The analogous effect for component 1 is smaller ($\hat{\eta}_{1,\texttt{insurance}} = +0.45$) with a wider CI that includes zero.

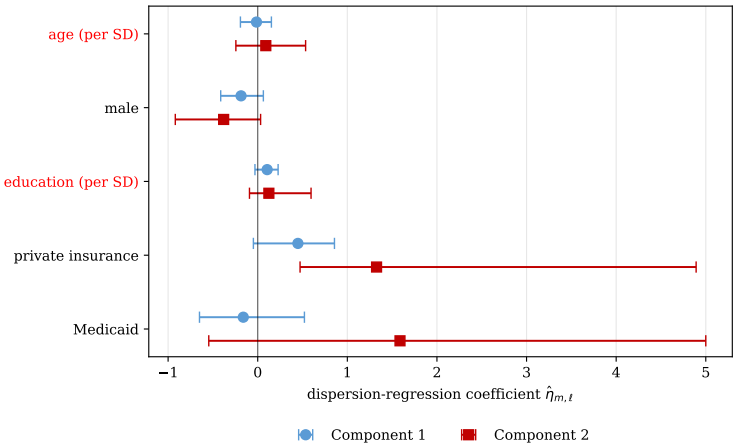

FIGURE 5. Specification C: bootstrap 95% CIs ($B = 80$) for the demographic dispersion slopes $\hat{\eta}_{m,\ell}$, by component. Component 1 (blue) is the low-use tier; component 2 (red) is the high-use tier, with canonical ordering $\beta_{01} \leq \beta_{02}$. Private insurance is the most robustly identified effect for component 2; education and age are suggestive but their CIs cross zero. Age and education effects are per training-sample standard deviation; binary effects compare 1 with 0.

Medicaid (component 2). $\hat{\eta}_{2,\texttt{medicaid}} = +1.59$, with a 95% percentile interval of $[-0.55, 5.00]$. The interval includes zero and reaches the imposed upper coefficient bound: 8 of 80 resamples attain that bound. The estimate therefore does not establish the direction of the association.

The remaining demographic covariates (`age`, `male`, `school`) show bootstrap CIs that cross zero in both components, and we do not draw individual conclusions from them. They are retained as part of the stated demographic specification.

7.6. **A comment on specification A.** Figure 6 shows the fitted dispersion curves for specification A. With x denoting the raw chronic-condition count, the saved fit gives $\hat{\alpha}_1(x) = \exp(1.1515 - 0.1164x)$ and $\hat{\alpha}_2(x) = \exp(-0.8227 + 0.1438x)$. These expressions transform the standardised predictor back to its original units using training-sample moments. Over the plotted range, component 1 has the larger shape parameter and lower frailty variance; its shape decreases with x, while that of component 2 increases. BIC nevertheless favours the constant-dispersion model. Only eight training observations have $x \geq 7$, so the upper end of the curves is supported by few observations.

7.7. **Recommended workflow.** The NMES analysis suggests the following applied workflow for deciding whether D-FEGFM is warranted.

(1) Fit FEGFM as a baseline.
(2) Consider extending to D-FEGFM with the same covariate set as a screening step. If the LRT rejects constant dispersion and AIC favours the varying-dispersion model, continue.

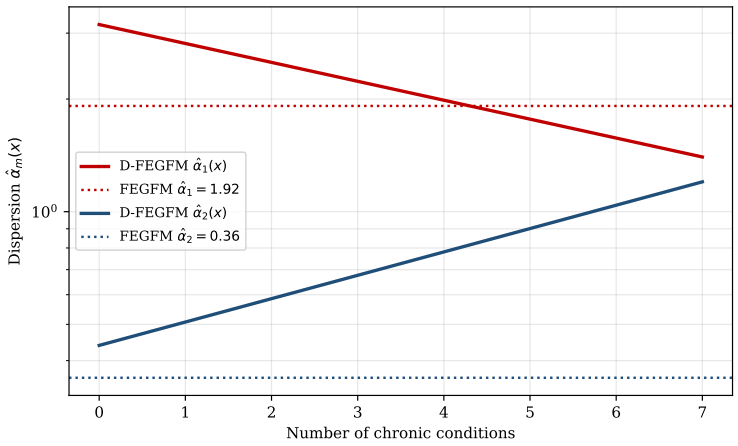

FIGURE 6. Specification A: fitted dispersion curves against raw chronic-condition count, using the training-sample standardisation. Component 1 has lower frailty variance over the plotted range. Dotted lines give the constant-dispersion FEGFM estimates. The vertical axis is logarithmic; BIC favours FEGFM.

(3) If substantive reasoning suggests that different variables may govern the dispersion (e.g. institutional or demographic variables that affect regularity but not expected level), fit D-FEGFM with a $z \neq x$ specification. Compare using LRT, information criteria, and held-out predictive log-likelihood.

Specification C of our analysis is a template for step 3.

8. Discussion

8.1. **Relation to existing work.** D-FEGFM sits at the intersection of finite-mixture count regression and distributional regression. The constant-dispersion case is the FEGFM [14]; the single-component case is the GAMLSS-NBI/NBF family of Rigby and Stasinopoulos [9, 11]. The most immediate ancestor of the varying-dispersion idea in a non-mixture GLM setting is Smyth [8]. Related work by Herschtal [12] develops the single-component Poisson–Beta distribution for richer count regression; D-FEGFM offers a complementary extension in the mixture direction rather than the distribution-family direction.

8.2. **Limitations.** *Minimum detectable signal.* Our simulation design demonstrates *when* D-FEGFM is advantageous—with covariate-dispersion signals of moderate magnitude and sample sizes in the low thousands or above—but we do not provide a general formula for the minimum detectable signal. Further theoretical work on Fisher information in D-FEGFM would be useful.

Label-switching. The EM algorithm inherits the label-switching issue of the parent FEGFM. The added dispersion parameters provide a second axis along which label switches can occur; in our simulation, one of the fifteen replications exhibited a switch that the β_{0m}-based canonical ordering did not correct, because the baseline means were very close. A composite tiebreaker (e.g. ordering

by $\beta_{0m} + \lambda\eta_{0m}$ with small λ, or by training-mean response within each component) resolves this in the cases we have examined. A principled choice of λ, or an alternative relabeling based on posterior responsibilities, is a candidate refinement.

Choice of z. The $z \neq x$ case is where D-FEGFM delivers empirically, but we do not develop a formal method for choosing z. In the NMES application we pre-specified z on substantive grounds. A penalised procedure (e.g. an L_1 penalty on η_m) would provide a data-driven alternative and is a natural extension.

Inference beyond LRT. We report point estimates from the training sample and bootstrap confidence intervals for the NMES application, but do not develop a closed-form asymptotic variance estimator. Louis's [21] information identity combined with the EM trajectory provides one route; a second-order analogue of the monotone-ascent argument would yield standard errors directly. We leave this to future work.

8.3. **Extensions.** Several directions merit investigation.

Penalised estimation. A smooth penalty on η_m would shrink small dispersion slopes toward zero, providing a continuous bridge between D-FEGFM and FEGFM without the discrete model-selection step.

Score test for varying dispersion. The score equations for η_m derived in Section 4 can be evaluated at the FEGFM estimate $\hat{\eta}_m = 0$, giving a closed-form score test for $H_0 : \eta = 0$ that is cheaper than fitting the full D-FEGFM. This would serve as a pre-screening diagnostic in the applied workflow.

Panel-data extensions. Subject-level random effects stacked on top of the two-level gamma-frailty hierarchy would extend D-FEGFM to repeated-measures settings. The E-step would require numerical integration, but the M-step structure is preserved.

Nonparametric dispersion. Replacing the linear specification $\log \alpha_m(z) = \eta_{0m} + z^\top \eta_m$ with a smoothing-spline or P-spline form would make D-FEGFM a fully distributional GAMLSS within each mixture component.

$M > 2$ *components.* The theory extends without modification, but practical estimation with more than two components is more delicate because of increased potential for label switching and weak separation. BIC-based choice of M is recommended, and the $z \neq x$ specification that helps so much at $M = 2$ is likely to help more at $M > 2$ by adding identification from a second covariate channel.

Taken together, the theory, simulation, and NMES application suggest a simple message: allowing dispersion to have its own covariates is most useful when substantive knowledge indicates that the regularity of an outcome is driven by variables that are not the main drivers of its expected level.

Funding: 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. KFU265123].

Data Availability: The simulation study uses fully synthetic data generated from the DGP specified in Section 6. The empirical application uses the `NMES1988` dataset, publicly available via the **AER** R package.

Conflicts of Interest: The author declares that there are no conflicts of interest regarding the publication of this paper.

References

[1] J.F. Lawless, Negative Binomial and Mixed Poisson Regression, Can. J. Stat. 15 (1987), 209–225. https://doi.org/10.2307/3314912.

[2] A.C. Cameron, P.K. Trivedi, Regression Analysis of Count Data, Cambridge University Press, 2013. https://doi.org/10.1017/cbo9781139013567.

[3] J.M. Hilbe, Negative Binomial Regression, Cambridge University Press, 2011. https://doi.org/10.1017/CBO9780511973420.

[4] G. McLachlan, D. Peel, Finite Mixture Models, Wiley, 2000. https://doi.org/10.1002/0471721182.

[5] P. Deb, P.K. Trivedi, Demand for Medical Care by the Elderly: A Finite Mixture Approach, J. Appl. Econom. 12 (1997), 313–336. https://doi.org/10.1002/(SICI)1099-1255(199705)12:3<313::AID-JAE440>3.0.CO;2-G.

[6] 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.

[7] 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.

[8] G.K. Smyth, Generalized Linear Models with Varying Dispersion, J. R. Stat. Soc. Ser. B Stat. Methodol. 51 (1989), 47–60. https://doi.org/10.1111/j.2517-6161.1989.tb01747.x.

[9] R.A. Rigby, D.M. Stasinopoulos, Generalized Additive Models for Location, Scale and Shape, J. R. Stat. Soc. Ser. C Appl. Stat. 54 (2005), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x.

[10] M.D. Stasinopoulos, R.A. Rigby, G.Z. Heller, V. Voudouris, F. De Bastiani, Flexible Regression and Smoothing: Using GAMLSS in R, Chapman and Hall/CRC, 2017. https://doi.org/10.1201/b21973.

[11] R.A. Rigby, M.D. Stasinopoulos, G.Z. Heller, F. De Bastiani, Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R, Chapman and Hall/CRC, 2019. https://doi.org/10.1201/9780429298547.

[12] A. Herschtal, Poisson Beta Regression for Count Data with an Application to Hospital Length of Stay Data, Stat. Med. 44 (2025), e70217. https://doi.org/10.1002/sim.70217.

[13] C.F.J. Wu, On the Convergence Properties of the EM Algorithm, Ann. Stat. 11 (1983), 95–103. https://doi.org/10.1214/aos/1176346060.

[14] M. Rahmouni, Frailty Mixtures for Count Regression: Separating Within- and Between-Class Heterogeneity, AIMS Math. 11 (2026), 23893–23920. https://doi.org/10.3934/math.2026962.

[15] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1977), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[16] R.M. Neal, G.E. Hinton, A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants, in: M.I. Jordan (Ed.), Learning in Graphical Models, Springer Netherlands, Dordrecht, pp. 355–368, (1998). https://doi.org/10.1007/978-94-011-5014-9_12.

[17] H. Teicher, Identifiability of Finite Mixtures, Ann. Math. Stat. 34 (1963), 1265–1269. https://doi.org/10.1214/aoms/1177703862.

[18] S.J. Yakowitz, J.D. Spragins, On the Identifiability of Finite Mixtures, Ann. Math. Stat. 39 (1968), 209–214. https://doi.org/10.1214/aoms/1177698520.

[19] C. Hennig, Identifiability of Models for Clusterwise Linear Regression, J. Classif. 17 (2000), 273–296. https://doi.org/10.1007/s003570000022.

[20] H. Chen, J. Chen, J.D. Kalbfleisch, A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (2001), 19–29. https://doi.org/10.1111/1467-9868.00273.

[21] T.A. Louis, Finding the Observed Information Matrix When Using the EM Algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 44 (1982), 226–233. https://doi.org/10.1111/j.2517-6161.1982.tb01203.x.