Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Persistent and Transient Efficiency in Four-Component Stochastic Frontier Models: An Integrated Nested Laplace Approximation Approach

Yuheng LING

lyhneocc@gmail.com

Hainan Normal University

Peiyi WU

Hainan Normal University

Lei JIANG

Guangzhou University

Research Article

Keywords: Four-component stochastic frontier model, latent non-Gaussian models, double skew-normal reparameterization, integrated nested Laplace approximation (INLA), persistent and transient technical efficiency

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11729587/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Persistent and Transient Efficiency in Four-Component Stochastic Frontier Models: An Integrated Nested Laplace Approximation Approach

Yuheng LING[*a], Peiyi Wu[a], Lei JIANG[*b]

a. College of Geography and Environmental Science, Key Laboratory of Earth Surface Processes and Environmental Change of Tropical Islands, Hainan Normal University, Haikou, 571158, China

b. School of Geography and Remote Sensing, Center for Human Geography and Urban Development, Guangzhou University, Guangzhou 510006, China

* Corresponding author: Yuheng LING, Lei JIANG. Email: lyhneocc@gmail.com

Abstract

This paper develops an integrated nested Laplace approximation (INLA) based estimator for a four-component stochastic production frontier model that separates time-invariant heterogeneity, persistent inefficiency, statistical noise, and transient inefficiency. Estimation relies on a double skew-normal reparameterization of the model, which combines the four stochastic components into persistent and transient composites. The transient composite is handled through a skew-normal likelihood, while the non-Gaussian persistent composite is represented using a diffuse Gaussian carrier and an auxiliary skew-normal likelihood, yielding a formulation compatible with INLA. Monte Carlo experiments show accurate estimation of model parameters across a wide range of panel dimensions and variance configurations. Recovery of technical efficiency depends strongly on the relative strength of the inefficiency signals, where persistent efficiency is recovered more accurately when persistent inefficiency is large relative to time-invariant heterogeneity, whereas transient efficiency improves as transient inefficiency strengthens relative to statistical noise. Larger panels improve efficiency recovery when persistent and transient inefficiency are sufficiently strong relative to heterogeneity and statistical noise, respectively, but cannot fully resolve weak separation between these components. Computational experiments show that INLA requires less computation time and its runtime increases more slowly with panel size than the alternative estimators considered. Applications to two real-world datasets illustrate differences in the relative importance of persistent versus transient inefficiency.

Keywords: Four-component stochastic frontier model; latent non-Gaussian models; double skew-normal reparameterization; integrated nested Laplace approximation (INLA); persistent and transient technical efficiency.

JEL classification C11 C23 D24

Funding Declaration

This work was supported by the Hainan Provincial Natural Science Foundation [grant number 225QN296]; Guangdong Philosophy and Social Science Foundation [grant number GD25CYJ07].

1. Introduction

Stochastic frontier (SF) models provide a widely used framework for measuring technical efficiency (TE) by decomposing deviations from the production frontier into statistical noise and inefficiency. In panel data, however, persistent differences across production units need not represent inefficiency alone. Time-invariant (or unit) heterogeneity arising from differences in technology, operating environments, or other unobserved characteristics may also shift the frontier. Failure to distinguish such heterogeneity from inefficiency can confound persistent technological differences with long-run productive inefficiency.

The four-component SF model (Colombi et al., 2014) addresses this issue by decomposing the disturbance into time-invariant heterogeneity, persistent inefficiency, statistical noise, and transient inefficiency. This decomposition provides a richer characterization of productive performance by distinguishing long-run from short-run inefficiency while separating both from their corresponding symmetric disturbance. The additional flexibility, however, gives rise to estimation difficulties. The persistent inefficiency term must be separated from time-invariant heterogeneity, and the transient inefficiency term from statistical noise, while Bayesian estimation typically requires inference over a high-dimensional collection of latent variables. These difficulties become increasingly important as the panel dimension grows.

This study develops an estimator for the four-component SF model using the integrated nested Laplace approximation (INLA) methodology introduced by Rue et al. (2009). INLA provides a deterministic alternative to simulation-based Bayesian methods, offering both inferential and computational advantages for the estimation of four-component SF models. To the best of our knowledge, this is its first application to this model. The main methodological contribution includes a double skew-normal reparameterization that combines statistical noise and transient inefficiency into an observation-level skew-normal disturbance and time-invariant heterogeneity and persistent inefficiency into a unit-specific skew-normal random effects component. The former can be incorporated directly through a skew-normal likelihood, whereas the latter remains non-Gaussian and therefore falls outside the conventional latent Gaussian model (LGM) framework required by INLA. We address this issue using a diffuse Gaussian carrier and an auxiliary skew-normal likelihood. The resulting representation

substantially reduces the latent dimension and permits deterministic Bayesian inference without iterative posterior simulations.

The proposed estimator is evaluated through Monte Carlo experiments covering different panel dimensions and configurations of the four stochastic components. The results show that the frontier coefficients are estimated accurately across the designs considered, whereas recovery of the latent efficiency components depends critically on the strength of the corresponding inefficiency signals. Persistent efficiency becomes increasingly recoverable as persistent inefficiency strengthens relative to time-invariant heterogeneity, while transient-efficiency recovery improves as transient inefficiency strengthens relative to statistical noise. Larger panels improve recovery when these signals are sufficiently informative but cannot fully overcome weak inefficiency signals.

We also conduct an experiment comparing INLA with alternative estimators as the panel dimension increases. INLA scales more favorably than the competing estimators, reflecting the reduction in latent dimension and the use of deterministic approximation rather than repeated simulation of latent inefficiency components. Empirical applications to Spanish dairy farms and Brazilian manufacturing firms illustrate the applicability of the four-component SF specification. The results reveal contrasting patterns in the composition of persistent and transient inefficiency across applications.

The remainder of the paper is organized as follows. Section 2 reviews the literature on estimation of four-component SF models. Section 3 develops the double skew-normal representation and the INLA estimator. Section 4 presents the Monte Carlo and computational experiments. Section 5 presents empirical applications to Spanish dairy farms and Brazilian manufacturing firms, while Section 6 concludes the paper.

2. Literature Review

Estimation of the four-component SF model has developed along both frequentist and Bayesian lines. Within the frequentist framework, the estimation has progressed from simple multi-stage procedures to increasingly likelihood-based methods.

Kumbhakar et al. (2014) proposed a sequential estimator that first fits a conventional panel-data model and then estimates the time-invariant and time-varying components separately using

standard SF techniques. Although straightforward to implement, the sequential treatment of the components may allow estimation errors to propagate across stages, resulting in greater bias than likelihood-based estimators.

Colombi et al. (2014) derived a maximum likelihood estimator (MLE) by exploiting the skew-normal distributions of the persistent and transient composite components and the closed skew-normal (CSN) distribution of the overall disturbance. This permits joint estimation of the frontier parameters and all four scale parameters, together with conditional prediction of persistent and transient inefficiency. The computational cost, however, is high since likelihood evaluation involves multivariate normal probabilities whose dimension increases with the panel length, while inference on inefficiency requires additional treatment of nonlinear parameter uncertainty.

More recently, Filippini & Greene (2016) proposed maximum simulated likelihood (SML), replacing high-dimensional integration in the MLE with simulation to simplify optimization and reduce estimation time while retaining joint parameter estimation. Further computation is still needed for inefficiency confidence intervals or standard errors.

Bayesian methods treat heterogeneity and inefficiency terms as latent variables and provide a natural framework for joint posterior inference. E. G. Tsionas & Kumbhakar (2014) develop a data-augmentation Gibbs sampler for the generalized true random-effects (GTRE) specification, introducing two complementary reparameterizations, the δ-parameterization, which combines unit heterogeneity with persistent inefficiency, and the ξ-parameterization, which combines unit heterogeneity with the noise component. An accept-reject step is then used to alleviate the strong posterior dependence among the latent components. Although this approach delivers full posterior inference, conventional data augmentation can experience from slow mixing, particularly in large samples. More recent Markov Chain Monte Carlo methods (MCMC), including particle MCMC (PMCMC) (Tsionas & Mallick, 2019) and Hamiltonian Monte Carlo (HMC) (Ng'ombe & Lambert, 2021) , have been proposed to improve the computational efficiency of posterior sampling.

An alternative to posterior simulation is approximate Bayesian computation (ABC). Ramírez–Hassan et al. (2025) avoid direct likelihood evaluation by drawing parameters from the prior,

generating pseudo-data, and retaining draws for which selected summary statistics, typically residual moments, are sufficiently close to their observed counterparts. ABC is flexible and relatively straightforward to implement, but its accuracy depends on the choice of summary statistics and tolerance level and may deteriorate as the dimension of the approximation problem increases. In their application, performance also weakens at high signal-to-noise ratios (denoted by λ, see Subsection 4.1).

A different approach to Bayesian inference relies on deterministic approximations rather than posterior simulation. One example is variational Bayes (VB). Hajargasht & Griffiths (2018) develop VB methods for estimating stochastic frontier models by approximating posterior inference as an optimization problem that minimizes the Kullback-Leibler divergence between a tractable approximating distribution and the target posterior. Although VB offers computational advantages over simulation-based methods, its accuracy depends on the choice of variational family (Blei et al., 2017). In particular, mean-field VB may fail to capture posterior dependence among latent components, potentially leading to underestimated posterior uncertainty (Giordano et al., 2018). This limitation is particularly relevant to four-component stochastic frontier models, where heterogeneity and inefficiency components may be strongly dependent on the posterior.

Against this background, the proposed INLA-based approach provides another deterministic Bayesian approximation method by incorporating the non-Gaussian latent components through auxiliary skew-normal likelihoods and a Gaussian carrier. This reparameterization enables posterior marginals for model parameters and latent effects to be computed using Laplace approximations and numerical integration. Unlike MCMC methods, INLA offers several computational advantages: (1) it does not require iterative posterior sampling; (2) it avoids the convergence and mixing issues associated with Markov chain simulation; and (3) it avoids the loss of sampling efficiency caused by serially correlated posterior draws. Together, these features can significantly reduce the computational burden of estimation for four-component SF models.

3. Model and Estimation

This section develops the INLA-based estimator. We first combine the four stochastic

components into two normal–half-normal composite terms, yielding a double skew-normal reparameterization that separates the persistent and transient dimensions of the four stochastic components. This substantially reduces the latent dimension, but the unit-specific composite component remains non-Gaussian and hence falls outside the conventional LGM framework. We address this difficulty using a diffuse Gaussian carrier and an auxiliary skew-normal likelihood, thereby retaining a latent Gaussian field while approximating the centered skew-normal distribution of the persistent composite component. Finally, we describe the INLA estimation procedure and the posterior transformations used to recover the original four components.

3.1. Four-component SF model

Consider the panel-data four-component stochastic production frontier model

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{0i} - u_{0i} + v_{it} - u_{it}, \qquad i = 1, \dots, N, \quad t = 1, \dots, T, \tag{1}$$

where y_{it} denotes the logarithm of output, $\boldsymbol{x}_{it}$ is a $K \times 1$ vector containing the logarithm of production inputs and possibly other observed covariates, and $\boldsymbol{\beta}$ is the corresponding vector of fronter coefficients.

The four stochastic components are assumed to be mutually independent, with

$$v_{0i} \sim \mathcal{N}(0, \sigma_{v0}^2), \qquad u_{0i} \sim \mathcal{N}^+(0, \sigma_{u0}^2), \qquad v_{it} \sim \mathcal{N}(0, \sigma_v^2), \qquad u_{it} \sim \mathcal{N}^+(0, \sigma_u^2), \tag{2}$$

where $\mathcal{N}^+(0, \sigma_{(\cdot)}^2)$ denotes a normal distribution truncated below at zero. All stochastic components are also assumed to be independent of the regressors.

The four components have distinct interpretations. The term v_{0i} captures time-invariant unit heterogeneity arising from unobserved characteristics that may shift the production frontier, while u_{0i} represents persistent, or long-run, technical inefficiency. At the observation level, v_{it} is idiosyncratic noise and u_{it} represents transient, or short-run, technical inefficiency. The specification thus separates unit heterogeneity from persistent inefficiency and statistical noise from transient inefficiency.

This decomposition follows the GTRE specification. Its main advantage is that the two time-invariant components are given distinct interpretations. A conventional random-effects SF model may absorb persistent inefficiency into unobserved heterogeneity, whereas treating the

entire time-invariant component as inefficiency may confound heterogeneity with long-run inefficiency.

TE therefore has persistent (PTE) and transient (RTE) parts,

$$PTE_i = \exp(-u_{0i}), \qquad (3)$$

and

$$RTE_{it} = \exp(-u_{it}). \qquad (4)$$

Overall technical efficiency (OTE) is

$$OTE_{it} = \exp[-(u_{0i} + u_{it})] = PTE_i \times RTE_{it}. \qquad (5)$$

Estimation of the model therefore requires not only inference on $\boldsymbol{\beta}$, but also sufficient separation of the four stochastic components to recover persistent and transient TE.

3.2. Double skew-normal reparameterization

Rather than applying Bayesian data augmentation to handle the four stochastic components separately (E. G. Tsionas & Kumbhakar, 2014), we group the time-invariant and time-varying components into two composite terms. Stacking the T observations for firm i, Eq. (1) can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + (v_{0i} - u_{0i})\mathbf{1}_T + (\boldsymbol{v}_i - \boldsymbol{u}_i) \equiv \boldsymbol{X}_i\boldsymbol{\beta} + \xi_i\mathbf{1}_T + \boldsymbol{\varepsilon}_i, \qquad (6)$$

where

$$\xi_i = v_{0i} - u_{0i}, \qquad \boldsymbol{\varepsilon}_i = \boldsymbol{v}_i - \boldsymbol{u}_i.$$

Here $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})'$, $\boldsymbol{X}_i$ is the corresponding $T \times K$ regressor matrix, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})'$, and $\mathbf{1}_T$ is a $T \times 1$ vector of ones.

Under the distributional assumptions in Eq. (2), both composite terms follow skew-normal distributions (Azzalini, 2005). Their probability density functions are

$$f_\varepsilon(\varepsilon) = \frac{2}{\omega_\varepsilon}\phi\left(\frac{\varepsilon}{\omega_\varepsilon}\right)\Phi\left(-\lambda\frac{\varepsilon}{\omega_\varepsilon}\right), \quad f_\xi(\xi) = \frac{2}{\omega_\xi}\phi\left(\frac{\xi}{\omega_\xi}\right)\Phi\left(-\lambda_0\frac{\xi}{\omega_\xi}\right), \quad (7)$$

where

$$\omega_\varepsilon^2 = \sigma_v^2 + \sigma_u^2, \qquad \lambda = \frac{\sigma_u}{\sigma_v},$$

and

$$\omega_\xi^2 = \sigma_{v_0}^2 + \sigma_{u_0}^2, \qquad \lambda_0 = \frac{\sigma_{u_0}}{\sigma_{v_0}}.$$

Eqs. (6) and (7) define the key double skew-normal reparameterization. The transient

component pair (v_{it}, u_{it}) is represented by the observation-level skew-normal disturbance $\boldsymbol{\varepsilon}_i$, while the time-invariant pair (v_{0i}, u_{0i}) is represented by the unit-specific skew-normal disturbance ξ_i. The four original scale parameters remain identified through the scale and skewness of these two composite terms.

The reparameterization also changes the computational structure of the model. Direct data augmentation would introduce separate NT-dimensional transient inefficiency and statistical noise, together with two sets of N-dimensional time-invariant components (Tsionas & Kumbhakar, 2014). Under Eq. (6), the transient components are absorbed into the skew-normal observation-level disturbance, while the two time-invariant components are replaced by only one N-dimensional persistent composite component. Thus, explicit treatment of the four original latent components is avoided during estimation.

The reduction in latent dimension does not, however, eliminate the difficulties associated with simulation-based Bayesian inference. The persistent composite component must still be estimated jointly with the frontier and distributional parameters, potentially generating strong posterior dependence and highly autocorrelated MCMC draws. As discussed by Tsionas & Kumbhakar (2014) , such dependence can lead to slow mixing and inefficient posterior exploration, motivating tailored reparameterizations and sampling schemes. In this respect, the double skew-normal representation is particularly suitable for INLA, which uses deterministic Laplace approximations and numerical integration rather than simulation. Consequently, the proposed approach avoids MCMC-specific convergence and mixing issues while benefiting from the reduced latent dimension.

For implementation, we use the mean-centered representation of the SF specification. Eq. (6) is therefore written as (see Appendix B.1 for the derivation),

$$\boldsymbol{y}_i = \beta_0^m \mathbf{1}_T + \widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{\beta}} + \xi_i^c \mathbf{1}_T + \boldsymbol{\varepsilon}_i^c, \qquad (8)$$

where ξ_i^c and each element of $\boldsymbol{\varepsilon}_i^c$ have zero mean. Eq. (8) is the parameterization used in the subsequent INLA formulation.

Although Eq. (8) is simpler than the GTRE specification, it is not yet an LGM. The vector of persistent composite effects,

$$\boldsymbol{\xi}^c = (\xi_1^c, \dots, \xi_N^c)',$$

is non-Gaussian. Under the independence assumptions in Eq. (2), its joint density factorizes as

$$f_{\xi}(\boldsymbol{\xi}^c) = \prod_{i=1}^{N} f_{\xi}\,(\xi_i^c),$$

where each ξ_i^c follows a skew-normal distribution[1]. The double skew-normal SF specification therefore remains outside the standard LGM framework (Martins and Rue, 2014; Cabral et al., 2024) and motivates the Gaussian-carrier approximation introduced in the next subsection.

3.3 From a latent non-Gaussian model to a diffuse Gaussian-carrier representation

The transient composite disturbance ε_{it}^c can be accommodated directly through the observation likelihood, since INLA permits non-Gaussian likelihoods provided that the latent field is conditionally Gaussian. In contrast, the persistent composite component ξ_i^c enters the linear predictor as a latent effect but follows a non-Gaussian distribution. Let

$$\boldsymbol{w} = (\boldsymbol{\beta}', \xi_1^c, \dots, \xi_N^c)'$$

and denote the desired centered skew-normal density of the persistent composite component $f_{\xi}\left(\xi_i^c \mid \boldsymbol{\theta}_{\xi}\right)$. Ignoring constants, the posterior kernel of the reparametrized model is

$$\pi(\boldsymbol{\beta}, \boldsymbol{\xi}^c, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{N} f_{\varepsilon}\left(\boldsymbol{y}_i - \beta_0^m \mathbf{1}_T - \widetilde{\mathbf{X}}_i\widetilde{\boldsymbol{\beta}} - \boldsymbol{\xi}^c \mid \boldsymbol{\theta}_{\varepsilon}\right) \times \prod_{i=1}^{N} f_{\xi}\left(\xi_i^c \mid \boldsymbol{\theta}_{\xi}\right) \pi(\boldsymbol{\beta})\, \pi(\boldsymbol{\theta}),$$

where

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}_{\varepsilon}, \boldsymbol{\theta}_{\xi}\right).$$

The second product specifies the non-Gaussian distribution of the persistent composite component, which prevents the model from satisfying the standard LGM assumptions required by INLA. Martins & Rue (2014) develop a correction-term approach for extending INLA to models containing near-Gaussian latent components. Their basic tool is to introduce a Gaussian approximation to the non-Gaussian latent distribution and correct the approximation through the ratio of the target density to the Gaussian density. They explicitly consider distributions that add skewness or kurtosis around a Gaussian benchmark. For example, introducing a Gaussian carrier $g_c(\xi_i^c) = N(\xi_i^c; 0; \tau_c^{-1})$, the target density can be factored as

[1] This independent skew-normal construction can also be viewed as a special case of the multivariate CSN family.

$$f_{\xi}\left(\xi_i^c \mid \boldsymbol{\theta}_{\xi}\right) = g_c(\xi_i^c)\,\frac{f_{\xi}\left(\xi_i^c \mid \boldsymbol{\theta}_{\xi}\right)}{g_c\left(\xi_i^c\right)}.$$

Define the exact correction term

$$CT_i\left(\xi_i^c, \boldsymbol{\theta}_{\xi}\right) = \frac{f_{\xi}\left(\xi_i^c \mid \boldsymbol{\theta}_{\xi}\right)}{g_c\left(\xi_i^c\right)}.$$

Then, the Gaussian factor can be assigned to the latent layer, while the correction term is moved into an augmented first stage.

Our approach adopts a computationally simpler approximation to the exact correction construction. We replace the exact density-ratio correction of Martins & Rue (2014) with a highly diffuse Gaussian carrier (prior) and an auxiliary skew-normal likelihood. The resulting representation preserves a Gaussian latent field while avoiding explicit evaluation of the correction term. Specifically, the persistent composite component is assigned the diffuse Gaussian prior[2],

$$\xi_i^c \sim N(0, \tau_c^{-1}), \tau_c = 10^{-6},$$

while its desired skew-normal shape is imposed through a second, auxiliary skew-normal likelihood.

Martins and Rue's (2014) approach employs a density-ratio correction, which involves evaluating the ratio of the skew-normal density to the Gaussian density for each latent effect during INLA's numerical approximation. Implementing this correction requires an additional likelihood component within the INLA framework[3]. However, evaluating the density ratio may become numerically challenging when the Gaussian carrier is highly diffuse. The proposed approximation avoids explicit evaluation of the density ratio while retaining the same latent dimension.

To implement the approximation, we also augment the actual equation with an auxiliary

[2] The prior variance is fixed at 10^6.

[3] In R-INLA, the *inla.stack* function is used to organize and combine multiple likelihood contributions, their associated observations, and latent effects into a unified model representation. The density-ratio correction can be incorporated as an additional likelihood block in this stacked representation, provided that an appropriate likelihood implementation is available.

likelihood whose purpose is to encode the target distribution of the persistent composite component. This auxiliary likelihood is not associated with any observed outcome; it is a computational device that provides a second likelihood channel for ξ_i^c. This idea is analogous to the zero trick used in MCMC (Lunn et al., 2012), where pseudo-observations are introduced solely to impose a custom distribution within a standard likelihood.

Specifically, one artificial observation is constructed for each unit

$$z_i = 0, i = 1, \dots, N,$$

with linear predictor

$$\mu_i^{(\xi)} = -\xi_i^c.$$

It follows that

$$z_i - \mu_i^{(\xi)} = \xi_i^c.$$

Assigning a centered skew-normal likelihood to this pseudo-observation therefore contributes exactly the desired density. Since $z_i = 0$,

$$f_{SN_c}\left(z_i, \mid \mu_i^{(\xi)}, \boldsymbol{\theta}_\xi\right) = f_{SN_c}\left(z_i - \mu_i^{(\xi)} \mid 0, \boldsymbol{\theta}_\xi\right) = f_{SN_c}\left(\xi_i^c, \mid 0, \boldsymbol{\theta}_\xi\right) = f_\xi\left(\xi_i^c \mid \boldsymbol{\theta}_\xi\right).$$

The auxiliary likelihood, therefore, contributes the target skew-normal density $f_\xi\left(\xi_i^c \mid \theta_\xi\right)$ to the posterior kernel. Although the pseudo-observation is zero, this does not imply that ξ_i^c should be close to zero. The centered skew-normal likelihood has mean zero after correcting for skewness, but its shape, including skewness and scale, is imposed on the latent effect. The pseudo-observation therefore does not shrink ξ_i^c toward zero as a Gaussian likelihood would, it instead reproduces the desired skew-normal distribution.

The operational representation is finally

$$\begin{aligned} \boldsymbol{y}_i \mid \xi_i^c, \boldsymbol{\theta}_\varepsilon &= \beta_0^m \mathbf{1}_T + \widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{\beta}} + \xi_i^c \mathbf{1}_T + \boldsymbol{\varepsilon}_i^c, & \varepsilon_{it}^c &\sim SN_c(\boldsymbol{\theta}_\varepsilon), \\ z_i \mid \xi_i^c, \boldsymbol{\theta}_\xi &\sim SN_c\left(-\xi_i^c, \boldsymbol{\theta}_\xi\right), & z_i &= 0, \\ \xi_i^c &\sim N(0, \tau_c^{-1}), & \tau_c &= 10^{-6}. \end{aligned}$$

In summary, the first skew-normal likelihood identifies the distribution of the transient composite disturbance; the second supplies the scale and skewness information for the persistent composite component, and the diffuse Gaussian carrier permits ξ_i^c to remain inside the latent Gaussian field.

Since τ_c is fixed, the normalizing constant of g_c is independent of the unknown model parameters and can be omitted from the posterior kernel. The only additional shape contribution induced by the carrier is then

$$g_c(\xi_i^c) \propto \exp\left[-\frac{\tau_c}{2}(\xi_i^c)^2\right].$$

Across all N units, the additional carrier contribution is

$$C_{\tau_c}(\xi^c) = \exp\left[-\frac{\tau_c}{2}\sum_{i=1}^{N}(\xi_i^c)^2\right].$$

As $\tau_c = 10^{-6}$, this factor is extremely close to 1 for all plausible values of ξ_i^c[4]. Thus, the carrier imposes only a negligible quadratic penalty to the log posterior, not substantive shrinkage. As $\tau_c \to 0$,

$$C_{\tau_c}(\boldsymbol{\xi}^c)\prod_{i=1}^{N} f_\xi\left(\xi_i^c \mid \boldsymbol{\theta}_\xi\right) \to \prod_{i=1}^{N} f_\xi\left(\xi_i^c \mid \boldsymbol{\theta}_\xi\right).$$

The approximate posterior kernel therefore approaches the target skew-normal latent specification. The carrier acts as a very weak Gaussian penalty whose strength is entirely controlled by τ_c. Appendix B.2 provides an analytical calibration and a Monte Carlo sensitivity analysis.

3.4 INLA estimation procedure

Conditional on the diffuse Gaussian-carrier, the four-component SF model is written as a three-level Bayesian hierarchical model.

Level 1: likelihood

Let $\boldsymbol{\beta}^m = \left(\beta_0^m, \widetilde{\boldsymbol{\beta}}'\right)'$ collect the frontier coefficients in the mean-centered representation. For the observed production outcomes,

$$y_{it} \mid \boldsymbol{\beta}, \xi_i^c, \boldsymbol{\theta}_\varepsilon \sim SN_c(\mu_{it}, \tau_\varepsilon, \gamma_\varepsilon) \quad i = 1, \dots, N, \quad t = 1, \dots, T,$$

where

$$\mu_{it} = \beta_0^m + \tilde{x}_{it}'\widetilde{\boldsymbol{\beta}} + \xi_i^c$$

and $\boldsymbol{\theta}_\varepsilon = (\tau_\varepsilon, \gamma_\varepsilon)$. For each unit, the auxiliary pseudo-observations satisfy

$$z_i = 0 \mid \xi_i^c, \boldsymbol{\theta}_\xi \sim SN_c\left(-\xi_i^c, \tau_\xi, \gamma_\xi\right),$$

[4] For example, even at $\mid \xi_i^c \mid = 100$, the per-observation factor is $\exp(-5 \times 10^{-3}) \approx 0.995$.

where $\boldsymbol{\theta}_\xi = (\tau_\xi, \gamma_\xi)$. $SN_c(\mu, \tau, \gamma)$ denotes the mean-parameterized skew-normal distribution with mean μ, precision τ, and standardized skewness γ. This corresponds to the parameterization of the skew-normal likelihood in R-INLA, for which the linear predictor is the conditional mean and γ is the standardized third central moment. Thus, $(\tau_\varepsilon, \gamma_\varepsilon)$ characterize the transient composite distribution, whereas (τ_ξ, γ_ξ) characterize the persistent composite distribution.

Let

$$\boldsymbol{Y}^* = (\mathrm{y}'_{it}, 0'_N)'$$

denote the augmented response vector containing the NT observed production outcomes and the N auxiliary zero observations. Conditional independence then gives the augmented joint likelihood

$$\pi(\boldsymbol{Y}^*|\boldsymbol{w}, \boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{t=1}^{T} f_{SN_c}(y_{it}|\mu_{it}, \tau_\varepsilon, \gamma_\varepsilon) \times \prod_{i=1}^{N} f_{SN_c}\left(0\middle|-\xi_i^c, \tau_\xi, \gamma_\xi\right). \qquad (9)$$

Level 2: Latent field

The latent field is

$$\boldsymbol{w} = [(\boldsymbol{\beta}^m)', (\boldsymbol{\xi}^c)']'. \qquad (10)$$

The regression coefficients are assigned to vague Gaussian priors,

$$\boldsymbol{\beta}^m \sim N_p(\mathbf{0}, 10^4 \boldsymbol{I}_p),$$

where p is the number of regression coefficients. The persistent composite component is assigned the diffuse Gaussian carrier introduced in Section 3.3,

$$\boldsymbol{\xi}^c \sim \mathcal{N}_N(\mathbf{0}, \tau_c^{-1}\boldsymbol{I}_N), \qquad \tau_c = 10^{-6}.$$

Hence,

$$\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{Q}^{-1}),$$

with block-diagonal precision matrix

$$\boldsymbol{Q} = \begin{pmatrix} 10^{-4}\boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \tau_c \boldsymbol{I}_N \end{pmatrix}.$$

Level 3: Hyperpriors

The hyperparameter vector is

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_\varepsilon, \boldsymbol{\theta}_\xi) = (\tau_\varepsilon, \gamma_\varepsilon, \tau_\xi, \gamma_\xi). \qquad (11)$$

For $j \in \{\varepsilon, \xi\}$, let

$$s_j = \tau_j^{-1/2}$$

denote the standard deviation of the corresponding centered composite distribution. We assign a penalized-complexity prior to s_j, calibrated by $P(s_j > 1) = 0.01$. The penalized-complexity prior penalizes deviation from a simpler base model through an interpretable tail-probability calibration (see Simpson et al. (2017) for discussion). The standardized skewness parameters and are assigned penalized-complexity priors centered on the Gaussian base model ($\varphi = 0$), with rate parameter $\lambda_\varphi = 10$. The Gaussian-carrier precision τ_c is fixed and is therefore not included in $\boldsymbol{\theta}$.

Under equations (9)-(11), the joint posterior distribution is

$$\pi(\boldsymbol{w}, \boldsymbol{\theta} | \boldsymbol{Y}^*) \propto \pi(\boldsymbol{Y}^* | \boldsymbol{w}, \boldsymbol{\theta})\, \pi(\boldsymbol{w})\, \pi(\boldsymbol{\theta}).$$

The latent field $\boldsymbol{w}$ is a high-dimensional Gaussian field, whereas the hyperparameter vector $\boldsymbol{\theta}$ is only four-dimensional. Such a hierarchical structure is well suited to INLA.

INLA obtains posterior marginals through nested deterministic approximations. For each value of $\boldsymbol{\theta}$, a Gaussian approximation is first constructed around the mode of $\pi(\boldsymbol{w}, \boldsymbol{\theta} | \boldsymbol{Y}^*)$. Denote this approximation by $\tilde{\pi}_G(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{Y}^*)$, the marginal posterior of the hyperparameters is then calculated by

$$\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y}^*) \propto \left. \frac{\pi(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{Y}^*)}{\tilde{\pi}_G(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{Y}^*)} \right|_{\boldsymbol{w} = \boldsymbol{w}^*(\boldsymbol{\theta})}.$$

where $\boldsymbol{w}^*(\boldsymbol{\theta})$ denotes the conditional mode. The marginal posterior of each latent component w_k is subsequently obtained by numerical integration,

$$\tilde{\pi}(w_k \mid \boldsymbol{Y}^*) = \int \tilde{\pi}\,(w_k | \boldsymbol{\theta}, \boldsymbol{Y}^*)\, \tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y}^*)\, d\boldsymbol{\theta}.$$

The resulting posterior marginals provide inference for the mean-centered coefficients, the persistent composite effects, and the hyperparameters of the two skew-normal distributions. Posterior transformation to the original four stochastic components is described in the following subsection.

3.5 Recovery of scale parameters and TE

Let γ_j denote the standardized skewness. The bounded skew-normal shape parameter δ_j is

obtained from γ_j using the inverse mapping given in Appendix C.1. The scale parameter of the corresponding uncentered skew-normal distribution is then

$$\omega_j^2 = \frac{s_j^2}{1 - 2\delta_j^2/\pi},$$

For the transient composite disturbance,

$$\sigma_u = \omega_\varepsilon \mid \delta_\varepsilon \mid, \qquad \sigma_v = \omega_\varepsilon \sqrt{1 - \delta_\varepsilon^2},$$

whereas for the persistent composite component,

$$\sigma_{u0} = \omega_\xi |\delta_\xi|, \qquad \sigma_{v0} = \omega_\xi \sqrt{1 - \delta_\xi^2}.$$

These transformations are applied to each joint posterior draw. The unconditional means of the two uncentered composite distributions are

$$m_\varepsilon = -\sigma_u \sqrt{\frac{2}{\pi}}, \qquad m_\xi = -\sigma_{u_0} \sqrt{\frac{2}{\pi}},$$

so that the original intercept is recovered as

$$\beta_0 = \beta_0^m - m_\varepsilon - m_\xi.$$

Similarly, the uncentered persistent composite effect is recovered from

$$\xi_i = \xi_i^c + m_\xi,$$

and the transient composite effect is

$$\varepsilon_{it} = y_{it} - \boldsymbol{x}_{it}' \mathbf{B} - \xi_i,$$

where $\mathbf{B}$ denotes the recovered coefficient vector in the original SF model, including β_0. Persistent and transient inefficiency are recovered conditionally on the composite components. Without loss of generality, consider a normal–half-normal composite component $z = g - h$, with $g \sim N(0, \sigma_g^2)$ and $h \sim N^+(0, \sigma_h^2)$. As shown in Appendix C.2, the conditional distribution $h \mid z$ is truncated normal. Its conditional mean provides the conventional inefficiency predictor of Jondrow et al. (1982). For the persistent component, we have

$$z = \xi_i, \qquad g = v_{0i}, \qquad h = u_{0i},$$

whereas for the transient disturbance,

$$z = \varepsilon_{it}, \qquad g = v_{it}, \qquad h = u_{it}.$$

Hence, the same conditional decomposition is applied separately to the persistent and transient

composite component.

TE is obtained from the conditional expectation of $\exp(-h)$ rather than by exponentiating the conditional mean of h[5]. For each joint posterior draw $s = 1, \dots, S$, this yields PTE and RTE predictions,

$$PTE_i^{(s)} = E\left[\exp(-u_{0i}) \mid \xi_i^{(s)}, \boldsymbol{\theta}^{(s)}\right],$$

and

$$RTE_{it}^{(s)} = E\left[\exp(-u_{it}) \mid \varepsilon_{it}^{(s)}, \boldsymbol{\theta}^{(s)}\right].$$

Posterior-averaged TEs are then

$$\widehat{PTE}_i = \frac{1}{S}\sum_{s=1}^{S} PTE_i^{(s)}, \qquad \widehat{RTE}_{it} = \frac{1}{S}\sum_{s=1}^{S} RTE_{it}^{(s)},$$

and

$$\widehat{OTE}_{it} = \frac{1}{S}\sum_{s=1}^{S} PTE_i^{(s)}\, RTE_{it}^{(s)}.$$

Posterior-averaged OTE is then constructed by multiplying PTE and RTE (see Eq (5)) within each joint posterior draw before averaging, preserving posterior dependence generated by common parameter uncertainty.

4. Simulations

We conduct Monte Carlo experiments to examine the finite-sample performance of the proposed INLA approach for the four-component SF model. Following Badunenko & Kumbhakar (2016), the experiments focus on different configurations of the scale parameters of time-invariant heterogeneity, persistent inefficiency, statistical noise, and transient inefficiency, yielding 16 disturbance-scale configurations. The estimator's performance is evaluated in terms of both parameter estimation and recovery of PTE, RTE and OTE. The experiments are conducted in R using the packages R-INLA (Lindgren & Rue, 2015).

4.1. Design of the experiment

We first describe the basic design of the experiment, including the data-generating process,

[5] The relevant closed-form expression is also given in Appendix C and corresponds to the conditional inefficiency prediction approach of Battese & Coelli (1988).

distributional assumptions, and other parameters used in the simulation. The production process employs two inputs, x_1 and x_2, to produce a single output, q. Following Badunenko & Kumbhakar (2016), we consider the Cobb-Douglas production function

$$q = \exp(b) x_1^{\alpha} x_2^{\gamma-\alpha}.$$

We assume constant returns to scale, $\gamma = 1$, and set $b = 0.3$, $\alpha = 0.4$. For unit i at time t, observed output in levels is generated as

$$q_{it} = \exp(b) x_{1it}^{\alpha} x_{2it}^{\gamma-\alpha} \exp\ (v_{0i} - u_{0i} + v_{it} - u_{it}).$$

Defining $y_{it} = \ln q_{it}$, the data-generating process is written as

$$y_{it} = b + \alpha \ln x_{1it} + (\gamma - \alpha) \ln x_{2it} + v_{0i} - u_{0i} + v_{it} - u_{it}.$$

In the experiment, each of σ_{v0}, σ_{u0}, σ_v and σ_u takes either 0.04 or 0.2. The resulting combinations are reported in Table 1.

Table 1. Combinations of λ and σ values used in the Monte Carlo simulations

Scenario	σ_{u0}	σ_u	σ_{v0}	σ_v	λ_0	λ	Λ
s1	0.04	0.04	0.04	0.04	1	1	1
s2	0.04	0.20	0.04	0.04	1	5	0.2
s3	0.20	0.04	0.04	0.04	5	1	5
s4	0.20	0.20	0.04	0.04	5	5	1
s5	0.04	0.04	0.04	0.20	1	0.2	1
s6	0.04	0.20	0.04	0.20	1	1	0.2
s7	0.20	0.04	0.04	0.20	5	0.2	5
s8	0.20	0.20	0.04	0.20	5	1	1
s9	0.04	0.04	0.20	0.04	0.2	1	1
s10	0.04	0.20	0.20	0.04	0.2	5	0.2
s11	0.20	0.04	0.20	0.04	1	1	5
s12	0.20	0.20	0.20	0.04	1	5	1
s13	0.04	0.04	0.20	0.20	0.2	0.2	1
s14	0.04	0.20	0.20	0.20	0.2	1	0.2
s15	0.20	0.04	0.20	0.20	1	0.2	5
s16	0.20	0.20	0.20	0.20	1	1	1

The scenarios are summarized by the following three ratios, $\lambda_0 = \frac{\sigma_{u0}}{\sigma_{v0}}, \lambda = \frac{\sigma_u}{\sigma_v}, \Lambda = \frac{\sigma_{u0}}{\sigma_u}$.

Here, λ_0 measures persistent inefficiency relative to individual heterogeneity, λ measures transient inefficiency relative to idiosyncratic noise, and Λ measures persistent inefficiency relative to transient inefficiency. Table 2 organizes the 16 scenarios according to λ_0 and λ.

Table 2. Combinations of λ_0 and λ values used in the simulation scenarios

	λ=0.2	λ=1	λ=5
λ_0=0.2	$s13$	$s9$, $s14$	$s10$
λ_0=1	$s5$, $s15$	$s1$, $s6$, $s11$, $s16$	$s2$, $s12$
λ_0=5	$s7$	$s3$, $s8$	$s4$

Low values of λ_0 indicate that persistent inefficiency is small relative to individual heterogeneity, whereas high values indicate the opposite. Similarly, low values of λ correspond to transient inefficiency being small relative to idiosyncratic noise, while high values indicate that transient inefficiency dominates noise. Hence, $s13$, with $\lambda_0 = \lambda = 0.2$, represents the case in which both inefficiency components are relatively weak, whereas $s4$, with $\lambda_0 = \lambda = 5$, represents the opposite extreme. The ratio Λ further distinguishes scenarios sharing the same values of λ_0 and λ.

For each replication, the inputs are generated from

$$f_b(x_{(\cdot)it}) = (b-1)^{-1}\exp[\ln(b) - x_{(\cdot)it}], \qquad 0 < x_{(\cdot)it} < \log(b),$$

with $b = 2$ for x_{1it} and $b = 10$ for x_{2it}.

We extend the original design of Badunenko & Kumbhakar (2016) by considering $N \in \{100,500,1000\}$ and $T \in \{10,20,40\}$. Lastly, each design is replicated $R = 1000$ times. For every simulated dataset, the same model specification and prior settings described in Subsection 3.2-3.4 are used, and 1000 joint posterior draws are obtained for posterior inference.

4.2. Comparative measures

We evaluate the proposed estimator in terms of frontier parameter estimation and technical TE recovery. For frontier parameters, we report relative bias (RB), root mean squared error (RMSE), and the empirical coverage probability of the 95% posterior credible intervals ($CP_{0.95}$). For TE recovery, we follow the four-component SF literature and assess relative bias and the Pearson correlation coefficient (PC). Let TE_{jr} and $\widehat{TE}_{jr}$ denote the true and estimated TE for observation j in replication r, respectively. Relative bias is calculated as

$$RB(TE) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{M}\sum_{j=1}^{M}\frac{\widehat{TE}_{jr} - TE_{jr}}{TE_{jr}},$$

where $M = N$ for persistent TE and $M = NT$ for transient and overall TE. The Pearson correlation coefficient is computed between the true and estimated TE values within each replication and then averaged across replications. These measures assess systematic estimation

bias and the ability to reproduce cross-sectional differences in efficiency.

To examine estimation accuracy across the TE distribution, we additionally calculate relative deviations at the 5th, 50th, and 95th percentiles. For each replication, observations are ordered according to their true TE values, and the estimated TE is matched to the same observation. The percentile-specific relative deviation is defined as

$$D_p(TE) = \frac{1}{R}\sum_{r=1}^{R} \frac{\widehat{TE}_{r,(p)} - TE_{r,(p)}}{TE_{r,(p)}}, p \in \{0.05, 0.50, 0.95\},$$

where $TE_{r,(p)}$ denotes the true TE at percentile p and $\widehat{TE}_{r,(p)}$ is the estimated TE for the corresponding observation. These measures assess estimation accuracy at the lower tail, median, and upper tail of the TE distribution.

4.3. Simulation results for frontier parameter estimates

We first examine the finite-sample performance of the frontier parameter estimates. Figure 1 reports the relative bias, RMSE, and coverage probability for the estimated coefficients of x_1 and x_2. For both coefficients, relative bias is negligible across the 16 scenarios and panel dimensions and shows no systematic relationship with λ_0, λ, or Λ. Moreover, relative biases are generally of order 10^{-4} and exhibit no systematic sign pattern. The estimates therefore show little evidence of finite-sample bias across the different scale configurations.

The RMSE results are more informative. The RMSE of estimated coefficients decreases markedly as either N or T increases, falling from about 0.006-0.007 in some of the smallest-panel designs to around 0.001 or below in the largest panels. The similarity of this pattern for both x_1 and x_2 indicates that it is a general feature of coefficient estimation rather than a regressor-specific result.

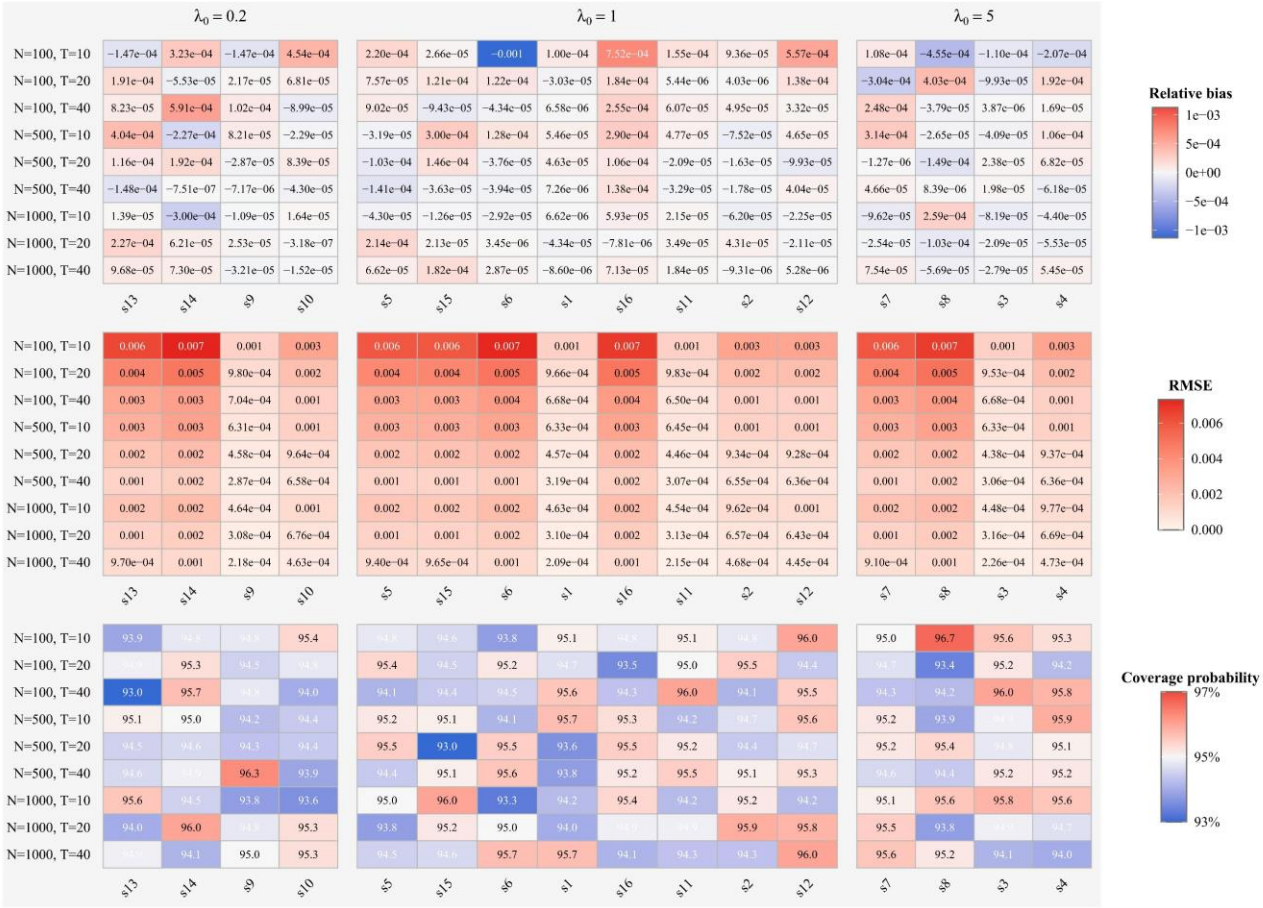
There is, however, noticeable variation in RMSE across scenarios even when N and T are held fixed. The precision of parameter estimates is more sensitive to the magnitude of the transient disturbance than to the persistent component. The same RMSE pattern is repeated across the three λ_0 blocks. Scenarios in which both time-varying components have small variation, such as s9, s1, s11, and s3, produce the smallest RMSEs. Precision deteriorates when σ_u or σ_v increases, and is lowest when both are relatively large, as in s14, s6, s16, and s8. This pattern is economically intuitive. Greater observation-specific disturbance variation

reduces the information available for estimating the systematic part of the frontier, even though it does not introduce appreciable bias. Importantly, changing λ_0 does not generate a comparable deterioration in the fixed-effect estimates.

Coverage probabilities are also stable across the designs. For both coefficients, empirical coverage of the nominal 95% credible intervals remains close to 95%, ranging approximately from 92.7% to 96.7% across all designs. There is no persistent tendency toward either under-coverage or over-coverage as λ_0, λ, N, or T changes. Moreover, the configurations associated with relatively large RMSE do not exhibit corresponding failures in coverage. Thus, the posterior uncertainty estimates appropriately reflect the loss of precision in the more variable designs.

In summary, the results provide strong evidence that estimation of structural parameter coefficients is highly robust across the designs considered. Across the full range of scale configurations, x_1 and x_2 exhibit negligible bias, decreasing RMSE as the panel expands, and coverage close to its nominal level.

Panel A Performance measures: x_1 coefficient



Panel B Performance measures: x_2 coefficient.

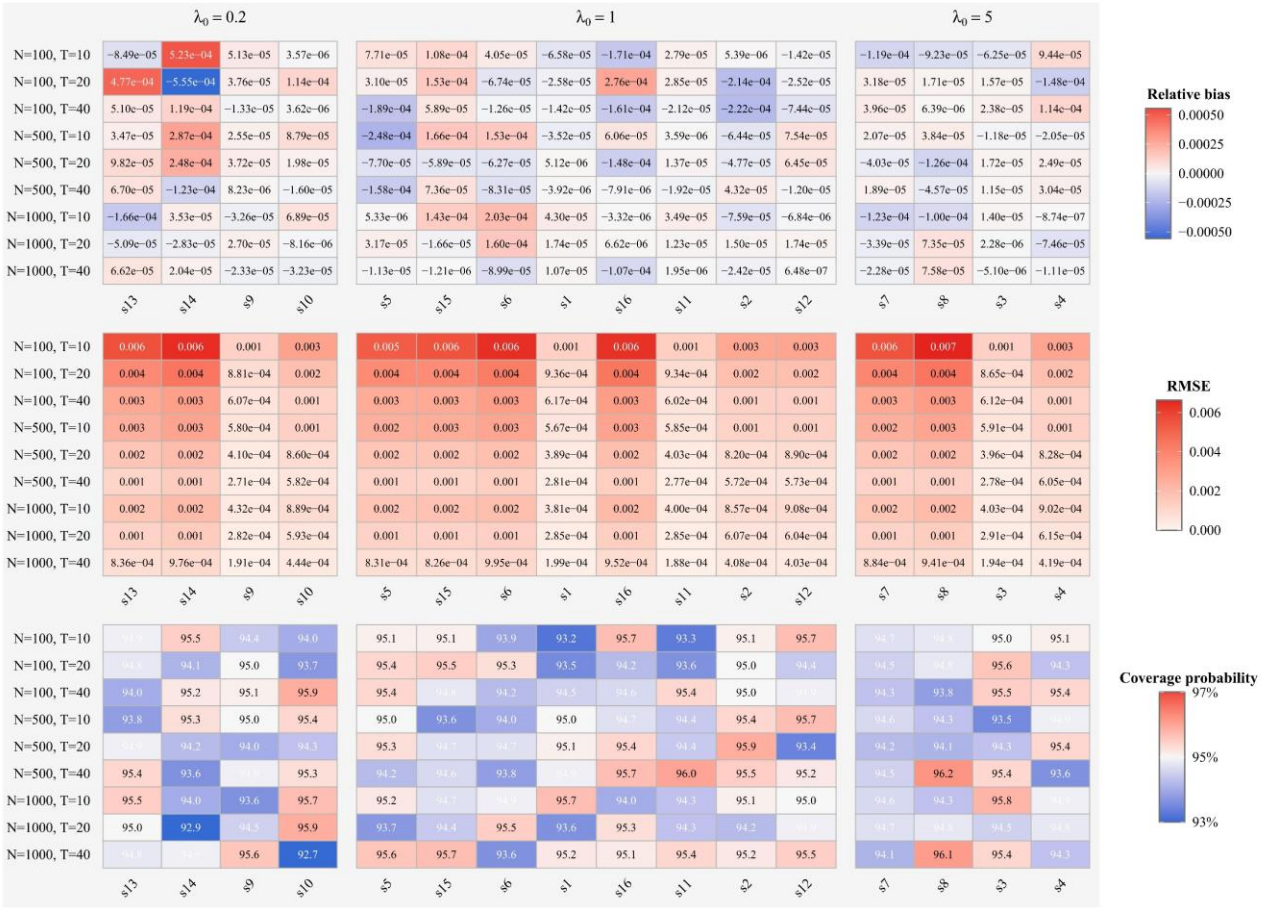


Figure 1 Finite-sample performance of the INLA estimator for the coefficient on x_1 and x_2. The figure reports the relative bias, RMSE, and empirical coverage probability of the nominal 95% posterior credible interval. The 16 simulation scenarios are grouped according to λ_0, while rows correspond to different combinations of N and T. Cell entries report the corresponding Monte Carlo performance measures.

4.4. Simulation results for TE estimates

The parameter estimation results conceal a more heterogeneous picture for the efficiency and inefficiency components. Table 3 and Figure 2 (The full information can be found in Tables A1-A3 in Appendix A) report some results of finite-sample performance of persistent, transient, and overall TE. The results are organized by λ_0, with the roles of λ, Λ, and the panel dimensions considered within each group.

The structural parameter results reveal a heterogeneous pattern for the recovery of TE. Table 3 summarizes the Pearson correlations between true and estimated persistent, transient, and

overall TE for the smallest and largest panel designs, while Figure 2 reports the corresponding relative biases across all 9 combinations of N and T. The complete simulation results, including correlations, relative bias, and deviations at the 5th, 50th, and 95th percentiles, are reported in Tables A1-A3 of Appendix A. The discussion below is organized by the persistent inefficiency-to-heterogeneity ratio, λ_0.

4.4.1. Weak persistent inefficiency relative to heterogeneity

The upper part of Table 3 and left panel of Figure 2 includes the cases of $\lambda_0 = 0.2$ (the full recovery information can be found in Table A1). The dominant feature of these results is the difficulty in recovering PTE. PTE correlations remain around $0.11\text{-}0.12$ across the four scenarios and show fairly no improvement as panel dimension grows. PTE is typically underestimated by only a few percentage points, yet the low correlations indicate that the cross-sectional ordering of PTE is poorly recovered. Hence, this could indicate the failure of separation of persistent inefficiency from unit heterogeneity.

For RTE, in contrast, performance changes sharply with λ. When $\lambda = 0.2$, as in s13, the correlation remains close to 0.12, so neither the PTE nor the transient RTE is recovered accurately across units. Raising λ to 1 produces correlations of approximately 0.5, while $\lambda = 5$ increases the correlation above 0.9. In s10, for example, the RTE correlation rises from 0.92 at $N = 100, T = 10$ to approximately 0.943 for the longer panels, while relative bias remains very small. Thus, the recovery of RTE improves markedly as transient inefficiency (σ_u) becomes larger relative to idiosyncratic noise (σ_v).

The detailed percentile results in Table A1 qualify this improvement. In the intermediate cases, small average bias can coexist with substantial errors in the tails of the TE distribution. In s14, for example, the 5th percentile of RTE is overestimated by roughly 20%, the 95th percentile is underestimated by about 10%, whereas the median is estimated more accurately. The corresponding distortions are considerably smaller in s9. Hence, the same value of λ need not imply the same distributional accuracy, and the allocation of variation between persistent and transient inefficiency also matters.

OTE reflects the recovery of both components. When one component is recovered more accurately than the other, overall performance is largely driven by the better-recovered

component. This is evident in s10, where RTE correlations rise from 0.920 to 0.943, while PTE correlations remain close to 0.11-0.12; the corresponding OTE correlation increases from 0.884 to 0.909. Thus, the high accuracy of OTE in this case does not imply equally accurate recovery of the persistent-transient decomposition.

4.4.2. Persistent inefficiency and heterogeneity of comparable magnitude

The second group corresponds to $\lambda_0 = 1$, with the complete results reported in Table A2. Compared with the case $\lambda_0 = 0.2$, PTE is recovered more accurately, although its performance remains sensitive to the scale configuration. Table 3 shows, for example, that the PTE correlation in s5 increases from 0.307 in the smallest panel dimension to 0.437 in the largest, while in s6 it rises from 0.278 to 0.416. The detailed results in Table A2 further show that increases in T are particularly useful for separating persistent inefficiency from time-invariant heterogeneity, whereas increases in N at a fixed T generally produce smaller changes in the PTE correlation.

RTE continues to depend primarily on λ. When $\lambda = 0.2$, its correlation remains close to 0.12, despite the improvement in PTE. When $\lambda = 1$, correlations are around 0.5, while $\lambda = 5$ raises them above 0.92. where RTE correlations increase to approximately 0.94 in the largest panels and relative biases remain close to 0. Hence, strengthening the persistent signal (σ_{u0}) does not by itself improve recovery of RTE; the latter remains governed primarily by transient inefficiency-to-noise ratio (λ).

The percentile results again show that small average bias can conceal large errors in the tails. In several scenarios, the median is recovered accurately while the 5th percentile is overestimated and the 95th percentile underestimated. This pattern is visible in s6 and s16. The comparison between s1 and s16 is informative since both have $\lambda_0 = \lambda = \Lambda = 1$, but differ in the absolute magnitude of standard deviation. Their correlations are very similar, whereas relative bias and percentile deviations are considerably larger in s16. Thus, the scale ratios play an important role in recovering cross-sectional differences in efficiency, whereas the absolute scale of the disturbances also affects the magnitude of prediction errors.

OTE again combines the information contained in the two components. In s12, for example, RTE is recovered with correlations above 0.91, whereas PTE correlations remain close to 0.52;

the resulting OTE correlation rises from 0.725 to 0.752. OTE can therefore be recovered reasonably well even when the persistent and transient components are recovered with markedly different levels of accuracy.

4.4.3. Persistent inefficiency dominant relative to heterogeneity

The final group corresponds to $\lambda_0 = 5$. This is the most favorable configuration for recovering PTE. Table 3 shows that PTE correlations are already high in the smallest panel dimension, ranging from 0.823 to 0.939 across the four scenarios, and rise further as the panel dimension expands. Figure 2 also shows that relative bias generally declines with increasing panel dimension. The contrast with the $\lambda_0 = 0.2$ cases is pronounced. When persistent inefficiency dominates time-invariant heterogeneity, the estimator recovers the cross-sectional variation in PTE more accurately.

In addition, increasing T again produces a particularly clear improvement in PTE correlation. In s7, for example, the correlation rises from 0.847 in the smallest panel size to 0.922 in the largest. Increases in N contribute more visibly to reductions in relative bias and percentile deviations. Figure 2 shows that persistent relative bias in s7 declines markedly as the panel expands, while Table A3 reports corresponding reductions in lower-tail prediction error.

The transient component continues to follow λ. With $\lambda = 0.2$ in s7, RTE correlation remains close to 0.12, despite the excellent recovery of PTE. At $\lambda = 1$, as in s8 and s3, the correlation rises to approximately 0.5. When both λ_0 and λ are large, as in s4, both PTE and RTE are recovered fairly accurately. PTE and RTE correlations exceed 0.9, while OTE correlation is approximately 0.93-0.95. Relative biases also become very small as the sample grows. For $N = 1000, T = 40$, they are approximately 0.2%, 0.3%, and 0.4% for PTE, RTE, and OTE, respectively.

The percentile results reinforce the same conclusion. When only one inefficiency component is strongly separated from its symmetric counterpart, the median may be recovered accurately while noticeable distortions remain in the tails of the other component. In s4, by contrast, the deviations at the 5th, 50th, and 95th percentiles are uniformly small. This scenario therefore provides the clearest evidence that PTE, RET, and OTE can all be recovered accurately when both inefficiency terms are sufficiently strong relative to their corresponding symmetric

idiosyncratic noise.

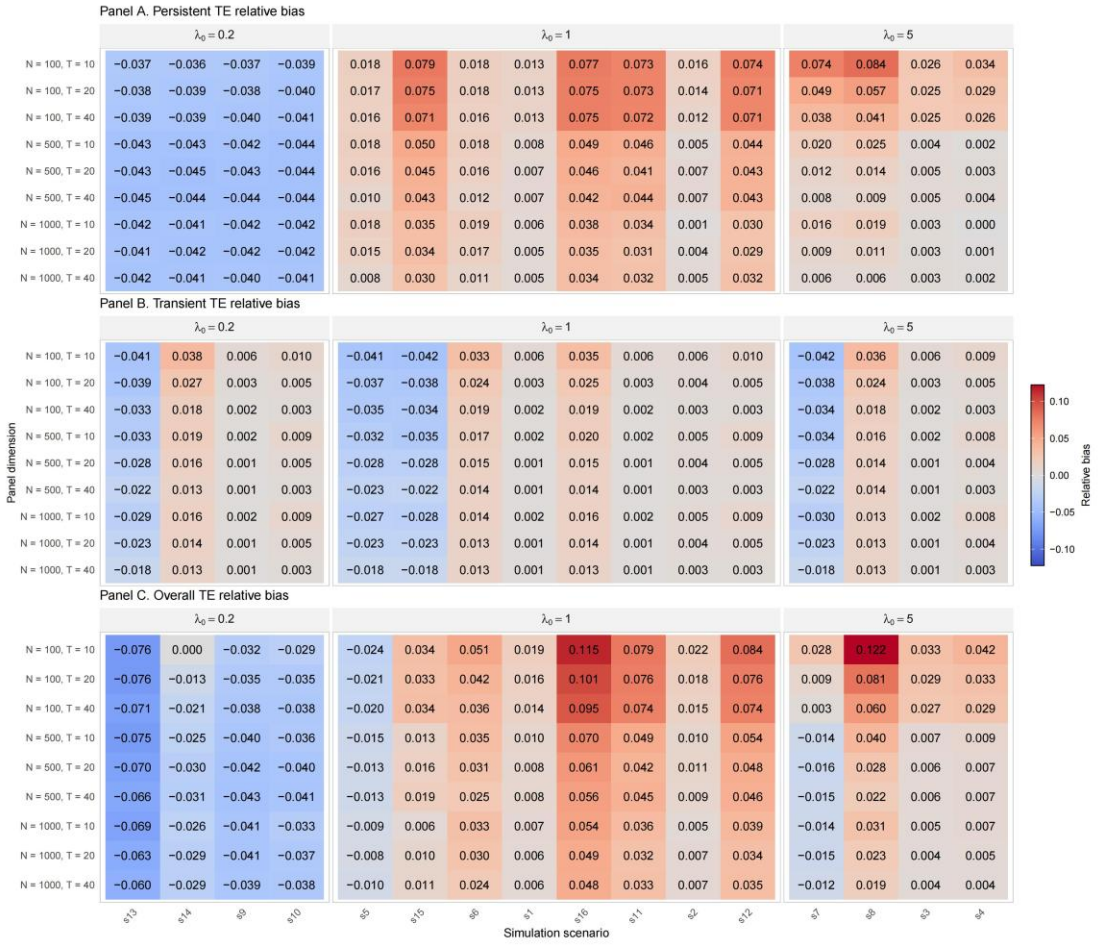


Figure 2. Relative bias of persistent, transient, and overall TE across simulation scenarios. Cells report the relative bias of posterior mean TE across 1,000 Monte Carlo replications. Scenarios are grouped by the persistent inefficiency-to-heterogeneity ratio, λ_0. Rows correspond to the nine combinations of $N$and T. Values closer to zero indicate smaller average bias.

Table 3 Correlation between true and estimated TE across simulation scenarios.

Scenario	λ_0	λ	Λ	Persistent Corr.	Transient Corr.	Overall Corr.
s1	1	1	1	0.506→0.526	0.508→0.526	0.499→0.515
s2	1	5	0.2	0.431→0.506	0.928→0.943	0.923→0.932
s3	5	1	5	0.939→0.948	0.505→0.526	0.923→0.934
s4	5	5	1	0.917→0.943	0.921→0.943	0.926→0.948
s5	1	0.2	1	0.307→0.437	0.115→0.119	0.158→0.263
s6	1	1	0.2	0.278→0.416	0.519→0.524	0.524→0.525
s7	5	0.2	5	0.847→0.922	0.115→0.119	0.707→0.902
s8	5	1	1	0.823→0.914	0.505→0.522	0.653→0.754
s9	0.2	1	1	0.117→0.118	0.504→0.526	0.243→0.278

s10	0.2	5	0.2	0.110→0.118	0.920→0.943	0.884→0.909
s11	1	1	5	0.521→0.527	0.505→0.526	0.499→0.518
s12	1	5	1	0.520→0.527	0.919→0.943	0.725→0.752
s13	0.2	0.2	1	0.117→0.118	0.114→0.119	0.118→0.111
s14	0.2	1	0.2	0.110→0.119	0.502→0.521	0.450→0.488
s15	1	0.2	5	0.502→0.523	0.112→0.118	0.388→0.501
s16	1	1	1	0.498→0.520	0.502→0.521	0.491→0.509

Notes: The table reports λ_0, λ, and Λ, together with the Pearson correlations between true and estimated PTE, RTE, and OTE. For each scenario, correlations are reported for the smallest and largest panel designs, $(N, T) = (100, 10) \rightarrow (1000, 40)$.

4.5 Discussion

The INLA estimator exhibits a clear finite-sample pattern across all scenarios. Its ability to recover PTE improves sharply with λ_0, while recovery of RTE improves similarly with λ. The corresponding correlations are close to 0.1 when the λ_0 is 0.2, around 0.5 when it is 1, and generally above 0.9 when it is 5. Larger N and T reduce bias and improve finite-sample accuracy when the inefficiency signal (σ_{u0} or σ_u) is sufficiently large, but they cannot fully compensate when the persistent inefficiency is difficult to distinguish from the time-invariant heterogeneity. The detailed percentile results in Tables A1-A3 further show that estimation is generally more accurate around the median than in the tails. Overall, recovery is governed primarily by the relative strength of each inefficiency component, deteriorating when persistent inefficiency is weak relative to time-invariant heterogeneity or when transient inefficiency is weak relative to idiosyncratic noise.

The INLA results are broadly consistent with those obtained using SML (Badunenko & Kumbhakar, 2016), particularly when the comparison is made at the common design $N = 500, T = 10$. In several scenarios the two estimators deliver almost identical recovery performance. For example, in s7 the correlations for PTE, RTE, and OTE are 0.853, 0.114, and 0.828 under INLA, compared with 0.85, 0.12, and 0.83 under SML. A similarly close correspondence is observed in s3, where the INLA correlations are approximately 0.943, 0.506, and 0.932, compared with 0.94, 0.51, and 0.93 under SML. These similarities could indicate INLA and SML deliver similar finite-sample behavior across different scale configurations. In other scenarios, however, INLA exhibits slightly stronger correlations

between estimated and true TE. In s4, for example, the correlations for PTE, RTE, and OTE are about 0.922, 0.922, and 0.947 under INLA, compared with 0.91, 0.91, and 0.94 under SML. Improvements are also visible in s16 and, particularly for overall TE, in s15. These gains are not uniform across all scenarios or performance measures. Overall, the comparison indicates broadly comparable finite-sample performance, with some evidence of stronger correlations between estimated and true TE under INLA in selected configurations.

The comparison with ABC reveals a different finite-sample pattern. Ramírez–Hassan et al. (2025) highlight that their two-stage ABC procedure performs well when the inefficiency-to-noise ratios are low or moderate, whereas relative bias and RMSE deteriorate as these ratios become large. Under INLA, by contrast, the correlations between estimated and true TE are weakest when λ_0 or λ is small and improve markedly as the corresponding inefficiency-to-noise ratio increases. The reported evidence therefore suggests different sensitivities to the strength of the inefficiency signal across the two approaches. This difference may reflect their distinct inferential strategies. ABC employs a two-stage likelihood-free procedure based on summary statistics, whereas INLA performs approximate Bayesian inference through the latent-model representation of the four-component SF model.

4.6 Computational aspects

This section assesses the computational burden of the proposed INLA estimator relative to alternative estimators for the four-component SF model as the panel dimension expands. We focus on the reparametrized MCMC estimator of Tsionas and Kumbhakar (TK), PMCMC, and SML.

The competing methods differ both in the treatment of the inefficiency terms and in the estimator used. TK is based on the ξ-parameterization, which employs an alternative reparameterization of GTRE intended to reduce dependence among the latent components. Estimation is carried out by data-augmentation MCMC, supplemented by an accept-reject step to improve mixing. PMCMC retains the original GTRE specification but replaces conventional latent-state updating with particle-based sampling within the MCMC algorithm. SML also uses the GTRE specification, while the latent components are integrated out by simulation and the resulting simulated likelihood is maximized numerically. The proposed INLA estimator instead

uses the double skew-normal representation developed above. These differences in model representation and estimation strategy lead to markedly different computational requirements (see Table A4 in Appendix A) as the panel dimension expands.

The computational experiment uses the representative case s16 ($\sigma_{u0} = \sigma_u = \sigma_{v0} = \sigma_v = 0.2$ and hence $\lambda_0 = \lambda = 1$) and considers three panel sizes, $(N, T) = (100, 10), (500, 20)$ and $(1000, 40)$. For each design, 100 datasets are generated, and the same dataset is supplied to all applicable estimators within each replication[6]. The resulting computation times are displayed in Figure 3.

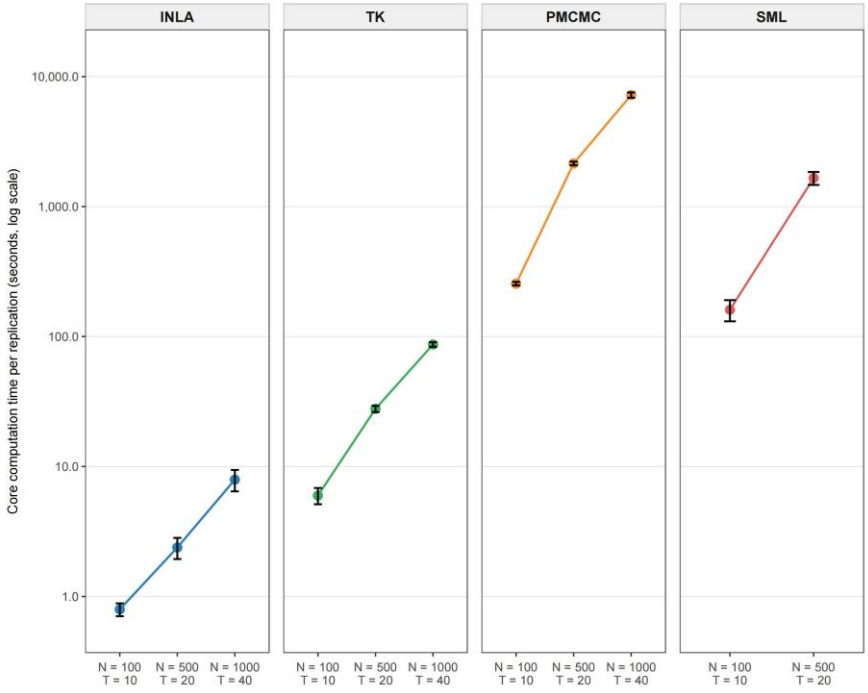


Figure 3. Computation time of alternative estimators for the four-component SF model. Points report mean computation time across 100 replications. The vertical axis is shown on a logarithmic scale. SML is not evaluated for $\mathrm{N} = 1000, \mathrm{T} = 40$ because of its huge computational cost.

Figure 3 shows a clear computational advantage for INLA. Its mean computation time rises from 0.8 seconds for $N = 100, T = 10$ to 2.38 seconds for $N = 500, T = 20$ and 7.92 seconds for $N = 1000, T = 40$. The corresponding times for TK are 5.98, 27.78, and 86.92 seconds, while PMCMC requires approximately 255, 2146, and 7215 seconds. SML takes about

[6] All estimators are run in single-threaded mode to ensure a consistent basis for comparing computational time.

161 seconds and 1656 seconds for the first two designs. Thus, even when the sample size increases from $NT = 1000$ to $NT = 40000$, INLA remains substantially faster and exhibits much weaker growth in computation time than the simulation-based alternatives. These results are consistent with the computational requirements summarized in Table A4. The proposed double skew-normal reparameterization avoids explicit simulation of the NT transient inefficiency terms within an MCMC loop and relies instead on sparse latent-Gaussian computation and fixed-dimensional deterministic integration.

5 Empirical illustrations

Having examined the performance of the proposed INLA estimator through Monte Carlo simulations, we now illustrate its application using two panel datasets, the well-known Spanish dairy farm (hereafter farm) dataset of Álvarez, Arias, and Greene (2004) and a dataset of Brazilian manufacturing firms (hereafter firm) obtained from PanelBox software (PanelBox Development Team, 2026). These datasets represent different scenarios in terms of the estimated values of λ_0, λ and Λ, allowing us to relate the results to the simulation evidence in Section 4[7].

5.1 Data and model

For the farm application, the dataset contains same 247 Spanish dairy farms from 1993 to 1998, providing 1482 observations in a balanced panel. The production frontier relates milk production to herd size, agricultural land, labor, and feed use. These variables are recorded in litres of milk, numbers of milking cows, hectares allocated to pasture and crops, man-equivalent labour units, and tonnes of feed, respectively.

For the firm application, the panel follows 500 firms annually from 2010 to 2019, providing 5000 observations with complete coverage for each firm. The production frontier relates value added to labour, capital, and materials, with both output and inputs expressed in logarithms.

Both applications employ a Cobb-Douglas production frontier with the normal–half-normal components defined in Section 3. For the farm panel, we estimate

$$\ln(y_{it})=\beta_0+\beta_{cows}\ln(x_{1it})+\beta_{land}\ln(x_{2it})+\beta_{labor}\ln(x_{3it})+\beta_{feed}\ln(x_{4it}) + v_{0i} - u_{0i} + v_{it} - u_{it}.$$

[7] The corresponding R code can be found in Appendix E.

For the firm panel, the production frontier is

$$\ln(y_{it})=\beta_0+\beta_{labor}\ln(x_{1it})+\beta_{capital}\ln(x_{2it})+\beta_{materials}\ln(x_{3it}) + v_{0i} - u_{0i} + v_{it} - u_{it}.$$

5.2 Frontier parameter estimates and component separation

Table 4 reports posterior means, standard deviations, and 95% credible intervals (CIs). For the farm panel, the estimated frontier output elasticities are 0.6421 for cows, 0.3991 for feed, 0.0348 for land, and 0.0351 for labour. All four inputs have positive elasticities, with their 95% CIs excluding zero. The posterior mean of the sum of these elasticities is 1.1111, with a 95% CI of $[1.0745, 1.1479]$. This result provides evidence of slight increasing returns to scale under the fitted Cobb-Douglas frontier.

For the firm panel, the estimated frontier output elasticities are 0.3549 for labour, 0.3035 for capital, and 0.3027 for materials. Their sum has a posterior mean of 0.9610 and a 95% CI of [0.9541,0.9680], providing evidence of slight decreasing returns to scale under the fitted Cobb-Douglas frontier.

The component scales reveal the more important contrast. For the farm case, $\sigma_{u0} = 0.1315$ and $\sigma_{v0} = 0.0837$, with a posterior mean ratio $\lambda_0 = 1.5871$ and a 95% CI of [1.1156,2.0913]. Persistent inefficiency is therefore larger than time-invariant heterogeneity on the underlying component-scale measure. The transient ratio λ=1.1819 has interval [0.9715,1.3941], placing transient inefficiency and noise at a comparable level. The posterior mean Λ=1.6904 also indicates a larger persistent than transient inefficiency scale.

The farm panel exhibits a different pattern in the relative scales of the four stochastic components. The posterior means of σ_{u_0} and σ_{v_0} are 0.1663 and 0.4222, respectively, while the posterior mean of their ratio, λ_0, is 0.3991, with a 95% CIs of [0.1602,0.7207]. Persistent inefficiency is small relative to heterogeneity, even though its estimated absolute scale exceeds that in the farm case. In contrast, $\sigma_u = 0.2929$ and $\sigma_v = 0.1952$ give $\lambda = 1.5017$ with interval [1.3890,1.6088]. The ratio Λ=0.5681, with interval [0.2286,0.9835], places the persistent inefficiency scale below the transient scale. (See Appendix D, which shows the robustness of the proposed INLA estimator.)

Table 4 Posterior estimates of frontier parameters and component-scales

Parameter	Mean	SD	95% CIs
Panel A			
$Intercept$	11.7447	0.0117	[11.7215, 11.7667]
β_{cows}	0.6421	0.0205	[0.6017, 0.6824]
β_{land}	0.0348	0.0132	[0.0091, 0.0612]
β_{labor}	0.0351	0.0171	[0.0018, 0.0676]
β_{feed}	0.3991	0.0108	[0.3784, 0.4201]

σ_{v0}	0.0837	0.006	[0.0723, 0.0952]
σ_{u0}	0.1315	0.0113	[0.1062, 0.1512]
σ_v	0.0663	0.002	[0.0625, 0.07]
σ_u	0.0781	0.0047	[0.068, 0.0871]
λ_0	1.5871	0.2352	[1.1156, 2.0913]
λ	1.1819	0.1022	[0.9715, 1.3941]
Λ	1.6904	0.1843	[1.3368, 2.0207]
Returns to scale	1.1111	0.0186	[1.0745, 1.1479]
Panel B			
$Intercept$	-3.7196	0.0702	[-3.8356, -3.5986]
β_{labor}	0.3549	0.0026	[0.3498, 0.36]
$\beta_{capital}$	0.3035	0.0019	[0.2997, 0.3072]
$\beta_{materials}$	0.3027	0.0022	[0.2984, 0.3069]
σ_{v0}	0.4222	0.0129	[0.4006, 0.4479]
σ_{u0}	0.1663	0.0795	[0.0697, 0.2887]
σ_v	0.1952	0.0032	[0.1894, 0.2014]
σ_u	0.2929	0.0063	[0.2798, 0.3047]
λ_0	0.3991	0.1989	[0.1602, 0.7207]
λ	1.5017	0.0543	[1.389, 1.6088]
Λ	0.5681	0.2717	[0.2286, 0.9835]
Returns to scale	0.961	0.0036	[0.9541, 0.968]

Notes: Panel A corresponds to the farm case; Panel B corresponds to the firm case. SD stands for standard deviation.

5.3 Estimated PTE and RTE

Table 5 summarizes PTE, RTE, and OTE. For the farm panel, the posterior means of average PTE and RTE are 0.9030 and 0.9411, while average OTE is 0.8498. The lower persistent TE indicates a larger average efficiency shortfall associated with the persistent component than with the transient component.

The firm panel displays the reverse pattern. Average PTE is 0.8823, while average RTE is 0.8046, yielding average OTE of 0.7099. Transient inefficiency therefore makes the larger component contribution to the estimated efficiency shortfall. Uncertainty is greater for average PTE, whose SD is 0.0523 (95% CIs$[0.8031, 0.9472]$), than for average transient TE, whose SD is 0.0039 (95% CIs $[0.7974, 0.8128]$). The broader uncertainty for the persistent component is consistent with the smaller λ_0 reported in Table 4.

Table 5 Posterior summaries of averaged TE

Dataset	TE component	Mean	SD	Min	Median	Max	95% CIs
Spanish dairy farm	Persistent	0.903	0.0085	0.7247	0.9027	0.9752	[0.8875, 0.9199]
	Transient	0.9411	0.0034	0.7385	0.9409	0.9874	[0.9351, 0.9479]
	Overall	0.8498	0.0084	0.6193	0.8496	0.9569	[0.8341, 0.8658]
Brazilian	Persistent	0.8823	0.0523	0.7796	0.8466	0.9311	[0.8031, 0.9472]

manufacturing firm	Transient	0.8046	0.0039	0.369	0.8044	0.9499	[0.7974, 0.8128]
	Overall	0.7099	0.0422	0.3102	0.6824	0.8616	[0.6455, 0.7595]

6 Conclusions

This study develops an INLA-based Bayesian estimator for the four-component SF model. The main contribution is to recast the model through a double skew-normal reparameterization, combining unit heterogeneity with persistent inefficiency and statistical noise with transient inefficiency. This reparameterization substantially lowers the latent dimensionality of the original four-component model specification. The non-Gaussian unit-specific component is accommodated through a diffuse Gaussian carrier with an auxiliary skew-normal likelihood, yielding a representation that can be estimated within the LGM framework of INLA. The approach therefore provides a deterministic alternative to simulation-based Bayesian estimation while preserving the economically important decomposition of persistent and transient inefficiency.

The Monte Carlo results show that the proposed INLA estimator performs well for the structural parameters across the designs considered. More importantly, they clarify the conditions under which the four stochastic components can be separated. Recovery of both persistent and transient inefficiency terms depends primarily on the corresponding signal-to-noise ratio, with stronger inefficiency signals leading to more accurate recovery. Increasing the panel dimensions improves recovery when the relevant inefficiency signal is sufficiently strong, but cannot fully compensate for weak identification when that signal is small. Thus, accurate estimation of structural parameters does not necessarily imply accurate recovery of the latent persistent and transient inefficiency terms.

The second finding is that INLA exhibits more favorable computational scaling with the panel dimension than the simulation-based estimators considered in the comparison. This advantage reflects both the double skew-normal reparameterization and INLA's deterministic approximation strategy. The transient composite disturbance is absorbed into the likelihood, while the remaining high-dimensional Gaussian latent field is handled through sparse approximations and numerical integration is confined to the low-dimensional hyperparameter space. The resulting computational gains make Bayesian estimation of the four-component

model more practical for larger panels.

The empirical applications to Spanish dairy farms and Brazilian manufacturing firms illustrate why retaining the four-component SF model matters in practice. The two datasets reveal contrasting patterns of inefficiency. PTE plays a larger role in the dairy-farm application, whereas RTE is more prominent in the manufacturing application. These findings highlight the importance of distinguishing PTE and RTE rather than relying solely on OTE.

However, the proposed approach also has limitations. The simulation evidence reinforces a limitation inherent to four-component SF models rather than to INLA specifically. When inefficiency is weak relative to the competing heterogeneity or noise component, its individual realization is difficult to recover even in larger samples. Extensions to alternative inefficiency distributions, richer dependence structures, and more flexible temporal or spatial specifications would therefore be useful directions for further research.

References

Azzalini, A. (2005). The Skew-normal Distribution and Related Multivariate Families. *Scandinavian Journal of Statistics*, *32*(2), 159–188. https://doi.org/https://doi.org/10.1111/j.1467-9469.2005.00426.x

Badunenko, O., & Kumbhakar, S. C. (2016). When, where and how to estimate persistent and transient efficiency in stochastic frontier panel data models. *European Journal of Operational Research*, *255*(1), 272–287.

Battese, G. E., & Coelli, T. J. (1988). Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics*, *38*(3), 387–399. https://doi.org/https://doi.org/10.1016/0304-4076(88)90053-X

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Colombi, R., Kumbhakar, S. C., Martini, G., & Vittadini, G. (2014). Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency. *Journal of Productivity Analysis*, *42*(2), 123–136. https://doi.org/10.1007/s11123-014-0386-y

Filippini, M., & Greene, W. (2016). Persistent and transient productive inefficiency: a maximum simulated likelihood approach. *Journal of Productivity Analysis*, *45*(2), 187–196.

Giordano, R., Broderick, T., & Jordan, M. I. (2018). Covariances, robustness and variational bayes. *J. Mach. Learn. Res.*, *19*(1), 1981–2029.

Hajargasht, G., & Griffiths, W. E. (2018). Estimation and testing of stochastic frontier models using variational Bayes. *Journal of Productivity Analysis*, *50*(1), 1–24. https://doi.org/10.1007/s11123-018-0531-0

Jondrow, J., Knox Lovell, C. A., Materov, I. S., & Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, *19*(2), 233–238. https://doi.org/https://doi.org/10.1016/0304-4076(82)90004-5

Kumbhakar, S. C., Lien, G., & Hardaker, J. B. (2014). Technical efficiency in competing panel data models: a

study of Norwegian grain farming. *Journal of Productivity Analysis*, *41*(2), 321–337.

Lindgren, F., & Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1–25. https://doi.org/10.18637/jss.v063.i19

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC press.

Martins, T. G., & Rue, H. (2014). Extending Integrated Nested Laplace Approximation to a Class of Near-Gaussian Latent Models. *Scandinavian Journal of Statistics*, *41*(4), 893–912. https://doi.org/https://doi.org/10.1111/sjos.12073

Ng'ombe, J., & Lambert, D. (2021). Using Hamiltonian Monte Carlo via Stan to estimate crop input response functions with stochastic plateaus. *Journal of Agriculture and Food Research*, *6*, 100226. https://doi.org/https://doi.org/10.1016/j.jafr.2021.100226

PanelBox Development Team. (2026). *PanelBox: Panel Data Econometrics for Python* (1.0.0). https://github.com/PanelBox-Econometrics-Model/panelbox

Ramírez–Hassan, A., Rengifo–Castro, J. D., Manzur, M., & Rueda-Ramírez, E. (2025). Approximate Bayesian computation to estimate persistent and transient efficiency in stochastic frontier panel data models. *Journal of Productivity Analysis*, *64*(2), 145–166. https://doi.org/10.1007/s11123-025-00765-3

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1–28. https://doi.org/10.48550/arXiv.1403.4630

Tsionas, E. G., & Kumbhakar, S. C. (2014). Firm heterogeneity, persistent and transient technical inefficiency: A generalized true random-effects model. *Journal of Applied Econometrics*, *29*(1), 110–132.

Tsionas, M., & Mallick, S. (2019). A Bayesian semiparametric approach to stochastic frontiers and productivity. *European Journal of Operational Research*, *274*(1), 391–402. https://doi.org/https://doi.org/10.1016/j.ejor.2018.10.026

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.pdf