

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Bayesian group LASSO for high-dimensional stochastic frontier analysis

Ancha Xu, Lijia Liu, Songzi Zhang & Ruiyue Lin

To cite this article: Ancha Xu, Lijia Liu, Songzi Zhang & Ruiyue Lin (25 Aug 2026): Bayesian group LASSO for high-dimensional stochastic frontier analysis, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2721727

To link to this article: https://doi.org/10.1080/24754269.2026.2721727

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 25 Aug 2026.

Submit your article to this journal

Article views: 328

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2721727

Taylor & Francis
by informa

REVIEW ARTICLE

OPEN ACCESS Check for updates

Bayesian group LASSO for high-dimensional stochastic frontier analysis

Ancha Xu [a,b], Lijia Liu[a], Songzi Zhang[a] and Ruiyue Lin[c]

[a]School of Statistics and Data Science, Zhejiang Gongshang University, Hangzhou, People's Republic of China; [b]Collaborative Innovation Center of Statistical Data Engineering, Technology & Application Zhejiang Gongshang University, Hangzhou, People's Republic of China; [c]School of Mathematics and Physics, Wenzhou University, Wenzhou, People's Republic of China

ABSTRACT

This study develops an innovative Bayesian stochastic frontier model that addresses two critical limitations in efficiency analysis: the curse of dimensionality in high-dimensional production environments and the failure to respect natural groupings among input variables. Our methodology combines group LASSO regularization with adaptive priors informed by data envelopment analysis, creating a unified framework that leverages both parametric and nonparametric efficiency information. The group-structured penalties preserve economically meaningful variable relationships while enabling automatic selection of relevant input clusters. An efficient Markov chain Monte Carlo algorithm facilitates posterior inference under this complex hierarchical specification. Extensive numerical experiments demonstrate remarkable improvements in estimation accuracy, particularly in high-dimensional and sparse settings. Application to commercial banking data reveals the method's ability to identify core operational drivers and uncover bimodal efficiency distributions, suggesting distinct strategic positioning among financial institutions.

ARTICLE HISTORY
Received 17 September 2025
Revised 3 December 2025
Accepted 17 August 2026

KEYWORDS
Data envelopment analysis; stochastic frontier analysis; group LASSO; Bayesian inference

1. Introduction

In modern competitive markets, accurately measuring technical efficiency across multiple production units is essential to diagnose performance gaps and inform optimal resource allocation. Two principal methodological frameworks dominate the field of efficiency analysis: the stochastic frontier model (SFM) and data envelopment analysis (DEA).

SFM models the production frontier parametrically and decomposes deviations into inefficiency and random noise components, typically assuming specific distributional forms for the inefficiency term, such as the half-normal distribution (Jondrow et al., 1982; Kumbhakar & Tsionas, 2011). While this framework allows for statistical inference and accommodates random disturbances, its reliance on strong parametric assumptions may lead to specification bias when real-world data violate these assumptions. Recent advances in Bayesian stochastic frontier analysis have introduced greater flexibility by incorporating prior information and

CONTACT Songzi Zhang 241432090@qq.com School of Statistics and Data Science, Zhejiang Gongshang University, 18 Xuezheng Street, Hangzhou 310018, Zhejiang, People's Republic of China

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

hierarchical structures (Tsionas, 2019), yet they still often depend on similar distributional assumptions.

DEA, in contrast, is a nonparametric, deterministic approach that constructs the production frontier through linear programming, requiring no assumptions about the functional form of the production technology (Simar & Wilson, 2007). This makes DEA highly flexible and adaptable. However, as emphasized by Simar and Zelenyuk (2020), the absence of a stochastic component in DEA means that it cannot explicitly account for random noise, making its inefficiency estimates sensitive to data imperfections and reducing their statistical reliability. Although some hybrid methods, such as empirical Bayes approaches (Tsionas, 2003), attempt to combine the strengths of SFM and DEA, they often fall short when confronted with high-dimensional covariates or latent structural heterogeneity.

An important structural issue that has received limited attention in the efficiency analysis literature is the fact that input variables in a firm's production process often operate not in isolation but as logically related groups. For instance, labor inputs stratified by skill level, or multiple indicators representing capital investment, are conceptually interconnected and should be treated as unified blocks rather than as unrelated individual variables. Traditional variable selection techniques, such as the classical LASSO (Tibshirani, 1996), treat each variable independently, potentially disrupting inherent relationships and leading to estimation bias when applied to data with group structures (Bakin, 1999). To address this, Yuan and Lin (2006) introduced the group LASSO, which extends the sparsity-inducing penalty from the individual to the group level. This method selects or discards entire groups of variables simultaneously, preserving structural coherence and enhancing interpretability. The group penalty, typically defined via the ℓ_2-norm of group coefficients, ensures that the variable selection respects natural partitions in the data. Subsequent developments have expanded the theory and applications of group LASSO in various domains (Friedman et al., 2010). Bayesian formulations of group LASSO have been explored in the literature, notably in nonparametric varying-coefficient models for functional genome-wide association studies (Li et al., 2015) and in general variable selection settings (Xu & Ghosh, 2015). These works primarily focus on inducing group-level sparsity in regression coefficients to enhance prediction and interpretation in high-dimensional settings. In contrast, our study introduces a novel approach by embedding a group LASSO prior within the inefficiency component of a stochastic frontier model, while simultaneously incorporating DEA-informed adaptive shrinkage on the inefficiency terms. This dual regularization (enforcing group sparsity on covariates and employing data-driven adaptation based on nonparametric DEA scores) effectively addresses firm-level inefficiency heterogeneity while preserving meaningful input group structures, an aspect not previously addressed in the Bayesian group LASSO literature.

In the context of technical inefficiency estimation across multiple firms, incorporating group sparsity is particularly compelling. Recent trends in efficiency analysis have emphasized the need to accommodate complex data structures, such as spatio-temporal heterogeneity (Emili & Galli, 2025). Moreover, multi-firm datasets often feature high-dimensional characteristics exacerbating the risk of overfitting, where the number of covariates p far exceeds the number of observations n (Chen et al., 2021; Lee & Cai, 2020). Although LASSO and its variants have been explored within SFM to address these challenges (Horrace et al., 2022; Parmeter et al., 2025), they often remain vulnerable to model instability. In contrast, group LASSO imposes structured regularization, effectively reducing the model's degrees of freedom and yielding more robust estimates by leveraging group-level sparsity.

To fill this gap, our study makes several key contributions to the field of efficiency analysis. This dynamic shrinkage mechanism captures heterogeneous inefficiency behaviour in a data-driven manner and resonates with recent efforts to impose LASSO-style regularization directly on inefficiency components or distributional parameters in SFM (Horrace et al., 2022; Tsionas, 2023).

(1) We propose an innovative Bayesian stochastic frontier model that synergistically combines the strengths of parametric SFM with nonparametric DEA. This unified framework effectively leverages information from both approaches to provide a more robust and comprehensive assessment of technical inefficiency.
(2) To address the curse of dimensionality and preserve economically meaningful relationships among inputs, we integrate a group LASSO prior into the SFM. This allows for automatic selection and shrinkage of entire groups of explanatory variables, ensuring more stable coefficient estimates and enhancing interpretability, particularly in high-dimensional and sparse settings where p is large relative to n.
(3) We introduce a data-driven adaptive prior for the technical inefficiency term, which is informed by pre-computed DEA inefficiency scores. This mechanism, inspired by the adaptive LASSO framework (Fan & Li, 2001; Zou, 2006), allows the prior distribution of inefficiency to dynamically adapt to each firm's observed efficiency level, thereby avoiding rigid distributional assumptions and enabling the accurate capture of heterogeneous inefficiency behaviour across firms.

The proposed methodology offers a comprehensive and distinctive approach, leading to robust estimation of firm-specific inefficiency (facilitating the accurate identification of both high- and low-efficiency firms) and providing actionable insights by pinpointing key groups of input variables that jointly influence output.

The remainder of the paper is organized as follows. Section 2 introduces the model and the corresponding regularized problem. Section 3 presents the technical details of Bayesian inference. Section 4 presents Monte Carlo simulations to assess the model's performance. Section 5 applies the proposed method to U.S. commercial bank data. Section 6 concludes with a discussion of key findings and future directions.

2. Model specification and regularization

We consider the classical stochastic frontier model (SFM) specified as follows:

$$y_i = x_i^\top \beta + v_i - u_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i denotes the output (dependent variable), $x_i \in \mathbb{R}^p$ is a vector of input (independent) variables, v_i represents statistical noise typically assumed to follow a normal distribution, $v_i \sim \mathcal{N}(0, \sigma_v^2)$, and $u_i \geq 0$ is the non-negative technical inefficiency term that captures the shortfall of observed output from the production frontier, representing manageable inefficiencies due to factors such as suboptimal resource allocation or managerial practices (Jondrow et al., 1982; Kumbhakar & Tsionas, 2011). Common distributional assumptions include half-normal or exponential, but these may lead to misspecification in heterogeneous settings as indicated by Tsionas (2023).

A central challenge in practical applications arises when input variables exhibit natural group structures. For instance, in banking applications, variables may naturally cluster into

categories such as human capital indicators (management ratios, staff qualifications), physical capital measures (branch networks, technology infrastructure), and funding composition variables (deposit types, wholesale funding sources). Traditional SFM frameworks treat each variable independently during estimation and selection, potentially disrupting these economically meaningful relationships and leading to overfitting in high-dimensional settings where the number of variables p approaches or exceeds the sample size n.

To address these limitations and effectively capture group-level contributions to production efficiency, we propose a regularized estimation framework that incorporates structured sparsity through group-level penalties. We assume that the p-dimensional input vector can be partitioned into G mutually exclusive groups: $x_i = [x_{i,1}^\top, x_{i,2}^\top, \ldots, x_{i,G}^\top]^\top$, where $x_{i,g} \in \mathbb{R}^{m_g}$ denotes the vector of m_g input variables corresponding to the gth group (for $g = 1, \ldots, G$), and $\sum_{g=1}^{G} m_g = p$. Correspondingly, the coefficient vector is partitioned as $\beta = [\beta_1^\top, \beta_2^\top, \ldots, \beta_G^\top]^\top$, where $\beta_g \in \mathbb{R}^{m_g}$ contains the coefficients for the gth group.

Drawing on the group LASSO methodology, we formulate the following optimization problem.

$$\min_{\beta \in \mathbb{R}^p, u \in \mathbb{R}_+^n} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{g=1}^{G} x_{i,g}^\top \beta_g + u_i \right)^2 + \lambda_u \sum_{i=1}^{n} w_i u_i + \sum_{g=1}^{G} \lambda_{\beta,g} \sqrt{m_g} \|\beta_g\|_2^2 \right\}. \quad (2)$$

The objective function in (2) consists of three components that collectively balance model fit, inefficiency regularization, and group-structured coefficient selection.

- The first component, $\frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{g=1}^{G} x_{i,g}^\top \beta_g + u_i)^2$, represents the sum of squared residuals that measures the model's fitting accuracy. This term corresponds to the negative log-likelihood under the assumption of Gaussian noise and ensures that the estimated model adequately explains the observed data.
- The second component, $\lambda_u \sum_{i=1}^{n} w_i u_i$, imposes an adaptive ℓ_1 penalty on the technical inefficiency terms (where the ℓ_1 penalty refers to the weighted sum of absolute values, equivalent to a weighted L_1 norm since $u_i \geq 0$). Following Tsionas (2023), the adaptive weights w_i are set as $(c + u_{\text{dea},i})^{-\gamma}$, where $u_{\text{dea},i} = 1 - \text{eff}_i$ represents the inefficiency score pre-computed using an input-oriented, variable returns to scale (VRS) DEA model (specifically, the BCC model), $c > 0$ is a smoothing parameter that ensures numerical stability by preventing division by zero for efficient units, and $\gamma > 0$ controls the sensitivity of the adaptive mechanism. Although u_i represents manageable technical inefficiency, unpenalized estimation risks overfitting by arbitrarily attributing residuals to inefficiency, especially in high-dimensional settings. The ℓ_1 penalty induces sparsity, encouraging most $u_i \approx 0$ while allowing evidence-driven positive values for truly inefficient units. The adaptive weights integrate DEA priors: highly inefficient DEA units (large $u_{\text{dea},i}$) receive lighter penalties (smaller w_i), permitting larger u_i in SFA, whereas efficient units face stronger zero-shrinkage. This bridges parametric and non-parametric methods, mitigates misspecification, and enhances robustness to inefficiency heterogeneity.
- The third component, $\sum_{g=1}^{G} \lambda_{\beta,g} \sqrt{m_g} \|\beta_g\|_2^2$, implements group-level shrinkage on the regression coefficients. The group-specific regularization parameter $\lambda_{\beta,g}$ controls the shrinkage intensity for the gth group, while the normalization factor $\sqrt{m_g}$ adjusts for differences in group size, preventing larger groups from being unduly penalized simply because they contain more variables (Yuan & Lin, 2006). This size-adjusted penalty

enables fair comparison across groups of varying dimensions and maintains selection consistency.

The group LASSO penalty offers several advantages over traditional variable-wise selection methods: (i) it maintains the interpretability of economically meaningful variable clusters, (ii) it reduces the effective dimensionality by operating at the group level rather than individual variables, and (iii) it provides more stable coefficient estimates by pooling information within groups. When $\lambda_{\beta,g}$ is sufficiently large, all coefficients within group g are simultaneously strongly shrunk toward zero, effectively removing the entire group from the model. This group-wise selection mechanism is particularly valuable in high-dimensional settings where preserving structural relationships among variables is crucial for economic interpretation.

3. Bayesian inference

To conduct comprehensive statistical inference on the optimization problem presented in Section 2 and quantify parameter uncertainty, we embed the regularized SFM within a Bayesian framework. This approach allows us to naturally incorporate the penalty terms as prior distributions and provides a principled mechanism for uncertainty quantification through posterior sampling.

3.1. Likelihood function

Within the Bayesian framework, the likelihood function describes the probability distribution of the observed data $\mathbf{y} = (y_1, \ldots, y_n)^\top$ conditional on the model parameters. Given the stochastic frontier specification in Equation (1), each observation y_i follows a normal distribution conditional on the regression coefficients β, inefficiency terms $\mathbf{u} = (u_1, \ldots, u_n)^\top$, and noise variance σ_v^2. The likelihood function is therefore expressed as

$$p(\mathbf{y} \mid \beta, \mathbf{u}, \sigma_v^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp\left(-\frac{(y_i - x_i^\top \beta + u_i)^2}{2\sigma_v^2} \right). \tag{3}$$

Taking the logarithm, we obtain

$$\log p(\mathbf{y} \mid \beta, \mathbf{u}, \sigma_v^2) = -\frac{n}{2} \log(2\pi\sigma_v^2) - \frac{1}{2\sigma_v^2} \sum_{i=1}^{n} (y_i - x_i^\top \beta + u_i)^2. \tag{4}$$

This establishes the direct correspondence between maximizing the log-likelihood and minimizing the sum of squared residuals in the optimization problem (2), providing the foundation for the Bayesian formulation.

3.2. Prior distributions

The key insight of the Bayesian approach lies in translating the regularization terms from the optimization problem into appropriate prior distributions. This translation enables automatic regularization through the Bayesian inference process while providing natural uncertainty quantification.

3.2.1. Group-structured priors for regression coefficients

To implement the group LASSO penalty within the Bayesian framework, we assign hierarchical priors to the coefficient vectors. For the gth group, we specify

$$\beta_g \mid \sigma_v^2, \lambda_{\beta,g} \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_v^2}{\sqrt{m_g}\lambda_{\beta,g}}\mathbf{I}_{m_g}\right), \quad g = 1, \ldots, G. \tag{5}$$

While a double exponential (Laplace) prior is commonly used in the Bayesian LASSO to induce individual coefficient sparsity, it is not suitable for group-structured regularization. The Laplace prior applies an L_1 penalty independently to each coefficient, which would break the desired 'group selection' property where all coefficients within a group are simultaneously shrunk or retained. In contrast, the multivariate normal prior with group-specific variance proportional to $\frac{\sigma_v^2}{\sqrt{m_g}\lambda_{\beta,g}}$ induces an L_2 penalty within each group, which exactly reproduces the adaptive group LASSO penalty when combined with the hierarchical Gamma prior on $\lambda_{\beta,g}$ (Xu & Ghosh, 2015; Yuan & Lin, 2006). This formulation ensures that groups of different sizes are fairly penalized and allows data-driven shrinkage of entire variable groups, aligning with the structural assumptions in efficiency analysis, where $\mathbf{I}_{m_g}$ denotes the $m_g \times m_g$ identity matrix. This prior specification has several important properties.

- The prior variance $\frac{\sigma_v^2}{\sqrt{m_g}\lambda_{\beta,g}}$ controls the degree of shrinkage for coefficients in group g. As $\lambda_{\beta,g}$ increases, the prior variance decreases, inducing stronger shrinkage toward zero.
- The normalization by $\sqrt{m_g}$ ensures that groups of different sizes are penalized appropriately, preventing larger groups from being disproportionately shrunk simply due to their size.
- The dependence on σ_v^2 creates a natural scaling relationship between the noise level and coefficient shrinkage, automatically adapting the regularization strength to the signal-to-noise ratio in the data.

3.2.2. Adaptive priors for inefficiency terms

To integrate non-parametric DEA information while maintaining distributional flexibility, we specify adaptive exponential priors for the inefficiency terms:

$$u_i \mid \lambda_u, c, \gamma \sim \text{Exp}(\lambda_u w_i), \tag{6}$$

where the adaptive weight $w_i = (c + u_{\text{dea},i})^{-\gamma}$ is constructed using pre-computed DEA inefficiency scores $u_{\text{dea},i}$, as introduced in Section 2. Recall that $u_{\text{dea},i} = 1 - \text{eff}_i$, where $\text{eff}_i \in (0, 1]$ is the efficiency score obtained from the BCC model (Banker et al., 1984). The parameters c and γ control the adaptiveness of the prior.

The adaptive mechanism operates as follows: For units identified as highly inefficient by DEA (large $u_{\text{dea},i}$), the weight w_i becomes smaller, leading to a larger prior expectation $E[u_i] = 1/(\lambda_u w_i)$. This allows the SFM to assign higher inefficiency values to these units, maintaining consistency with the DEA assessment. Conversely, for efficient units (small $u_{\text{dea},i}$), larger weights impose stronger shrinkage toward zero, encouraging the SFA inefficiency estimates to align with DEA's efficiency assessment.

3.2.3. Hyperparameter priors

To complete the hierarchical Bayesian specification, we assign weakly informative priors to all hyperparameters, ensuring that posterior inference remains primarily data-driven.

- Group regularization parameters:

$$\lambda_{\beta,g} \sim \text{Gamma}(a_{\lambda_{\beta,g}}, b_{\lambda_{\beta,g}}), \quad g = 1, \ldots, G. \tag{7}$$

- Inefficiency regularization parameter:

$$\lambda_u \sim \text{Gamma}(a_{\lambda_u}, b_{\lambda_u}). \tag{8}$$

- Error variance:

$$\sigma_v^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma). \tag{9}$$

- Adaptive weight parameters:

$$c \sim \text{Gamma}(a_c, b_c), \quad \gamma \sim \text{Gamma}(a_\gamma, b_\gamma). \tag{10}$$

The Gamma distribution is chosen for its conjugacy properties with the exponential and normal distributions, facilitating efficient posterior computation. The hyperparameter values $(a., b.)$ are set to induce weakly informative priors that provide minimal constraints while ensuring proper posterior distributions. This hierarchical specification allows all regularization parameters to be learned from the data rather than fixed a priori, providing adaptive regularization that automatically adjusts to the characteristics of each dataset. The resulting Bayesian model naturally quantifies uncertainty in all parameter estimates while incorporating both group structure and DEA information into the efficiency analysis framework.

3.3. Posterior inference

Combining the likelihood function with all prior distributions, we can obtain the joint posterior distribution, which represents the joint probability distribution of all unknown parameters given the observed data y. According to Bayes' theorem, the joint posterior distribution is proportional to the product of the likelihood function and all prior distributions.

$$\begin{aligned} p(\beta, u, \sigma_v^2, \lambda_\beta, \lambda_u, c, \gamma \mid y) \propto p(y \mid \beta, u, \sigma_v^2) \times \prod_{g=1}^{G} p(\beta_g \mid \sigma_v^2, \lambda_{\beta,g}) p(\lambda_{\beta,g}) \\ \times \prod_{i=1}^{n} p(u_i \mid \lambda_u, c, \gamma) p(\lambda_u) p(\sigma_v^{-2}) p(c) p(\gamma), \end{aligned} \tag{11}$$

where $\lambda_\beta = (\lambda_{\beta,1}, \ldots, \lambda_{\beta,G})$. As this joint posterior distribution is complex in form and cannot be solved analytically, we employ Gibbs sampling combined with the Metropolis-Hastings (MH) algorithm for efficient posterior inference, thereby generating posterior samples of the parameters.

3.3.1. Full conditional posterior distributions

To implement MCMC sampling, we need to derive the full conditional posterior distribution for each parameter (or block of parameters). These conditional distributions form the core of the Gibbs sampler; by sequentially drawing samples from these distributions, we can generate a Markov chain that converges to the joint posterior distribution.

- For the regression coefficient vector β_g of each group, its full conditional posterior distribution is a multivariate normal distribution:

$$\beta_g \mid \cdot \sim \mathcal{N}_{m_g}\left(\mu_{\beta_g}, \Sigma_{\beta_g}\right), \tag{12}$$

  where $\Sigma_{\beta_g}^{-1} = \sigma_v^{-2} X_g^\top X_g + \sigma_v^{-2}\sqrt{m_g}\lambda_{\beta,g} I_{m_g}$ and $\mu_{\beta_g} = \Sigma_{\beta_g}(\sigma_v^{-2} X_g^\top (y - \sum_{h\neq g} X_h\beta_h + u))$.
- The full conditional posterior distribution of the inefficiency term u_i is a truncated normal distribution on $\mathbb{R}_+ = [0, \infty)$:

$$u_i \mid \cdot \sim \mathcal{N}_+\left(x_i^\top \beta - y_i - \sigma_v^2 \lambda_u w_i, \sigma_v^2\right). \tag{13}$$

- The conditional distribution of the inverse-variance of the error, σ_v^{-2}, is a Gamma distribution:

$$\sigma_v^{-2} \mid \cdot \sim \mathcal{G}\left(\frac{n + \sum_g m_g}{2} + a_\sigma, \frac{\sum_{i=1}^n (y_i - x_i^\top \beta + u_i)^2 + \sum_g \sqrt{m_g}\lambda_{\beta,g}\|\beta_g\|_2^2}{2} + b_\sigma\right). \tag{14}$$

- The full conditional posterior distribution of $\lambda_{\beta,g}$ $(g = 1, \ldots, G)$ is

$$\lambda_{\beta,g} \mid \cdot \sim \mathcal{G}\left(a_{\lambda_{\beta,g}} + \frac{m_g}{2}, \frac{\sqrt{m_g}\|\beta_g\|_2^2}{2\sigma_v^2} + b_{\lambda_{\beta,g}}\right). \tag{15}$$

- The full conditional posterior distribution of λ_u is

$$\lambda_u \mid \cdot \sim \mathcal{G}\left(a_{\lambda_u} + n, b_{\lambda_u} + \sum_{i=1}^n w_i u_i\right). \tag{16}$$

- The full conditional posterior distributions of c and γ are as follows:

$$\begin{aligned} p(c \mid \cdot) &\propto \prod_{i=1}^n (\lambda_u (c + \tilde{u}_i)^{-\gamma}) \exp(-b_c c - \lambda_u (c + \tilde{u}_i)^{-\gamma} u_i), \\ p(\gamma \mid \cdot) &\propto \prod_{i=1}^n (\lambda_u (c + \tilde{u}_i)^{-\gamma}) \exp(-b_\gamma \gamma - \lambda_u (c + \tilde{u}_i)^{-\gamma} u_i). \end{aligned} \tag{17}$$

  Due to the non-standard form of the full conditional posteriors for c and γ in Equation (17), direct sampling methods are infeasible. We therefore employ a random walk Metropolis-Hastings (MH) algorithm, using a truncated normal distribution as

the proposal density under the constraint that both $c > 0$ and $\gamma > 0$. At iteration s, the proposals are drawn as

$$c' \sim \mathcal{N}_+\left(c^{(s)}, \sigma_c^2\right), \quad \gamma' \sim \mathcal{N}_+\left(\gamma^{(s)}, \sigma_\gamma^2\right), \tag{18}$$

where $\mathcal{N}_+(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2, truncated to the positive half-line $(0, \infty)$. The step sizes σ_c and σ_γ are initialized at 0.1. The acceptance probability for c is

$$\alpha_c = \min\left(1, \frac{p(c' \mid \cdot)}{p(c^{(s)} \mid \cdot)} \frac{q(c^{(s)} \mid c')}{q(c' \mid c^{(s)})}\right). \tag{19}$$

We recognize the strong posterior correlation between c, γ, and λ_u arising from the adaptive weight $w_i = (c + \tilde{u}_i)^{-\gamma}$, which directly scales the rate of the exponential prior on u_i. To mitigate poor mixing caused by this correlation, we implement adaptive tuning of the proposal variances every 100 iterations. The target acceptance rate is set to the interval $[0.2, 0.4]$: if the running acceptance rate falls below 0.2, the step size is reduced by 10%; if above 0.4, it is increased by 10%. Step sizes are bounded in $[0.01, 0.5]$ to avoid divergence or stagnation. This adaptive strategy substantially improves chain mixing. Convergence is monitored via trace plots, effective sample sizes (ESS > 200), and the Gelman-Rubin statistic $\hat{R} < 1.1$ across multiple chains. Simulation studies confirm stable mixing even under high correlation.

Through the above full conditional distributions, we are able to construct an efficient MCMC sampler. After the sampling is complete, we can analyze the generated posterior samples, typically using the posterior median as a point estimate for the parameters to obtain a robust estimate of the model parameters. The posterior samples can also be used to construct credible intervals. Algorithm 1 summarizes the procedure of the proposed Gibbs sampling.

3.3.2. Prior selection and numerical stability

The group-LASSO prior on β_g is induced hierarchically with $\lambda_{\beta,g} \sim \mathcal{G}(1.5, 0.1)$, a heavy-tailed distribution that permits both small and large regularization strengths. This weakly informative prior avoids over-shrinkage in moderate-dimensional problems. However, in high-dimensional settings ($p \gg n$), an extremely small posterior $\lambda_{\beta,g}$ can render the precision matrix

$$\Sigma_{\beta_g}^{-1} = \sigma_v^{-2} X_g^\top X_g + \sigma_v^{-2} \sqrt{m_g} \lambda_{\beta,g} I_{m_g}$$

nearly singular. To ensure numerical robustness, the Gibbs sampler incorporates two automatic safeguards.

(1) *Diagonal jitter*: A tiny constant $\epsilon = 10^{-4}$ is added to the diagonal of the precision matrix before inversion:

$$\Sigma_{\beta_g}^{-1} \leftarrow \Sigma_{\beta_g}^{-1} + \epsilon I_{m_g}.$$

(2) *Eigen-decomposition fallback*: If Cholesky decomposition fails, the inverse is computed via spectral decomposition with eigenvalues truncated at 10^{-8}:

$$\Sigma_{\beta_g} = V \operatorname{diag}(\max(D, 10^{-8}))^{-1} V^\top.$$

These mechanisms are triggered automatically within each MCMC iteration and do not alter the target posterior. Convergence diagnostics (trace plots of $\lambda_{\beta,g}$, effective sample sizes, and Gelman-Rubin statistics) in the simulations confirm stable mixing even when $p > n$.

Algorithm 1 Gibbs Sampling Algorithm

Set the total number of MCMC iterations N_{iter} and the number of burn-in iterations N_{burnin}.

for $s = 1, \ldots, N_{\text{iter}}$ **do**

Given the state at iteration $s-1$, generate the state for iteration s as follows:

1. For $g = 1, \ldots, G$, sample $\beta_g^{(s)}$ from $\mathcal{N}_{m_g}\left(\mu_{\beta_g}^{(s-1)}, \Sigma_{\beta_g}^{(s-1)}\right)$ in (12).

2. For $i = 1, \ldots, n$, sample $u_i^{(s)} \sim \mathcal{N}_+(\ldots)$ from the truncated normal distribution (13).

3. Sample $\sigma_v^{-2(s)}$ from the Gamma distribution (14).

4. For $g = 1, \ldots, G$, sample the group regularization parameter $\lambda_{\beta,g}^{(s)}$ from the Gamma distribution (15).

5. Sample the inefficiency regularization parameter $\lambda_u^{(s)}$ from the Gamma distribution (16).

6. Sample c and γ from (17) using an MH algorithm.

7. Update the weights $w_i^{(s)} \leftarrow (c^{(s)} + \tilde{u}_i)^{-\gamma^{(s)}}$.

end for

After N_{iter} iterations, discard the burn-in period samples ($s \leq N_{\text{burnin}}$), and use the rest as the posterior samples to obtain the posterior estimation.

4. Simulation studies

To comprehensively evaluate the performance of the proposed Bayesian group LASSO framework, we conduct extensive Monte Carlo experiments that systematically compare our approach against existing methods across various data scenarios. The simulation study addresses three key research questions: (1) How does the group LASSO perform relative to traditional LASSO methods in SFM? (2) What is the impact of data sparsity on method performance? (3) How do different clustering algorithms for group identification affect the results? These experiments employ realistic data-generating processes to assess both coefficient estimation accuracy and inefficiency term recovery across multiple performance metrics.

We develop three distinct simulation scenarios to comprehensively address different practical challenges in SFM.

(1) The benchmark comparison (Section 4.1) uses a low-dimensional non-sparse setting ($p = 5$) identical to Tsionas (2023), enabling a fair comparison with the current state-of-the-art Bayesian LASSO method and demonstrating the benefit of group structure even when traditional LASSO is not expected to fail due to sparsity.
(2) The sparsity study (Section 4.2) examines high-dimensional scenarios ($p = 30$) with varying proportions of zero coefficients, testing variable selection capability.
(3) The dimensionality study (Section 4.3) fixes moderate sparsity (20%) but increases p up to 100, evaluating scalability and curse-of-dimensionality mitigation. This progressive design isolates the impact of grouping from sparsity and dimensionality effects.

While methods with pre-defined group penalties exist in the penalized regression literature, they rely on *a priori* group structure knowledge, which is rarely available in SFM applications with high-dimensional inputs. Consequently, we benchmark against the Bayesian LASSO of Tsionas (2023), the current standard for individual-level variable selection in SFM. This comparison highlights the incremental value of the data-driven grouping strategy in a realistic scenario where group structures must be learned unsupervised and solely from the input variables.

4.1. Comparative evaluation against benchmark method

We begin by comparing our proposed approach against the LASSO-based method developed by Tsionas (2023), which serves as the current state-of-the-art for penalized SFM. This comparison is designed to demonstrate the advantage of incorporating the grouping structure for estimating inefficiency in low-dimensional and non-sparse scenarios. Although $p = 5$ is small and LASSO is not primarily designed for variable selection here, we deliberately follow the exact setting of Tsionas (2023) to ensure a fair benchmark comparison. More importantly, even in this non-sparse case, standard LASSO tends to over-shrink the regression coefficients because it mistakenly treats part of the one-sided inefficiency term u_i as reducible signal, leading to severely biased β estimates (often inflated variance by orders of magnitude). In contrast, the group LASSO protects correlated inputs within the same group from excessive shrinkage, resulting in dramatically lower MSE_β (1–3 orders of magnitude improvement).

Following the settings of Tsionas (2023), we simulate data according to the stochastic frontier model (1), where the experimental parameters are specified as follows:

- $x_i \in \mathbb{R}^5$ with each component drawn independently from Uniform$(0, 1)$.
- $\beta = (0.25, 0.25, 0.25, 0.25, 0.25)^\top$, representing a balanced contribution from all inputs.
- $v_i \sim \mathcal{N}(0, \sigma_v^2)$ with three noise levels representing different data quality scenarios: (I) low noise: $\sigma_v = 0.1$. (II) medium noise: $\sigma_v = 0.25$. (III) high noise: $\sigma_v = 0.5$.
- $u_i \sim$ Exponential(1), ensuring realistic inefficiency magnitudes.
- $n \in \{100, 250, 500, 1000\}$ to evaluate the performance across different sample sizes.
- Each experimental configuration is replicated 1000 times.

Each simulated dataset is analyzed using Bayesian LASSO method of Tsionas (2023) and the proposed Bayesian group method,where variables are partitioned into two groups by four clustering algorithms: hierarchical clustering (`hclust`), K-means clustering (`kmeans`), model-based clustering (`model`) and partitioning around medoids (`pam`). We evaluate method performance using the mean squared error (MSE) of regression coefficient estimates

($\text{MSE}_\beta = \frac{1}{p}\|\hat{\beta} - \beta\|_2^2$), the mean absolute error (MAE) of inefficiency estimates ($\text{MAE}_u = \frac{1}{n}\sum_{i=1}^n |\hat{u}_i - u_i|$ and $\text{MAE}_r = \frac{1}{n}\sum_{i=1}^n |e^{-\hat{u}_i} - e^{-u_i}|$) and regression sum of squares ($\text{SSR} = \sum_{i=1}^n (y_i - x_i^\top \hat{\beta} + \hat{u}_i)^2$). We use MSE for β because it is more sensitive to large deviations in continuous parameter estimates (Fan & Li, 2001). For the non-negative inefficiency $u_i \sim$ Exponential(1), which exhibits a heavy right tail, MAE is more robust than MSE against outliers and is widely adopted in SFM literature for inefficiency evaluation (Horrace et al., 2022; Jondrow et al., 1982; Tsionas, 2023).

The results are shown in Figure 1, in which several important findings demonstrate the superiority of the proposed Bayesian group LASSO approach.

(1) The Bayesian group LASSO methods consistently outperform the LASSO method across all metrics, with particularly dramatic improvements in estimation accuracy. The MSE_β improvements span 1–3 orders of magnitude. For instance, in the scenario of ($n = 100, \sigma_v = 0.1$), the LASSO achieves $\text{MSE}_\beta = 1454.60 \times 10^{-4}$, while group LASSO methods average approximately 1.8×10^{-4}, a nearly 800-fold improvement. This striking difference arises because LASSO shrinks all coefficients uniformly and confounds the one-sided inefficiency with random noise, producing excessively large bias in finite samples. Group LASSO, by contrast, performs coordinated shrinkage within correlated groups, preserving true signal and yielding near-oracle performance even without sparsity.
(2) The superior coefficient estimates translate directly into better inefficiency recovery. Under moderate conditions ($n = 500$, $\sigma_v = 0.25$), the group LASSO reduces MAE_u from 0.727 (LASSO) to approximately 0.213, a 70% improvement that indicates more accurate identification of production inefficiencies.
(3) Under high noise conditions ($\sigma_v = 0.5$), group LASSO maintains its substantial advantage, with MSE_β remaining more than two orders of magnitude lower than LASSO. This robustness stems from the group-wise shrinkage mechanism, which pools information within groups to better distinguish signal from noise.
(4) The four clustering methods show remarkably consistent performance, with standard deviations in MSE_β typically below 0.001. This stability indicates that automatic group identification is reliable across different algorithmic approaches, supporting the practical applicability of our framework.

4.2. Effects of sparsity level

Before applying the group LASSO, we first identify groups using unsupervised clustering algorithms on the design matrix, without using response information. This ensures that grouping is data-driven and reflects inherent correlations among input variables.

To evaluate performance across different levels of coefficient sparsity, we conduct a systematic sparsity comparison study. The simulation framework is extended to higher dimensions to Model (1), where $x_i \in \mathbb{R}^{30}$ with $x_{i,j} \sim$ Uniform(0, 1). The sparsity level s (the proportion of non-zero coefficients) is set as 0, 0.1, 0.2, 0.3, 0.5 and 0.8, respectively. The values of non-zero coefficients are set as 0.25. σ_v is 0.1, and the sample size is 200. For the group LASSO, variables are partitioned into $G = 5$ groups using five clustering methods (`hclust`, `kmeans`, `model`, `pam` and spectral clustering (`spectral`)).

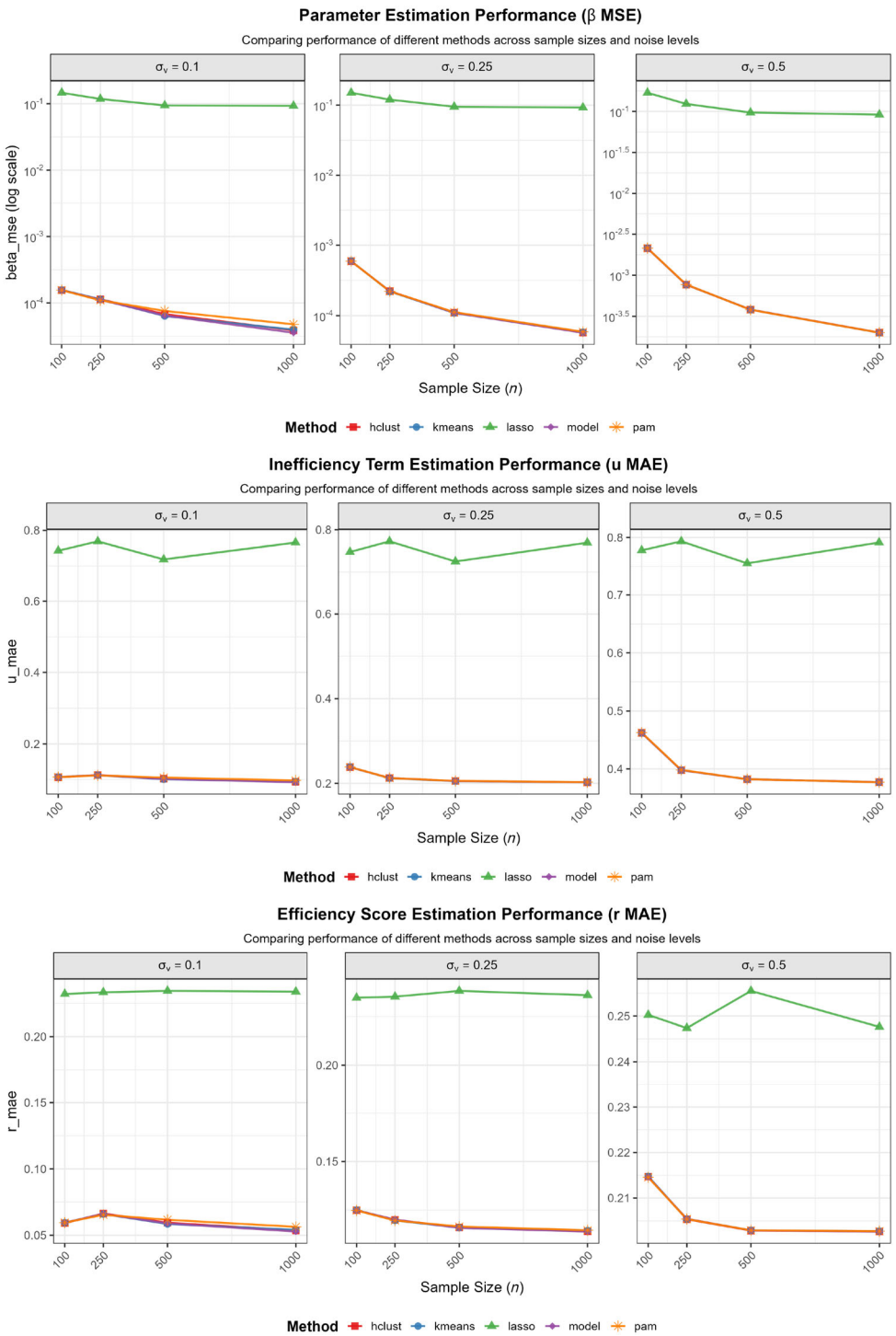


Figure 1. Performance comparison across different sample sizes and noise levels. The 3 × 3 grid compares five methods (`hclust`, `kmeans`, `LASSO`, `model`, `pam`) across three noise levels (columns) and three performance metrics (rows). The first row shows MSE for parameter estimates (MSE_β, log scale), the second row displays MAE for inefficiency terms (MAE_u), and the third row presents MAE for efficiency values (MAE_r).

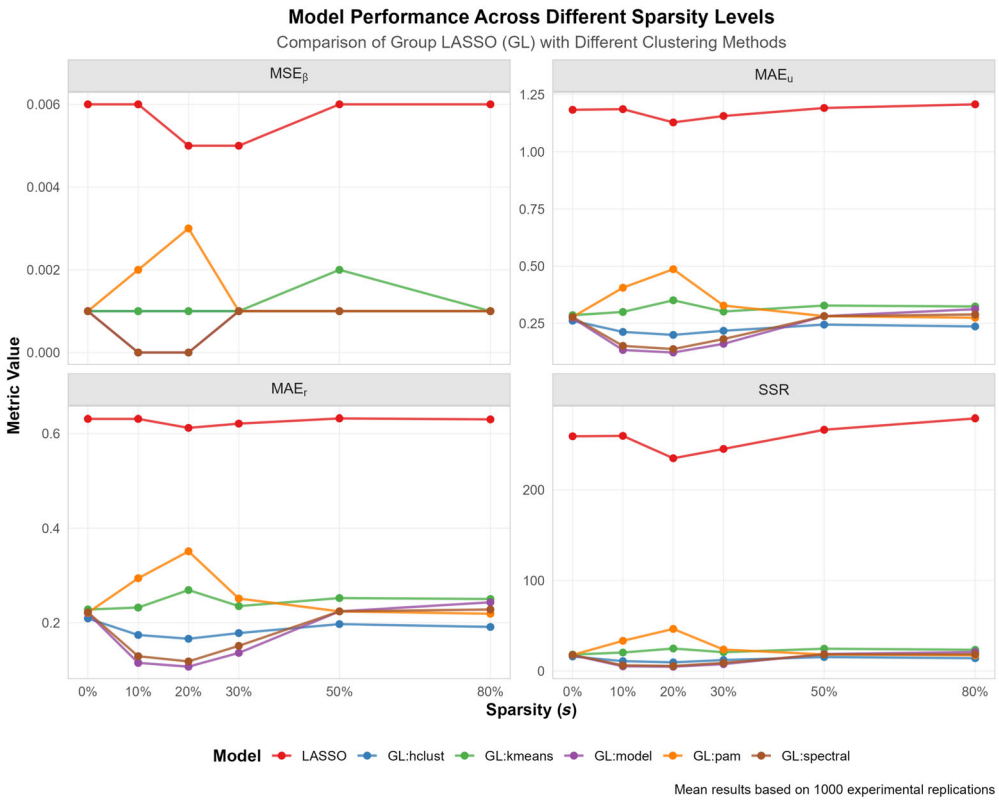


Figure 2. Performance comparison across sparsity levels (based on 1000 Monte Carlo replications) under different methods.

Figure 2 demonstrates the results of group LASSO as well as LASSO method under different levels of sparsity. The differences among the five group LASSO variants in the figure arise directly from the choice of clustering algorithm used to define the groups. As can be seen in Figure 2, group LASSO maintains substantial advantages over LASSO across all sparsity levels, which suggests that group LASSO is particularly well-suited for efficiency analysis applications, where the number of truly important input groups is typically small relative to the total number of measured variables. Different clustering algorithms show varying performance across sparsity levels. At low sparsity ($s \leq 0.2$), `model` and `spectral` clustering perform optimally, likely due to their sophisticated handling of high-dimensional, sparse data structures. For moderate to high sparsity levels, `hclust` demonstrates consistent robustness, while `kmeans` and `pam` exhibit greater variability.

The selection of the clustering algorithm is crucial in the group LASSO framework, as it directly determines the quality of the variable grouping and the efficacy of group-level sparsity enforcement. We briefly elaborate on the five methods as follows.

- `hclust` (hierarchical clustering) naturally captures nested structures and performs robustly across sparsity levels due to its agglomerative nature.
- `kmeans` is computationally efficient but sensitive to initialization and spherical cluster assumptions, leading to higher variability.

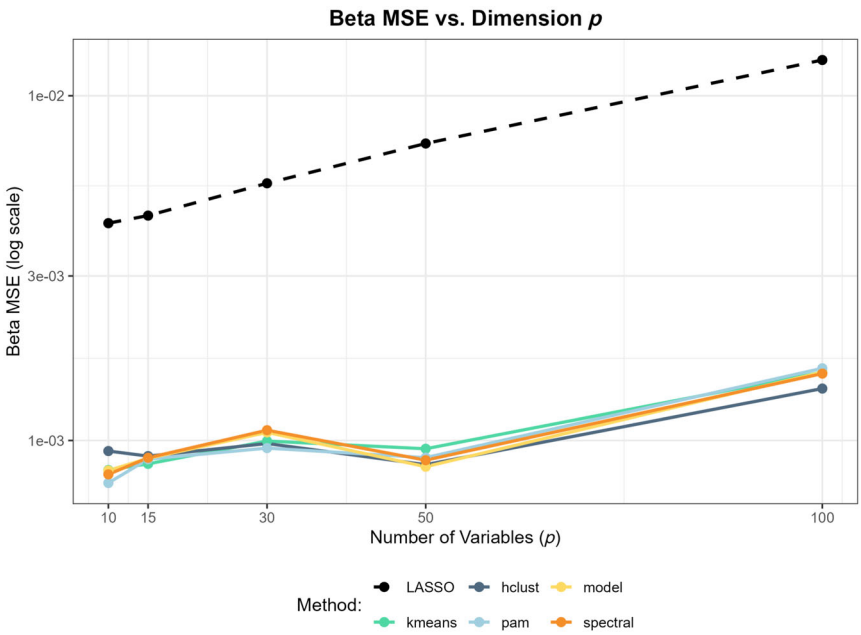


Figure 3. MSE of $\boldsymbol{\beta}$ estimates versus dimension p.

- `model` (model-based clustering using Gaussian mixtures) excels in low-sparsity regimes by flexibly modelling complex distributions.
- `pam` (partitioning around medoids) enhances robustness to outliers compared to `kmeans`.
- `spectral` clustering leverages eigenvectors of the similarity graph, making it particularly powerful for detecting non-convex clusters in high-dimensional sparse data.

These methodological distinctions clarify why `model` and `spectral` perform best at low sparsity levels ($s \leq 0.2$), while `hclust` maintains stable performance even as signal density increases ($s \geq 0.3$). In practice, the clustering method should be selected according to the underlying correlation structure and sparsity level of the data.

4.3. Dimensionality comparison

This experiment evaluates the performance of different methods across varying dimensions while maintaining a fixed sparsity level. We set the proportion of non-zero coefficients at 0.2 and vary the dimensionality $p \in \{10, 15, 30, 50, 100\}$ to assess performance in high-dimensional settings. The data generating process is as Model (1), where $x_i \in \mathbb{R}^p$ with $x_{i,j} \sim$ Uniform$(0, 1)$, β has 20% non-zero elements, each equal to 0.25, and other settings are the same as Section 4.2.

Figures 3 and 4 display the MSE of the coefficient estimates (MSE$_\beta$) and the MAE of the inefficiency estimates (MAE$_u$) under varying dimensionality levels, comparing the performance of LASSO and group LASSO combined with different clustering strategies.

As shown in Figure 3, the group LASSO significantly improves the estimation accuracy of β, achieving MSE values at least one order of magnitude lower than those of LASSO.

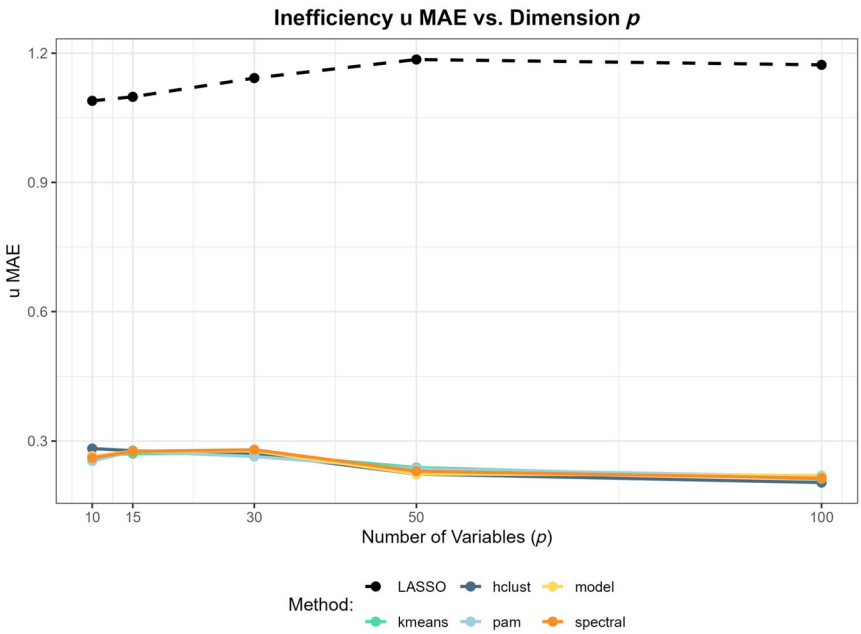


Figure 4. MAE of inefficiency term u estimates versus dimension p.

This improvement stems from its ability to exploit variable correlations through group structures, thereby mitigating bias in high-dimensional settings. Although the MSE increases with dimensionality p, group LASSO consistently outperforms LASSO across all cases. At low dimensionality, the performance of different group LASSO variants is comparable. However, as p increases, the `hclust`-based grouping strategies outperform the others, likely due to its better noise resilience.

Figure 4 shows that the MAE in estimating u is substantially lower than the baseline and exhibits a decreasing trend as p increases–from 0.28 to 0.22–suggesting that additional covariates aid in disentangling inefficiency from noise. While the choice of clustering method has a relatively minor impact, `spectral` and `hclust` clustering show a slight advantage at higher dimensions.

Overall, these results highlight that incorporating group structure not only alleviates the curse of dimensionality but also enhances estimation robustness and efficiency accuracy in high-dimensional environments.

4.4. Robustness and sensitivity analysis

We conduct two additional simulation studies to assess the robustness of the Bayesian group LASSO framework, specifically examining its sensitivity to hyperparameter settings and outliers. This evaluation is particularly important given the method's reliance on DEA scores, which can be influenced by extreme values.

4.4.1. Hyperparameter sensitivity

The proposed hierarchical model introduces hyperparameters for the adaptive weights, specifically (a_c, b_c) for the parameter c, and (a_γ, b_γ) for the scaling parameter γ. To assess

the sensitivity to these prior specifications, we compare the default setting used in previous sections against two alternative configurations.

- **Base (Default):** $a_c = 2, b_c = 1, a_\gamma = 3, b_\gamma = 1$. This setting represents a moderate informative prior, aiming to balance data-driven updates with regularization.
- **Prior A:** $a_c = 1, b_c = 0.5, a_\gamma = 1, b_\gamma = 1$. A weakly informative prior simulates a scenario almost entirely dominated by the data.
- **Prior B:** $a_c = 5, b_c = 2, a_\gamma = 5, b_\gamma = 1$. A strongly informative prior concentrates prior mass more tightly.

Figure 5 presents performance trajectories across sample sizes under different prior settings. Interestingly, 'Prior A' produces the lowest estimation errors for both β and u, suggesting that a more diffuse prior allows the model to fit the specific data realization more closely in this simulation. Nevertheless, we retain the moderately informative 'Base' setting as the recommended default for two reasons. First, although 'Prior A' performs best in this controlled setting, the 'Base' prior offers necessary regularization to guard against overfitting in real-world applications where the signal-to-noise ratio is unknown. Second, the performance gap between 'Base' and 'Prior B' is narrow, and 'Base' consistently achieves low MSE/MAE values, demonstrating that the method functions reliably without requiring extensive hyperparameter tuning. The consistent performance patterns across clustering methods further confirm the structural robustness of the approach.

4.4.2. Robustness to outliers

Since DEA scores utilize the convex hull of the data, they can be sensitive to outliers. To test whether this sensitivity propagates bias into the SFM estimation, we introduce high-leverage points into the covariate space. Specifically, for 5% of the sample, we magnify the input variables by a factor of 10, generating the response variable based on these extreme inputs. This creates a high-heterogeneity scenario with potential leverage points.

The results, displayed in Figure 6, reveal a notable phenomenon: the model performance in the presence of outliers (blue curve) is superior to the benchmark case (red curve), particularly for MSE_β. This counter-intuitive result can be attributed to the nature of the introduced outliers. Because these points strictly follow the data-generating process, they act as *good leverage points*. They expand the range of the regressors, thereby increasing the signal-to-noise ratio and allowing the regression plane to be estimated with greater precision. Crucially, this result confirms the robustness of our approach. The method effectively utilizes the structural information contained in extreme observations rather than breaking down due to numerical instability or DEA-induced bias. The parallel trends in MAE_u indicate that the inefficiency recovery remains stable even when the covariate space is severely distorted.

5. Real data analysis

5.1. Data

To evaluate the practical effectiveness of our proposed methodology, we conduct an empirical analysis using real-world banking data. The dataset is sourced from the Call Schedule RCO 09302024 (1 of 2), published by the U.S. Federal Financial Institutions Examination

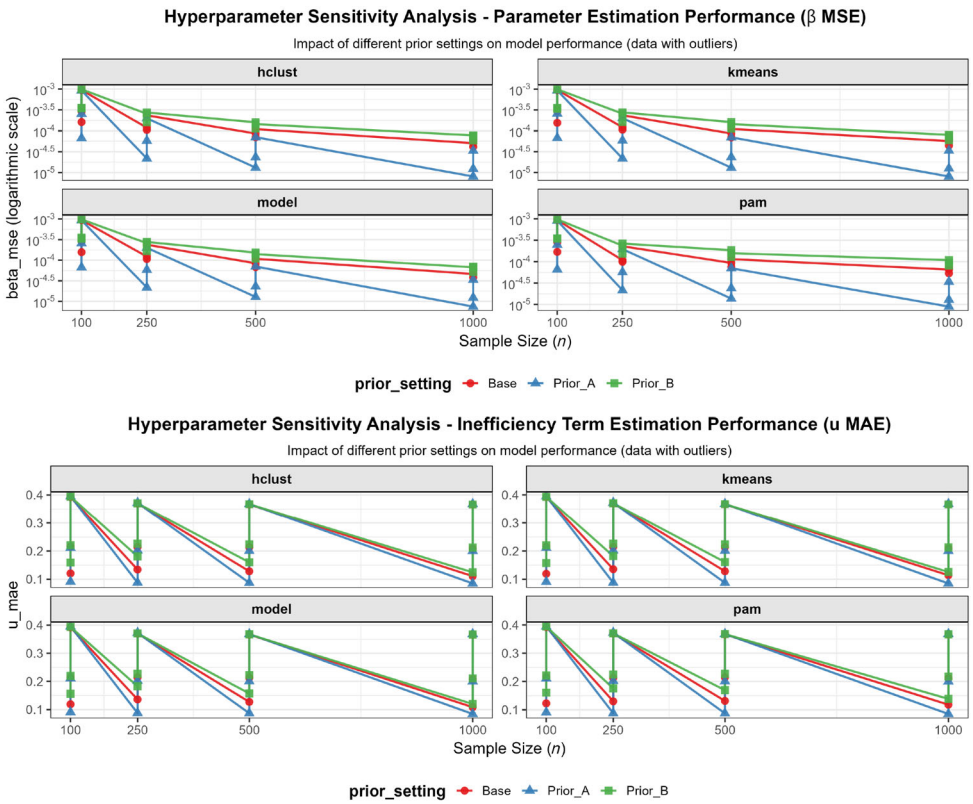


Figure 5. Sensitivity analysis of hyperparameters comparing three priors across different initialization strategies.

Council (FFIEC). This comprehensive dataset contains quarterly financial reports for commercial banks and serves as the foundation for regulatory risk assessment and supervisory evaluation.

We focus on 4752 financial institutions as of the third quarter of 2024. These institutions collectively account for 78% of total industry assets, making them representative of mainstream banking sector trends and ensuring the broad practical relevance of our findings. To maintain focus on core domestic banking operations, we exclusively utilize variables from the RCON series, which capture domestic activities, while excluding RCFD variables that represent international business operations.

The variable pool comprises 39 carefully selected RCON variables (listed in Table A1), chosen based on established theoretical foundations in banking efficiency literature. Following the seminal work of Berger and Humphrey (1997), who emphasized the critical importance of incorporating core balance sheet components (such as deposits, loans, and capital) as input and output variables, our selection captures the essential elements of a bank's production process and risk management framework. This approach ensures theoretical completeness, operational integrity, and high data quality. We employ RCON2170 as the dependent variable, which serves as a comprehensive measure of a bank's total asset size. This choice is theoretically consistent with the intermediation approach pioneered by Sealey

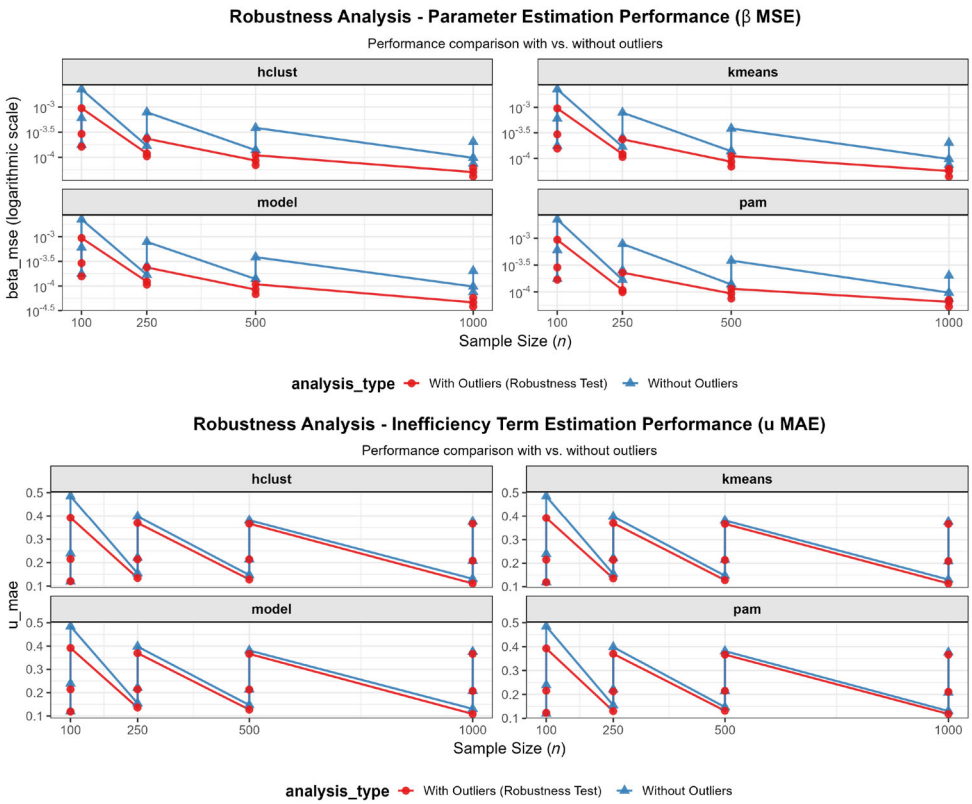


Figure 6. Robustness analysis comparing model performance with and without outliers.

and Lindley (1977), which views banks as financial intermediaries that transform funding sources into loans and other earning assets.

5.2. Model specification

To assess the impact of grouping strategy, we implement a comparative analysis contrasting LASSO (serving as the baseline method) with our proposed group LASSO approach. Variable groupings are generated through a data-driven approach employing five distinct clustering algorithms: K-means (`kmeans`), hierarchical clustering (`hclust`), partitioning around medoids (`pam`), model-based clustering (`model`), and spectral clustering (`spectral`). All variables undergo standardization using Z-score normalization to ensure comparability across different scales. We benchmark against standard Bayesian LASSO rather than predefined group methods because the latter assume known groupings, which contradicts the goal of unsupervised, data-driven group discovery. The substantial gains reported thus reflect the pure contribution of automated clustering in a comparable Bayesian SFM framework.

The estimation process operates within a comprehensive Bayesian framework designed to incorporate prior information about bank efficiency levels. As described in Sections 2 and 3, adaptive priors for inefficiency terms are constructed using pre-computed DEA inefficiency scores. For this analysis, $u_{\text{dea},i} = 1 - \text{eff}_i$ is obtained by the BCC model to the 38 RCON input variables and RCON2170 as the output. These scores then inform the adaptive prior weights

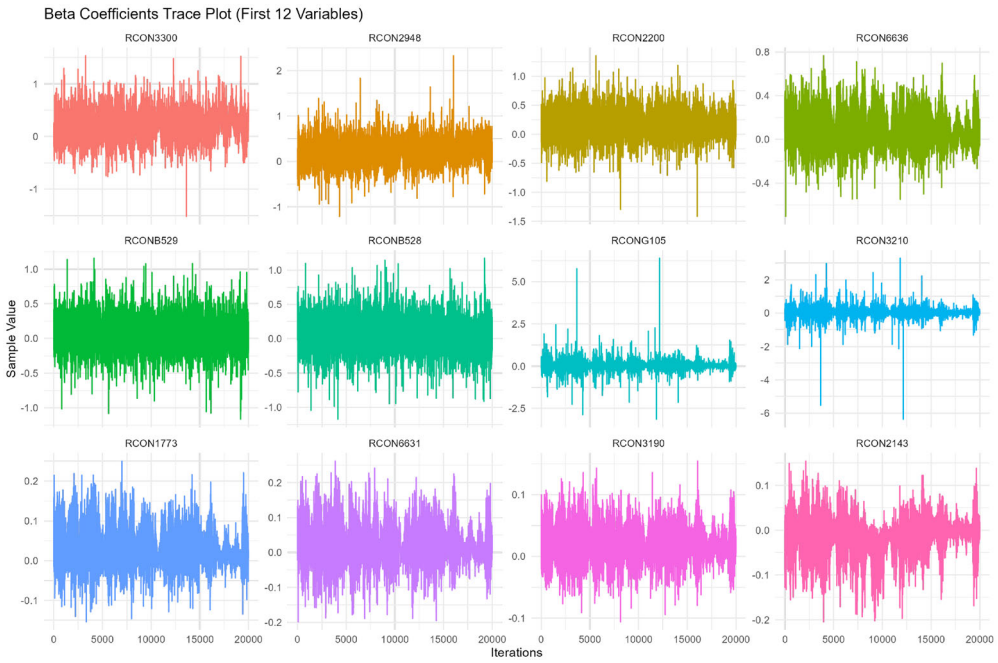


Figure 7. MCMC convergence diagnostics. The trace plots for main predictors β exhibit rapid mixing and stable fluctuation, indicating good convergence.

$w_i = (c + u_{\mathrm{DEA},i})^{-\gamma}$. Posterior samples are generated using the Gibbs sampling algorithm with the following specifications: 105,000 total iterations, 5000 burn-in periods, and retention of every fifth sample, yielding 20,000 independent posterior samples for inference.

Model performance is assessed across three dimensions: the regression coefficients β (significance and economic intuition), the posterior distribution of the inefficiency term u (converted to technical efficiency with e^{-u}), and the SSR (a goodness-of-fit metric).

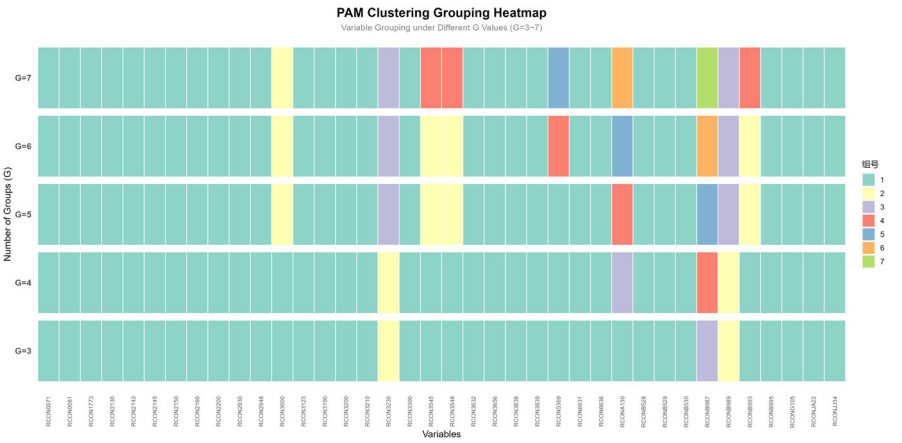
To ensure the reliability of the Bayesian inference, we conducted a diagnostic of the MCMC chains. Figure 7 presents the trace plots for the key regression coefficients β. These plots exhibit rapid mixing and fluctuate stably around their means, indicating good convergence.

5.3. Results

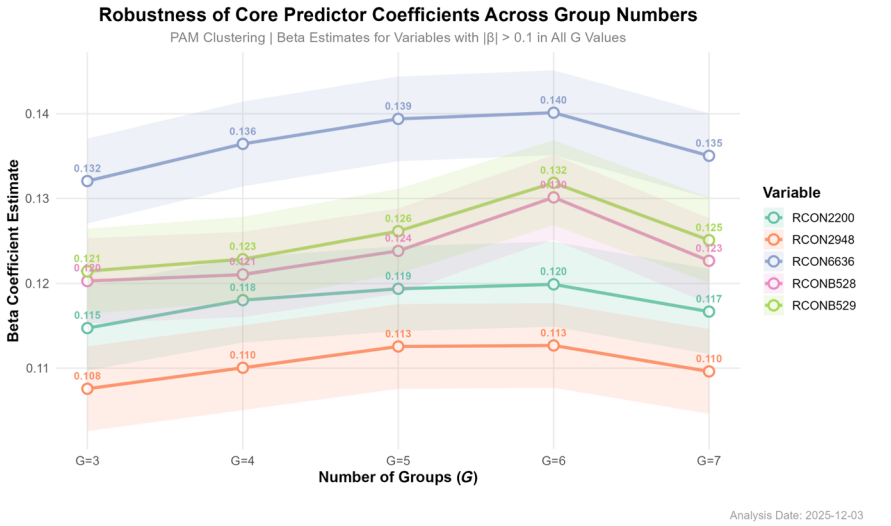
5.3.1. Variable grouping structure and sensitivity analysis

Since unsupervised clustering relies on a pre-specified number of clusters (G), we conducted a sensitivity analysis by varying G from 3 to 7 using the PAM algorithm. This analysis serves two purposes: (i) to examine how the PAM algorithm partitions the banking variables under different values of G, and (ii) to assess the robustness of the main findings.

The results are presented in Figure 8. Panel (a) displays the grouping heatmap, offering a direct visualization of the variable clusters. The liability-side variables (left side, Group 1) consistently form a stable cluster across all values of G, indicating strong homogeneity in funding structures. In contrast, the asset-side variables (right side) display a hierarchical splitting pattern as G increases, revealing finer distinctions within loan portfolios. Panel (b)



(a) PAM Grouping Heatmap ($G = 3 \sim 7$)



(b) Robustness of Core Coefficients

Figure 8. Analysis of grouping structure and robustness. Panel (a) explicitly visualizes the variable clusters generated by PAM, showing stable liability groups and evolving asset groups. Panel (b) demonstrates that the coefficients of core variables are robust to the choice of *G*. (a) PAM Grouping Heatmap ($G = 3 \sim 7$) and (b) Robustness of Core Coefficients.

provides further evidence of the robustness of the proposed model by displaying the estimated coefficients for the core predictors (those with $|\beta| > 0.1$) across different choices of G. Despite changes in the group structure, the estimated economic effects of the key drivers remain remarkably stable. Based on the stability of the coefficients around $G = 5$ and the interpretability of the corresponding clusters in Panel (a), we adopt $G = 5$ for the main analysis.

5.3.2. Estimation results

Using $|\beta| > 0.1$ as the threshold for variable selection, all models consistently identify six core variables that exhibit strong economic relevance in banking operations. These variables include `RCON2200` (Total deposits, $\beta \approx 0.119$), `RCON2948` (Total liabilities, $\beta \approx 0.113$), `RCON3300` (Total liabilities, reserves, and capital, $\beta \approx 0.103$), `RCON6636` (Interest-bearing deposits from domestic and foreign institutions, $\beta \approx 0.139$), `RCONB528` (Loans net of unearned income, $\beta \approx 0.124$), and `RCONB529` (Loans net of income and reserves, $\beta \approx 0.126$). The threshold of $|\beta| > 0.1$ is selected based on standardized regression coefficients (both X and y are Z-score normalized), where a coefficient magnitude exceeding 0.1 corresponds to a meaningful economic effect (approximately 10% of a standard deviation change in the outcome per standard deviation change in the predictor). This criterion aligns with conventional benchmarks in applied econometrics for identifying substantive predictors in high-dimensional settings (Gelman, 2008; McElreath, 2020).

These results reinforce fundamental banking principles. The positive coefficients on liability-side variables (e.g., deposits and institutional funding) validate the notion that deposit mobilization enhances credit creation capacity. The positive coefficients on net loan variables suggest that, after controlling for deposits, liabilities, capital, and other financial indicators, larger loan and lease portfolios are positively associated with the total-asset frontier.

Based on the posterior samples, the distribution of the inefficiency term u is estimated using kernel density methods, as illustrated in Figure 9. As shown, the posterior distribution of u under the `hclust`, `pam`, `model`, and `spectral` methods consistently displays a pronounced bimodal pattern. This bimodal structure reflects substantial heterogeneity in technical efficiency among large U.S. commercial banks, revealing two distinct efficiency clusters. The first mode, centred around $u \approx 0.2$ ($\text{TE} = e^{-0.2} \approx 0.82$), represents high-efficiency institutions (approximately 60% of the sample), typically large, digitally advanced banks with strong retail franchises. The second mode, at $u \approx 0.8$ ($\text{TE} \approx 0.45$), captures low-efficiency banks (about 40%), often characterized by reliance on traditional corporate lending, higher operating costs, and greater exposure to economic cycles. This polarization is consistent with prior banking literature documenting persistent efficiency gaps (Berger & Humphrey, 1997; Wheelock & Wilson, 2012), and is effectively uncovered by group LASSO due to its ability to preserve correlated variable structures within efficiency drivers. In contrast, the posterior distribution derived from the standard LASSO method exhibits a unimodal shape, suggesting that it fails to effectively distinguish between high- and low-efficiency institutions.

To quantitatively assess model performance, we computed the SSR values under all methods, which are presented in Table 1. From Table 1, we find that all group LASSO methods significantly outperform traditional LASSO in terms of SSR values. The `pam`-based group LASSO emerges as the top performer with an SSR of 273.92, representing a remarkable 75.8% improvement over traditional LASSO (SSR = 1131.31). The performance hierarchy among clustering methods follows `pam` > `hclust` > `model` > `spectral` > `kmeans`. The superior performance of `pam` stems from its exceptional ability to capture local data structures while maintaining robustness to outliers and noise.

6. Conclusion

This study proposes a Bayesian group LASSO framework that addresses key limitations in traditional efficiency evaluation methods by integrating the strengths of SFM and DEA. The

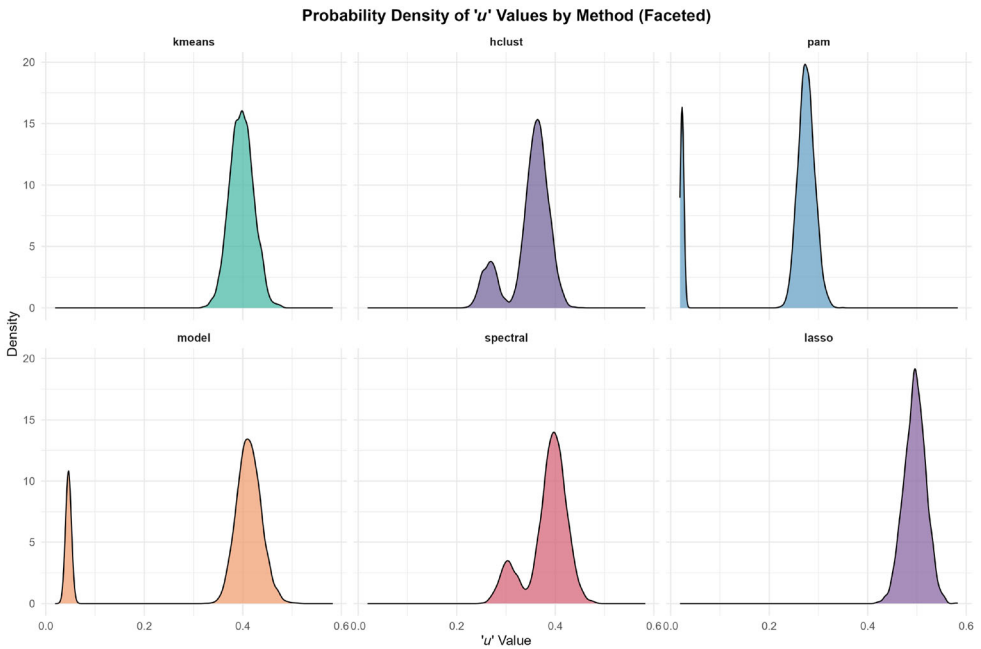


Figure 9. Posterior distribution of inefficiency term *u* under different methods.

Table 1. SSR comparison under different methods.

Method	SSR	Improvement (%)
LASSO	1131.31	–
Group LASSO (`pam`)	273.92	75.8
Group LASSO (`hclust`)	310.45	72.6
Group LASSO (`model`)	335.67	70.3
Group LASSO (`spectral`)	358.92	68.3
Group LASSO (`kmeans`)	425.18	62.4

framework introduces two main innovations: an adaptive prior mechanism that leverages DEA information to avoid distributional misspecification, and group LASSO regularization that respects economic variable relationships while handling high-dimensional data effectively.

Monte Carlo simulations demonstrate that our method consistently outperforms traditional LASSO across various scenarios in both coefficient and inefficiency estimation accuracy. The empirical application to U.S. commercial banking data reveals core financial drivers of bank output and identifies a striking bimodal efficiency distribution, indicating clear industry segmentation into high-efficiency and low-efficiency clusters. Group LASSO methods, particularly those with data-driven clustering, achieve substantially lower prediction errors than traditional approaches, confirming the importance of incorporating variable group structures.

Future research could extend this cross-sectional framework to dynamic panel data models to capture efficiency evolution over time, explore more sophisticated endogenous grouping mechanisms for greater flexibility, and apply the methodology to other industries

with complex production processes to establish broader generalizability. The framework provides a theoretically robust and practically effective tool for efficiency analysis that opens new avenues for understanding structural heterogeneity in economic data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research is supported by National Natural Science Foundation of China [grant numbers 72571246 and 72571198], Zhejiang Provincial Natural Science Foundation of China [grant number LZ24A010002], the Fundamental Research Funds for the Provincial Universities of Zhejiang [grant number 2025ZDAPY01], the Summit Advancement Disciplines of Zhejiang Province (Zhejiang Gongshang University—Statistics), and Collaborative Innovation Center of Statistical Data Engineering Technology & Application.

ORCID

Ancha Xu http://orcid.org/0000-0003-3289-2720

References

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems* [Doctoral dissertation]. Australian National University. https://doi.org/10.25911/5d78db4c25dbb

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*(9), 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078

Berger, A. N., & Humphrey, D. B. (1997). Efficiency of financial institutions: International survey and directions for future research. *European Journal of Operational Research*, *98*(2), 175–212. https://doi.org/10.1016/S0377-2217(96)00342-6

Chen, Y., Tsionas, M. G., & Zelenyuk, V. (2021). LASSO+DEA for small and big wide dataLasso+dea for small and big wide data. *Omega*, *102*, Article 102419. https://doi.org/10.1016/j.omega.2021.102419

Emili, S., & Galli, F. (2025). Modelling spatio-temporal dynamics in multi-output stochastic frontiers for the European agribusiness industry. *Journal of Agricultural, Biological and Environmental Statistics*, *30*(2), 540–575. https://doi.org/10.1007/s13253-025-00680-y

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group LASSO and a sparse group LASSO. arXiv preprint. https://arxiv.org/abs/1001.0736arXiv:1001.0736.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873. https://doi.org/10.1002/sim.v27:15

Horrace, W. C., Jung, H., & Lee, Y. (2022). LASSO for stochastic frontier models with many efficient firms. *Journal of Business & Economic Statistics*, *41*(4), 1132–1142. https://doi.org/10.1080/07350015.2022.2110881

Jondrow, J., Lovell, C. A. K., Materov, I. S., & Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, *19*(2–3), 233–238. https://doi.org/10.1016/0304-4076(82)90004-5

Kumbhakar, S. C., & Tsionas, E. G. (2011). Some recent developments in efficiency measurement in stochastic frontier models. *Journal of Probability and Statistics*, *2011*(1), Article 603512. https://doi.org/10.1155/jpas.v2011.1

Lee, C. Y., & Cai, J. Y. (2020). LASSO variable selection in data envelopment analysis with small datasets. *Omega*, *91*, Article 102008. https://doi.org/10.1016/j.omega.2018.12.008

Li, J., Wang, Z., Li, R., & Wu, R. (2015). Bayesian group LASSO for nonparametric varying-coefficient models with application to functional genome-wide association studies. *The Annals of Applied Statistics*, *9*(2), 640–664. https://doi.org/10.1214/15-AOAS808

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

Parmeter, C. F., Prokhorov, A., & Zelenyuk, V. (2025). The post double LASSO for efficiency analysis. arXiv preprint. https://arxiv.org/abs/2505.14282arXiv:2505.14282.

Sealey, C. W., & Lindley, J. T. (1977). Inputs, outputs, and a theory of production and cost at depository financial institutions. *Journal of Finance*, *32*(4), 1251–1266. https://doi.org/10.1111/jofi.1977.32.issue-4

Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, *136*(1), 31–64. https://doi.org/10.1016/j.jeconom.2005.07.009

Simar, L., & Zelenyuk, V. (2020). Improving finite sample approximation by central limit theorems for estimates from data envelopment analysis. *European Journal of Operational Research*, *284*(3), 1002–1015. https://doi.org/10.1016/j.ejor.2020.01.036

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tsionas, M. G. (2003). Combining DEA and stochastic frontier models: An empirical Bayes approach. *European Journal of Operational Research*, *147*(3), 499–510. https://doi.org/10.1016/S0377-2217(02)00248-5

Tsionas, M. G. (2019). Robust Bayesian inference in stochastic frontier models. *Journal of Risk and Financial Management*, *12*(4), Article 183. https://doi.org/10.3390/jrfm12040183

Tsionas, M. G. (2023). Combining data envelopment analysis and stochastic frontiers via a LASSO prior. *European Journal of Operational Research*, *304*(3), 1158–1166. https://doi.org/10.1016/j.ejor.2022.04.029

Wheelock, D. C., & Wilson, P. W. (2012). Do large banks have lower costs? New estimates of returns to scale for U.S. banks. *Journal of Money, Credit and Banking*, *44*(1), 171–199. https://doi.org/10.1111/jmcb.2012.44.issue-1

Xu, X., & Ghosh, M. (2015). Bayesian variable selection and estimation for group LASSO. *Bayesian Analysis*, *10*(4), 909–936. https://doi.org/10.1214/14-BA929

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1430. https://doi.org/10.1198/016214506000000735

Appendix A

Table A1. List of 39 RCON variables (domestic operations).

Variable Code	Variable Name
RCON0071	Interest-bearing balances due from depository institutions
RCON0081	Non-interest-bearing cash and currency
RCON1773	Available-for-sale securities
RCON2130	Investments in unconsolidated subsidiaries and associated companies
RCON2143	Intangible assets
RCON2145	Premises, equipment and leased assets (including right-of-use assets)
RCON2150	Other real estate owned
RCON2160	Other assets
RCON2170	Total assets
RCON2200	Total deposits
RCON2930	Total other liabilities
RCON2948	Total liabilities
RCON3000	Minority interest in consolidated subsidiaries
RCON3123	Allowance for credit losses
RCON3190	Other borrowed money
RCON3200	Subordinated notes and debentures
RCON3210	Total shareholders' equity
RCON3230	Common stock
RCON3300	Total liabilities, allowances and capital accounts
RCON3545	Total trading assets
RCON3548	Trading liabilities
RCON3632	Undistributed profits and capital reserves
RCON3656	Direct or indirect real estate investments
RCON3838	Perpetual preferred stock and related surplus
RCON3839	Surplus
RCON5369	Loans and leases held for sale
RCON6631	Non-interest-bearing deposits
RCON6636	Total interest-bearing deposits in domestic and foreign offices
RCONA130	Other components of equity capital
RCONB528	Loans and leases, net of unearned income
RCONB529	Loans and leases, net of unearned income and allowance
RCONB530	Accumulated other comprehensive income
RCONB987	Securities purchased under agreements to resell
RCONB989	Securities sold under agreements to repurchase
RCONB993	Federal funds purchased in domestic offices
RCONB995	Federal funds sold in domestic offices
RCONG105	Total equity capital including noncontrolling interests
RCONJA22	Equity securities with readily determinable fair values
RCONJJ34	Held-to-maturity securities