Components Identification in Finite Mixture Model

Chen Ting[1] *Phoong Seuk Yen*[1]*

[1] Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia,

* Corresponding Author: phoong@fsmt.upsi.edu.my

Abstract

Finite Mixture Model (FMM) are widely utilized in statistical analysis to model unobserved heterogeneity within complex datasets. A key challenge in their implementation is selecting the optimal number of components, as this directly affects the model's interpretability, stability, and forecasting power. This study focuses on component selection strategies in FMMs, with particular emphasis on two widely used information criteria: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). While AIC favors model flexibility, BIC emphasizes parsimony and often performs better in finite samples. We apply the FMM framework to monthly data on China's rice trade, investigating the relationship among exchange rates, import/export prices, and trade volumes. Empirical findings suggest that a two-component model offers the best fit according to the BIC, revealing distinct market regimes likely shaped by structural conditions and policy interventions.

Keywords: Finite Mixture Model; components; Akaike Information Criterion; Bayesian Information Criterion; Exchange Rate; Agricultural Trade

DOI: 10.63802/grhas.v1.i3.44

1.Introduction

Finite Mixture Model (FMM) serve as a flexible statistical approach for modeling heterogeneous populations, where the observed data are assumed to arise from a mixture of multiple latent subgroups. Their use has expanded across various fields such as economics, finance, medicine, and genetics. In economic research, FMMs have proven effective in modeling regime shifts, market segmentation, and structural changes, enabling the detection of patterns not visible through standard econometric models (Sawadogo et al., 2024; Suk et al., 2021).

A fundamental challenge in applying FMMs lies in determining the appropriate number of components. Overestimating the number of components can lead to overfitting, resulting in unnecessarily complex models that generalize poorly. Conversely, underestimating it can oversimplify the data structure and obscure important latent patterns (Schlattmann, 2003). Therefore, robust model selection procedures are essential to balance model fit with parsimony.

Among the most commonly used tools for component selection are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria offer different perspectives: AIC tends to select models with better fit by allowing more complexity, while BIC imposes a stronger penalty on the number of parameters, favoring simpler models that perform well in smaller samples. Their comparative performance provides valuable guidance in empirical settings where the true data-generating process is unknown.

This study applies the FMM framework to analyze China's rice trade, focusing on how trade behavior may vary under different latent regimes driven by changes in exchange rates, import and export prices, and trade volumes. By identifying the optimal number of components through AIC and BIC, we aim to uncover hidden structures in rice trade patterns and interpret them within the broader context of economic policy and global market dynamics.

The remainder of this paper is organized as follows: Section 2 presents the theoretical foundation and applications of FMMs. Section 3 discusses model selection methods with a focus on AIC and BIC. Section 4 provides empirical evidence using China's rice trade data. Finally, Section 5 concludes the study and outlines implications for future research.

2.Literature Review

The FMM first emerged in modern statistical literature in the 19th century, introduced by Newcomb (Newcomb, 1886) as a method for handling outliers. A few years later, Pearson (Pearson, 1894) applied the method of moments to estimate FMM parameters, analyzing a dataset of a thousand crabs using two univariate Gaussian mixture distributions. This foundational work laid the groundwork for the broader adoption of FMM in data analysis, leading to its application across various probability distributions, such as Poisson, Weibull, and normal distributions.

Three decades have passed since McLachlan and Basford (McLachlan & Basford, 1988) published

their influential monograph, which underscored the potential of FMM for inference and clustering. Since then, this model have attracted significant research attention and have been widely utilized in diverse and evolving fields. As a result, the academic discourse surrounding FMM has expanded considerably, reflecting its growing impact and versatility in statistical modeling.

The application of FMM has extended into the fields of medicine and genetics. Haji Aghabozorgi and Eskandari (Haji Aghabozorgi & Eskandari, 2024) investigate the integration of FMM with Bivariate Ranked Set Sampling (BRSS) to refine clustering and parameter estimation in medical research, particularly in breast cancer studies. Their analysis of a breast cancer dataset illustrates notable advancements in distinguishing cancer subtypes, which play a crucial role in personalized treatment and prognosis.

In the realm of economics and finance, Schneider and Scharfenaker (Schneider & Scharfenaker, 2020) highlight the effectiveness of FMM in analyzing income distribution, demonstrating its ability to capture the underlying complexity and diversity in income data. By applying this model to household income datasets, they identify multiple latent groups corresponding to different income sources, such as wage employment, self-employment, and capital returns. Their research showcases FMM's capability to reveal hidden economic structures, offering deeper insights into income inequality and supporting the development of more targeted policy interventions.

In tourism research, Mortazavi and Lundberg (Mortazavi & Lundberg, 2020) utilized FMM to classify tourists in Venice based on their spending behaviors while accounting for hidden heterogeneity. Their analysis uncovered distinct spending patterns across categories such as accommodation, dining, and entertainment, offering a more precise understanding of tourist behavior. By capturing these variations, their study provided valuable insights for targeted marketing and service optimization. Their findings highlight the effectiveness of FMM in identifying unobserved heterogeneity and enhancing strategic tourism management.

In social science, Jenkins and Rios-Avila (Jenkins & Rios-Avila, 2023) utilize FMM to address the complexities of integrating survey data with administrative records. Their study introduces advanced estimation techniques designed to manage the inherent heterogeneity in such datasets. By identifying latent classes, they offer a more detailed understanding of population subgroups. This research makes a significant contribution by providing a robust analytical framework for linked data, uncovering hidden structures and relationships that conventional methods may overlook. Their approach proves particularly valuable for policy-making and advancing social science research.

FMM has also found applications in agriculture, as demonstrated by Fei et al. (Fei et al., 2023). Their research introduces an innovative method for assessing pollution levels and identifying high-risk zones for heavy metal contamination in soil. By integrating FMM with Bayesian Maximum Entropy (BME) theory, they develop a comprehensive

framework for environmental risk evaluation. This approach generates detailed maps that pinpoint areas susceptible to contamination, which is crucial for targeted soil remediation strategies.

In summary, the inherent flexibility of FMM has enabled its widespread application across various disciplines, including agriculture, tourism, medicine, finance, economics, and social sciences. These models have proven particularly useful in capturing latent heterogeneity, improving classification accuracy, and enhancing inference in complex datasets. The reviewed studies illustrate the effectiveness of FMM in uncovering hidden structures, modeling heterogeneous populations, and refining analytical approaches across diverse domains.

Despite extensive applications, a recurring challenge remains: how to correctly identify the number of components. An improper specification may lead to overfitting or underfitting, compromising both interpretability and predictive accuracy (Manole & Khalili, 2021; Schlattmann, 2003).

Multiple statistical criteria have been developed for component selection in FMM. Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) are the most widely used, offering penalized-likelihood based trade-offs between model complexity and fit. While AIC minimizes information loss, BIC emphasizes parsimony and model consistency as sample size grows (Agiakloglou & Tsimpanos, 2023).

Beyond AIC and BIC, other techniques have been proposed. The Integrated Completed Likelihood (ICL) criterion incorporates both likelihood and posterior uncertainty to penalize models that generate ambiguous classifications (Biernacki et al., 2002). Bootstrap likelihood ratio tests (Feng & McCulloch, 1996) provide a resampling-based inference framework, which has gained attention for its empirical robustness. Cross-validation approaches offer a data-driven way to assess out-of-sample predictive performance (Smyth, 2000). While no method is universally superior, each has distinct advantages depending on model purpose, sample size, and underlying data complexity.

Given the mixture of parsimony and empirical relevance required in this study, we prioritize AIC and BIC as practical, interpretable, and widely accepted criteria. However, the limitations of these criteria are acknowledged, and alternative methods are discussed to frame future extensions.

Although substantial work applies FMM to diverse domains, few empirical studies systematically address component selection in real-world agricultural trade under exchange rate volatility. Particularly for China’s rice trade, where global competitiveness, policy adjustments, and market openness have evolved dramatically since WTO accession, identifying latent regimes remains underexplored. Prior work often relies on traditional linear or panel models, overlooking hidden heterogeneity that mixture models can uncover. This study contributes to filling this gap by applying FMM to China's rice trade, incorporating both AIC and BIC, and explicitly investigating latent trade regimes associated with exchange rate movements.

3. Components Identification

3.1 Model Selection in FMM: AIC and BIC

A key challenge in applying FMM is determining the optimal number of components, as this choice directly influences the model's reliability, interpretability, and precision (Schlattmann, 2003). Accurately selecting the number of components is essential for ensuring robust and meaningful results, as both overfitting and underfitting may distort the model's representation of underlying data heterogeneity. To address this issue, model selection criteria serve as effective tools for identifying the most appropriate model structure by balancing model fit and complexity. Among the widely adopted approaches, the AIC and BIC are commonly used penalized-likelihood methods for model comparison (Chakrabarti & Ghosh, 2011).

3.2 Alternative Criteria: ICL, BLRT, and Cross-Validation

Although several alternative model selection approaches exist for determining the number of components in FMM, including Integrated Completed Likelihood (ICL), Bootstrap Likelihood Ratio Tests (BLRT), and Cross-Validation (CV), these methods were not applied in the present study due to both theoretical considerations and practical constraints that are closely related to the empirical context and dataset characteristics. The specific nature of China's rice trade data and the objectives of this research necessitated careful evaluation of these methods, and ultimately led to the prioritization of information-based criteria that balance both empirical suitability and computational efficiency.

3.3 Rationale for Excluding Alternative Methods

The ICL criterion incorporates a classification-based penalty term into the Bayesian Information Criterion (BIC), thereby favoring models that yield well-separated components with high posterior classification certainty (Biernacki et al., 2002). While ICL is particularly advantageous when the primary research objective centers on cluster membership and sharply delineated component boundaries, its conservative nature often leads to underfitting when mixture models exhibit some degree of component overlap (McLachlan & Peel, 2000). In the empirical context of China's rice trade, price and quantity fluctuations are simultaneously shaped by structural trade policies and unpredictable external shocks, resulting in an inherent degree of overlap between latent components. Applying ICL's stronger classification penalty in such a setting may oversimplify the actual market heterogeneity and inadvertently eliminate economically meaningful trade regimes. Since the core objective of this study lies in identifying latent trade structures and regime shifts in relation to exchange rate volatility—rather than assigning definitive cluster membership—the ICL criterion was deemed overly restrictive for the empirical goals of the current analysis, so it was not adopted.

Cross-validation, on the other hand, evaluates out-of-sample predictive performance by partitioning the

data into multiple training and validation subsets (Smyth, 2000). In principle, CV offers a data-driven framework that does not depend on asymptotic distributional assumptions and allows for model evaluation through empirical generalization. Nevertheless, its implementation in finite mixture modeling presents significant computational and methodological complications. Because mixture models are estimated via iterative algorithms such as the Expectation-Maximization (EM) algorithm, each partitioned dataset requires repeated model convergence for every candidate component structure. This greatly increases computational demand, especially for moderate-to-large datasets such as the present study, which contains 1,135 monthly observations. Furthermore, CV-based results are often sensitive to random partitioning, potentially producing unstable component selection outcomes unless an extensive number of replications are conducted. These concerns are amplified in longitudinal trade data, where temporal dependencies may compromise random data splits, thereby distorting model estimation and undermining stability. Given these factors, cross-validation was not adopted in this study due to both its computational infeasibility and its limited reliability in preserving consistent regime classification under complex time-dependent trade patterns.

The BLRT procedure evaluates competing nested models, specifically testing models with k components against those with $k-1$ components, by resampling the likelihood ratio statistic to obtain empirical p-values (Feng & McCulloch, 1996). While BLRT is recognized for its superior statistical power in finite sample contexts, its computational demands are considerably higher than other criteria, as it requires numerous bootstrap replications for each model comparison. Every bootstrap sample necessitates independent EM optimization, substantially increasing processing time and computational burden, particularly when evaluating multiple competing models across several variables simultaneously. Beyond these computational concerns, BLRT outcomes can also exhibit sensitivity to starting values, convergence difficulties, and parameter identification issues common to mixture models, especially when component separation is limited or variances are highly heterogeneous (McLachlan & Peel, 2000). In preliminary trials conducted during the early stages of this research, BLRT yielded unstable and inconsistent results across multiple initializations, reinforcing our decision to avoid reliance on this method in favor of more stable information-based selection criteria.

3.4 Summary of Selection Strategy

In summary, while ICL, BLRT, and CV provide valuable alternative perspectives in certain mixture model applications, their respective limitations—whether related to sensitivity to overlapping components (ICL), computational feasibility and partition stability (CV), or convergence robustness (BLRT)—rendered them less appropriate for the present study's objectives and empirical data structure. Given these challenges, AIC and BIC, both of which are widely accepted, computationally efficient, and theoretically well-established within the finite mixture modeling literature, were

ultimately selected as the most suitable and practical tools for identifying the latent component structures that characterize China's rice trade.

Preliminary trials using CV and BLRT during the early analysis phase yielded unstable component solutions or excessive computation times, supporting the decision to rely primarily on AIC and BIC.

Akaike proposed the AIC, which became the first extensively applied criterion for model selection in statistical analysis (Akaike, 1974). AIC is generally expressed as:

$$AIC = -2lnL(\hat{X}) + 2r \qquad (1)$$

where $L(\hat{X})$ denotes the maximum likelihood estimate, while r signifies the total number of parameters in the model. In a more compact form, the AIC formula can be expressed as:

$$AIC = -2ln(m) + 2r \qquad (2)$$

where m represents the maximum value of the likelihood function and r denotes the number of parameters in the model. The model with the smallest AIC value is regarded as the optimal selection.

Schwarz (Schwarz, 1978) proposed the BIC, which is formulated as:

$$BIC = -2lnL(\hat{X}) + kln(n) \qquad (3)$$

where $L(\hat{X})$ represents the maximized value of the likelihood function, k represents the number of parameters in the model, and n stands for the sample size. The BIC equation can also be expressed as:

$$BIC = -2lnL(m) + kln(n) \qquad (4)$$

where m denotes the maximum likelihood function value, k represents the number of parameters in the model, and n refers to the sample size. The model with the lowest BIC value is considered the best choice.

AIC and BIC are both commonly used model selection criteria, yet they differ in their theoretical basis and the way they penalize model complexity (Kuha, 2004). AIC, rooted in information theory, seeks to minimize information loss by balancing model fit and complexity. It imposes a penalty on the number of parameters but does not explicitly account for sample size. Designed to assess the relative quality of models within a given dataset, AIC generally favors more complex models, especially when the sample size is large, as the penalty for additional parameters remains relatively modest.

Conversely, BIC is based on Bayesian probability theory and explicitly incorporates sample size into its penalty term. It applies a stricter penalty for model complexity, particularly as the sample size increases, since the penalty grows with the logarithm of the number of observations. As a result, BIC often favors more parsimonious models compared to AIC, especially in large datasets. Moreover, BIC is consistent, meaning it tends to identify the true model as the sample size grows, whereas AIC may lean toward slightly overfitted models (Arevalo et al., 2017).

Both criteria offer distinct advantages, and their application depends on the trade-off between model

fit and simplicity, aligned with specific research objectives (Vrieze, 2012). This study employs both AIC and BIC for model selection to capitalize on their complementary strengths. AIC is utilized for its effectiveness in balancing goodness of fit and complexity, making it particularly useful for identifying models that closely approximate the true data-generating process, especially in predictive modeling. Meanwhile, BIC is incorporated for its explicit consideration of sample size and stronger complexity penalty, which favors simpler, more parsimonious models, particularly in large datasets. By integrating both criteria, the study ensures a comprehensive model evaluation, mitigating the risk of overfitting associated with AIC while benefiting from BIC's emphasis on simplicity and consistency (Agiakloglou & Tsimpanos, 2023).

4. Empirical application using China's rice trade data

4.1Data Description and Preprocessing

Building on the discussion of model selection criteria, this section applies AIC and BIC to determine the optimal number of components in the FMM using China's rice trade data. This empirical analysis not only demonstrates the practical application of FMM but also provides valuable insights into the trade patterns and influencing factors in China's rice market.

This study utilizes monthly data obtained from the agriculture trade section of the China Stock Market and Accounting Research (CSMAR) Database, covering the period from July 2005 to May 2024, resulting in a total of 227 observations. The dataset includes five core variables: Import_Quantity and Export_Quantity (measured in tons), Import_Price and Export_Price (measured in USD per ton), and Exchange_Rate (measured in monthly average RMB/USD).

Among these variables, price data were log-transformed to reduce skewness and better approximate normality, which is important for finite mixture modeling. The transformation helps stabilize variance and improve model convergence when dealing with potentially heavy-tailed economic data. The quantity and exchange rate variables were retained in their original scales.

Data inspection revealed no missing values across the dataset. To ensure robustness and mitigate the influence of extreme values, Import_Price and Export_Price were winsorized at the 1st and 99th percentiles. This preprocessing step helps reduce distortion caused by price spikes or outliers, especially during periods of market shocks.

Table 1. Descriptive Statistics of Key Variables

Variable	Mean	Std. Dev.	Min	Max	Obs
Import_Quantity	195237.79	159511.53	4166	720000	227

Import_Price	518.14	106.49	320.63	920.95	227
Export_Quantity	99773.53	81094.15	641	350675	227
Export_Price	634.47	307.29	277.67	2466.16	227
Exchange_Rate	6.79	0.52	6.1	8.23	227

Table 1 presents the descriptive statistics for the key variables. The average import quantity was approximately 195,238 tons, while the export quantity averaged 99,774 tons. Average import and export prices were 518.14 USD/ton and 634.47 USD/ton, respectively, with considerable standard deviations, indicating price volatility over the sample period. The average exchange rate was 6.71 RMB/USD, reflecting a relatively stable currency trend with mild fluctuation.

To assess the distributional characteristics of the price variables, their empirical distributions were examined and found to deviate from unimodal symmetry. Specifically, both import and export prices displayed signs of multimodality and asymmetry, with visible clustering around different value ranges across time. These characteristics suggest the presence of latent subgroups or regime-switching behavior in China's rice trade. Such distributional patterns support the application of FMM, which are particularly well-suited for identifying hidden heterogeneity and modeling structural shifts in economic data.

All data processing and empirical analysis were conducted using Stata 17. Reproducibility of the results is ensured, and the dataset, along with the analysis scripts, is available upon request to facilitate independent validation of the findings.

4.2 Model Selection Results

Selecting the appropriate number of components is a crucial step before applying the FMM, as an incorrect specification can result in unreliable estimates and misleading interpretations (Manole & Khalili, 2021). In this study, the optimal component count was determined for all chosen variables related to China's rice trade, using the AIC and BIC as key evaluation metrics. The analysis was carried out in Stata 17, with model selection based on the lowest AIC and BIC values, ensuring the best possible fit in line with theoretical expectations (Liu et al., 2023).

The component analysis for China's rice import and export prices, quantities, and exchange rates is summarized in Tables 1-4.. Identifying the optimal number of components is crucial, as it directly affects the distribution weights obtained. These tables display the computed values of -2 log-likelihood, AIC, and BIC for models with k=2,3 and 4. The choice of these component numbers reflects a balance between model complexity and interpretability. Higher-component models, such as those with 5 or more components, were excluded to prevent overfitting, which could reduce the model's generalizability (Fraley & Raftery, 2002). Furthermore, preliminary assessments indicated that increasing the number of components beyond four did not yield significant

improvements in model fit based on -2 log-likelihood, AIC, and BIC. Restricting the analysis to 2, 3, and 4 components ensured computational efficiency while preserving result robustness and relevance. In this study, models with these component numbers were assessed, as an excessive number of components could lead to overfitting (Wang & Yang, 2024). The final selection was based on AIC and BIC values, with the model yielding the lowest values for these criteria identified as the best fit.

Table 2 Number of components for Rice Import Price with Exchange Rate

Number of Component	Number of Parameter	-2 log likelihood	AIC	BIC
2	7	-1317.070	-1303.070	-1279.09535
3	10	-1312.488	-1292.488	-1258.23850
4	15	-1300.632	-1270.632	-1219.25775

For the rice import price and exchange rate analysis, the model with two components was identified as the best choice based on the BIC criterion. In a FMM, parameters define each component's distribution and proportions, including key statistical measures such as mean, variance, and other distribution-specific characteristics. As the number of components increases, so does the number of parameters, leading to higher model complexity (Frühwirth-Schnatter et al., 2021) .

The -2 log-likelihood measures how well the model aligns with the observed data, with lower values indicating a better fit. However, relying exclusively on this metric can lead to overfitting, as adding more components inevitably improves model fit (Bainter, 2015). Both the number of parameters and the -2 log-likelihood play crucial roles in model selection. While additional parameters enhance fit, they also introduce greater complexity. Meanwhile, the -2 log-likelihood contributes to the computation of information criteria, such as AIC and BIC. Ultimately, model selection prioritizes minimizing these criteria, ensuring a balance between goodness-of-fit and model complexity, rather than focusing solely on likelihood values or parameter count.

According to the results in Table 1, the analysis examined the optimal number of components for modeling the relationship between rice import price and exchange rate in China. The findings indicated that the two-component model was the most appropriate choice based on the BIC criterion, as it yielded the lowest value compared to models with three or four components. Consequently, k=2 was determined to be the most reasonable selection for analysis.

Table 3 Number of components for Rice Import Quantities with Exchange Rate

Number of Component	Number of Parameter	-2 log likelihood	AIC	BIC
2	7	-2995.783	-2981.783	-2957.80835
3	11	-2963.409	-2941.409	-2903.73455
4	12	-2955.648	-2931.648	-2890.54860

For rice import quantities and exchange rate, the model with two components was identified as the best choice based on the BIC criterion. As shown in Table 3, statistical analysis confirmed that the two-component model was the most appropriate for capturing the relationship between rice import quantities and exchange rate. This selection was justified by the lowest BIC value among all tested models, reflecting a balance between model fit and complexity. Similarly, the AIC value was also minimized when the number of components was set to two. Therefore, the results indicate that k=2 was the most suitable choice for this study.

Table 4 Number of components for Rice Export Price with Exchange Rate

Number of Component	Number of Parameter	-2 log likelihood	AIC	BIC
2	7	-1538.202	-1524.202	-1500.22735
3	11	-1485.012	-1463.012	-1425.33755
4	15	-1479.866	-1449.866	-1398.49175

For rice export price and exchange rate, the model with two components was identified as the most suitable based on the BIC criterion. As presented in Table 3, the two-component model was the best choice for analyzing this relationship, as it achieved the lowest BIC value among all tested models, ensuring an optimal balance between fit and complexity. Furthermore, the AIC value was also

minimized for this model, reinforcing its selection as the most appropriate option.

Table 5 Number of components for Rice Export Quantities with Exchange Rate

Number of Component	Number of Parameter	-2 log likelihood	AIC	BIC
2	7	-2817.375	-2803.375	-2779.40035
3	10	-2815.074	-2795.074	-2760.82450
4	15	-2806.129	-2776.129	-2724.75475

For rice export quantities and exchange rate, the two-component model was identified as the optimal choice based on the BIC criterion. Table 4.4 presents the component analysis results, where the two-component model achieved the lowest BIC value, indicating the best balance between simplicity and model fit. Moreover, its AIC value was also lower than those of models with three or four components, further supporting its selection as the most appropriate model.

4.3 Robustness Checks

To evaluate the robustness of the empirical findings, several sensitivity checks were conducted. First, the full sample was divided into two sub-periods: July 2005–December 2013 and January 2014–May 2024. The FMM was re-estimated separately for each sub-period. In both cases, the two-component structure remained the optimal model based on the Bayesian Information Criterion (BIC), confirming that the core finding of dual trade regimes is not sensitive to temporal variation.

Second, the model was re-estimated using alternative initialization seeds within the Expectation-Maximization (EM) algorithm. The results consistently converged to a two-component solution across different random starts, indicating stable model convergence and robustness to initial parameter settings.

Finally, a restricted variable set was tested by estimating the model using only import price and exchange rate. Even under this simplified specification, the two-component model remained optimal under both AIC and BIC criteria. These tests collectively support the stability of the component structure and confirm that the presence of latent regimes in China’s rice trade is not an artifact of model specification.

4.4 Interpretation of Latent Trade Regimes

In conclusion, this study employed a two-component mixture model, a choice well-suited for analyzing economic data such as rice trade and exchange rates, which often exhibit distinct patterns under different conditions, including stable and volatile periods (Dew-Becker et al., 2021). This approach aligns with the nonlinear and dynamic nature of real-world economic systems, where regime shifts frequently occur in response to both domestic policies and external shocks (Brock, 2018). Furthermore, Jochmans et al. (2017) highlighted the effectiveness of two-component mixture models in addressing tail behavior and extreme observations, making them particularly suitable for datasets that experience substantial fluctuations during periods of market turbulence, such as those observed in agricultural commodity markets.

To enhance the interpretation of the two-component structure, we examine the economic characteristics associated with each latent regime. The first component, which dominates during periods with relatively low price variability and stable exchange rates, likely corresponds to a "stable trade regime." This regime is characterized by government-controlled price mechanisms, predictable import quotas, and steady exchange rate policies. In contrast, the second component is marked by higher price volatility and irregular trade volumes, suggesting a "volatile regime" influenced by global economic shocks, policy transitions, and speculative pressures.

Although the model was not explicitly estimated with covariates, posterior probabilities from the mixture model suggest temporal clustering of each component. Periods such as 2008–2009 (global financial crisis), 2014–2016 (grain reserve system reforms), and 2020–2022 (COVID-19 disruptions) exhibit strong alignment with the volatile regime, as indicated by higher assignment probabilities to the second component. Conversely, the years 2010–2013 and 2017–2018 largely align with the stable regime. These empirical patterns confirm that the two-component structure is not only statistically valid but also economically meaningful, reflecting distinct underlying mechanisms in China's rice trade.

Building on this theoretical foundation, the two identified regimes in China's rice trade can be empirically linked to specific macroeconomic conditions and policy developments. The stable phase represents periods of relative economic equilibrium characterized by steady trade flows, moderate exchange rate variations, and the operation of long-standing domestic policies, including grain price supports and regulated import quotas. These conditions prevailed during years of balanced supply-demand dynamics, stable global market conditions, and predictable state reserve management.

In contrast, the volatile phase corresponds to episodes marked by heightened uncertainty and market disturbances, reflecting the sensitivity of rice trade to both global and domestic disruptions. For instance, the 2008 global financial crisis triggered sharp movements in exchange rates and commodity prices, while domestic reforms to China's grain reserve system between 2014 and 2016 introduced new uncertainties into stockpile levels and trade

flows. The COVID-19 pandemic (2020–2022) further disrupted international supply chains and trade logistics, exacerbating price and quantity volatility. Moreover, escalating trade tensions during the U.S.–China trade conflict period (2018–2019) generated additional external pressures that directly influenced trade patterns and exchange rate dynamics. These historical events jointly contributed to the empirical classification observed in the model, with the volatile regime capturing these periods of intensified external and internal shocks.

The ability of the FMM to distinguish between these underlying economic forces not only validates its appropriateness for modeling heterogeneous trade behavior but also offers valuable policy insights.

Table 6 Comparison of Stable and Volatile Trade Regimes

Feature	Stable Regime	Volatile Regime
Trade Characteristics	Predictable trade flows, low price volatility	Irregular volumes, high price fluctuations
Exchange Rate Trend	Stable or gradual variation	Sharp changes, highly responsive to external shocks
Policy Context	Price supports, import quota regulations	Policy shifts, grain reserve reforms, deregulation
Representative Periods	2010–2013, 2017–2018	2008–2009, 2014–2016, 2020–2022, 2018–2019
Key Influences	Internal policy control, demand-supply equilibrium	Global crisis, pandemic, trade conflicts, speculative market forces

This summary table provides a comparative overview of the two regimes identified through the FMM analysis. It highlights the contrasting economic conditions, policy settings, and external influences that shape each regime, reinforcing the empirical validity and interpretive value of the two-component model.

Recognizing when the system transitions into a volatile regime enables policymakers to adopt timely interventions, such as reserve adjustments, exchange rate stabilization measures, or trade facilitation policies aimed at mitigating the impact of global uncertainty on China’s rice market.

5.Conclusion

This study investigates the problem of determining the optimal number of components in FMM, combining both theoretical analysis and empirical application to China's rice trade under exchange rate fluctuations. The research underscores the importance of correctly specifying mixture components to enhance model interpretability, capture latent trade structures, and improve predictive accuracy.

From the theoretical perspective, we reviewed the foundational concepts of FMM and critically examined various model selection techniques. Among these, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were prioritized in our empirical application, given their wide adoption and sound theoretical underpinnings. AIC, while tending to favor more complex models, balances goodness of fit with complexity, whereas BIC provides a more conservative, parsimonious approach, especially valuable in finite samples. These criteria jointly offer a practical balance between model simplicity and representational accuracy.

Empirically, the study applies FMM to monthly data on China's rice imports and exports, incorporating trade prices, volumes, and exchange rate movements. The results demonstrate that mixture modeling effectively captures unobserved heterogeneity in trade behaviors, uncovering distinct market dynamics that would likely remain masked under conventional linear regression frameworks. Across all tested models, a two-component structure consistently emerged as optimal, as evidenced by the lowest BIC values. This suggests the presence of two trade regimes: a stable regime characterized by predictable trade flows and moderate exchange rate variations, and a volatile regime influenced by global financial shocks, domestic policy adjustments, and broader macroeconomic disruptions. Posterior classification further allowed us to temporally map these regimes to historical events such as the 2008 global financial crisis, grain reserve reforms, the COVID-19 pandemic, and international trade tensions.

Beyond theoretical and empirical contributions, this study carries important policy implications. The ability to identify latent trade regimes offers practical tools for policymakers to monitor exchange rate sensitivities, design more adaptive trade policies, and develop contingency frameworks that enhance trade resilience. In particular, the interaction between state reserve operations and regime shifts suggests that well-timed reserve interventions may help stabilize domestic markets during periods of heightened external risk. Furthermore, recognizing the existence of such trade heterogeneity can inform international trade negotiations, allowing for more flexible and responsive agreements under evolving global trade conditions.

Nonetheless, several limitations remain, which provide fertile ground for future research. Although AIC and BIC served effectively in this study, alternative model selection methods such as the Integrated Completed Likelihood (ICL), Bootstrap Likelihood Ratio Tests (BLRT), and cross-validation techniques could offer additional robustness checks under different data structures. Moreover, future work could explore Bayesian mixture modeling, where prior information may strengthen estimation precision and classification stability, particularly when component overlap is

substantial. Extending the framework to dynamic mixture models would further enable analysis of regime probabilities as evolving processes, capturing shifts in trade patterns over time. Finally, applying similar methodologies to other agricultural sectors or multi-country trade datasets would enhance the external validity and generalizability of the findings, broadening the relevance of FMM in both international trade and broader economic modeling.

References

econometric models. *Journal of Spatial Econometrics*, *4*(1), 2.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Arevalo, F., Cid, A., & Moya, J. (2017). AIC and BIC for cosmological interacting scenarios. *The European Physical Journal C*, *77*, 1-13.

Bainter, S. A. (2015). Advantages of a Bayesian approach for examining class structure in finite mixture models. *Multivariate behavioral research*, *50*(1), 127-127.

Biernacki, C., Celeux, G., & Govaert, G. (2002). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(7), 719-725.

Brock, W. A. (2018). Nonlinearity and complex dynamics in economics and finance. In *The economy as an evolving complex system* (pp. 77-97). CRC Press.

Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. *Philosophy of statistics*, 583-605.

Dew-Becker, I., Giglio, S., & Kelly, B. (2021). Hedging macroeconomic and financial uncertainty and volatility. *Journal of Financial Economics*, *142*(1), 23-45.

Fei, X., Lou, Z., Lv, X., Ren, Z., & Xiao, R. (2023). Pollution threshold assessment and risk area delineation of heavy metals in soils through the finite mixture distribution model and Bayesian maximum entropy theory. *Journal of Hazardous Materials*, *452*, 131231.

Feng, Z. D., & McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(3), 609-617.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the american statistical association*, *97*(458), 611-631.

Frühwirth-Schnatter, S., Malsiner-Walli, G., & Grün, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, *16*(4), 1279-1307.

Haji Aghabozorgi, H., & Eskandari, F. (2024). Clustering and estimation of finite mixture models under bivariate ranked set sampling with application to a breast cancer study. *Statistical Papers*, *65*(2), 705-736.

Jenkins, S. P., & Rios-Avila, F. (2023). Finite mixture models for linked survey and

administrative data: Estimation and postestimation. *The Stata Journal*, *23*(1), 53-85.

Jochmans, K., Henry, M., & Salanié, B. (2017). Inference on two-component mixtures under tail restrictions. *Econometric Theory*, *33*(3), 610-635.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research*, *33*(2), 188-229.

Liu, Q., Charleston, M. A., Richards, S. A., & Holland, B. R. (2023). Performance of akaike information criterion and Bayesian information criterion in selecting partition models and mixture models. *Systematic Biology*, *72*(1), 92-105.

Manole, T., & Khalili, A. (2021). Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure. *The annals of statistics*, *49*(6), 3043-3069.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 38). M. Dekker New York.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models* (Vol. 299). John Wiley & Sons.

Mortazavi, R., & Lundberg, M. (2020). Expenditure-based segmentation of tourists taking into account unobserved heterogeneity: The case of Venice. *Tourism Economics*, *26*(3), 475-499.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American journal of Mathematics*, 343-366.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, *185*, 71-110.

Sawadogo, R., Sawadogo, P. N., & Kinda, R. S. (2024). Financial Inclusion, Globalization and Structural Transformation in Developing Countries. *Journal of Economic Integration*, *39*(2), 420-453. https://doi.org/10.11130/jei.2024013

Schlattmann, P. (2003). Estimating the number of components in a finite mixture model: the special case of homogeneity. *Computational statistics & data analysis*, *41*(3-4), 441-451.

Schneider, M. P., & Scharfenaker, E. (2020). Mixing it up: The case for finite mixture models to study the distribution of income. *The European Physical Journal Special Topics*, *229*(9), 1685-1704. https://doi.org/10.1140/epjst/e2020-900130-4

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and computing*, *10*(1), 63-72.

Suk, Y., Kim, J.-S., & Kang, H. (2021). Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes. *Journal of Educational and Behavioral Statistics*, *46*(3), 323-347.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the

differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, *17*(2), 228.

Wang, C., & Yang, Y. (2024). Estimating the Number of Components in Finite Mixture Models via Variational Approximation. *arXiv preprint arXiv:2404.16746*.

Manuscript received: June 8, 2025

Revised: June 23, 2025

Accepted: July 2, 2025

Author Biographies

CHEN TING, master in Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia. Email: chenting0418@gmail.com

Phoong Seuk Yen, Associate Professor in Department of Mathematics, Faculty of Science and Mathematics,Universiti Pendidikan Sultan Idris. Email: phoong@fsmt.upsi.edu.my