Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 13
AMMSAC 2026

A Multi-source Evaluation Fusion Model Based on MCMC Reverse Inference and DVSE Dynamic Variance Standardization

Yiqing Hu *, Bowen Zhou and Jiangyuxuan Dai

Xi'an Jiaotong University, Xi`an, China

* Corresponding Author Email: yiqinghu@stu.xjtu.edu.cn

Abstract. To address the issues of inconsistency between public scores and implicit voting scales, as well as significant fluctuations in the generation of competition results, this paper investigates stable ranking methods based on the fusion of multi-source evaluation data. First, historical competition rules are transformed into computable constraints, and the distribution of undisclosed votes is estimated using a Bayesian framework and MCMC sampling. Subsequently, consistency tests and mechanism backtesting are employed to compare the stability of cumulative ranking, proportional weighting, and composite rules. The results show that the model achieves an overall consistency rate of 66.3% with historical results, with the proportion-weighted rule accounting for 73.5% of the inconsistent samples; the proposed Dynamic Variance Standardization Equilibrium (DVSE) mechanism reduces the probability of extreme controversial results by 12.4%. The study demonstrates that DVSE unifies scoring and voting scales via Z-scores, suppressing the excessive amplification of high-variance data while preserving multi-source information, thereby providing a transferable algorithmic framework for similar competition evaluation and comprehensive ranking tasks.

Keywords: MCMC Reverse Inference; Dynamic Variance Standardization Equilibrium (DVSE); in Bayesian Inference.

1. Introduction

In multi-source evaluation and ranking tasks, different data dimensions often have varying scales, volatility, and information density. If a weighted sum or simple ranking fusion is applied directly, the final results are prone to being influenced by highly volatile dimensions, leading to reduced ranking stability. Therefore, how to suppress abnormal fluctuations while preserving multi-source information is a critical issue in the design of evaluation algorithms.

Existing methods typically fall into two categories: rank aggregation and ratio-weighted methods. Rank aggregation can mitigate the impact of outliers but compresses original differences, resulting in the loss of detailed information; proportional weighting preserves continuous variations but may amplify high-variance data sources, causing results to deviate from the comprehensive evaluation objectives. Furthermore, in practical applications, some key variables are often not directly observable, making the modeling and validation of evaluation mechanisms more challenging.

To address these shortcomings, this paper constructs an algorithmic model framework for heterogeneous evaluation data. First, under known output results and rule constraints, we employ a reverse inference method to recover unobservable information; second, we use mechanism backtesting to compare the stability of different aggregation rules; finally, we propose the Dynamic Variance Standardization Equilibrium (DVSE) mechanism, which maps different evaluation dimensions to a unified standardized space before performing weighted fusion, thereby reducing the interference of scale and variance differences on the results.

Taking the process by which professional scores and audience votes jointly determine rankings in competition-style programs as an example, this paper abstracts it as a multi-source ranking problem. This framework not only enhances the interpretability of hidden information but also improves the stability and applicability of complex evaluation systems under extreme input conditions.

2. MCMC-based Reverse Inference Model for Latent Voting

In the overall evaluation system of competitive reality shows, judges' scores mainly reflect the completion of dance movements, technical details, and stage performance, while fan voting more often reflects contestants' popularity base, public attention, and market appeal. These two types of information jointly determine the elimination result, but they are not equally observable. Judges' scores are usually public, while the exact number of fan votes is not fully disclosed, which makes the voting side a typical "black box".

Therefore, this section treats the weekly elimination result as the known output, regards judges' scores as partially observable inputs, and infers the latent fan voting in reverse within the constraint space defined by the competition rules. In other words, the hidden voting distribution needs to be recovered based on the public scores and the final elimination list. From a computational perspective, this can be understood as a nonlinear parameter estimation problem in a high-dimensional constraint space. The difficulty is that the same elimination result can often be produced by many different voting combinations, so there is no unique solution. For this reason, the model should not only provide a point estimate, but also describe the uncertainty of the estimation results. Based on this, this paper introduces a probabilistic inference method to estimate the latent fan voting distribution and to provide a data basis for the later mechanism backtesting.

2.1. Mathematical Constraint Representation of Competition Rules

To carry out reverse estimation, the elimination rules in the show first need to be converted into computable mathematical constraints. According to the scoring methods used in different seasons, this paper mainly discusses two types of aggregation logic: the rank-sum mechanism and the percentage-weighted mechanism. The former emphasizes relative ranking, while the latter directly keeps the continuous differences in scores and voting proportions. These two mechanisms are not equally sensitive to extreme popularity and technical score gaps.

2.1.1. Discrete Constraint Modeling of the Rank-Sum Mechanism.

In seasons 1–2 and 28–34, the show used a discrete rank-sum method. Let the set of competing couples in week w be $C = c_1, c_2, \ldots, c_n$. For contestant i, the judges' score ranking is denoted as $R_J(i,w) \in 1,2,\ldots,n$, and the unknown fan voting ranking is denoted as $R_V(i,w) \in 1,2,\ldots,n$. The combined ranking score is defined as:

$$S(i,w) = R_J(i,w) + R_V(i,w) \tag{1}$$

Under this mechanism, a larger value means a worse overall position. Therefore, if contestant E is eliminated, the combined ranking score should fall into the lowest competitive range among all contestants, which can usually be written as:

$$\exists \left\{R_V(i,w)\right\}_{i=1}^{n}, \quad \text{s.t.} \begin{cases} S(E,w) \geq S(i,w), \forall i \in C \\ \sum_{i=1}^{n} R_V(i,w) = \dfrac{n(n+1)}{2} \\ R_V(i,w) \neq R_V(j,w), \forall i \neq j \end{cases} \tag{2}$$

This constraint shows a key logic under the ranking system: if a contestant ranks high on the judges' side but is still eliminated, then this contestant's ranking in fan voting must be clearly low. In other words, although the rank-sum mechanism compresses the original score gaps, the elimination result can still be used to infer the possible range of insufficient fan support.

2.1.2. Continuous Constraint Modeling of the Percentage-Weighted Mechanism.

Starting from season 3, the show's mechanism gradually shifted to a percentage-weighted method. Let $J_{i,w}$ denote the total judges' score of contestant i in week w. Then the contestant's

contribution proportion from the judges' side is: $P_J(i,w) = J_{i,w} \Big/ \sum_{k=1}^{n} J_{k,w}$. Let $v_{i,w}$ denote the fan voting proportion of contestant i, which satisfies the simplex constraint: $\sum_{i=1}^{n} v_{i,w} = 1$. Under the percentage mechanism, the total proportion score of contestant i can be written as:

$$P_{\text{total}}(i,w) = \frac{1}{2} P_J(i,w) + \frac{1}{2} v_{i,w} \tag{3}$$

If contestant E is eliminated, then this contestant's total proportion score needs to be lower than that of the other competitors, namely:

$$P_J(E,w) + v_{E,w} < P_J(i,w) + v_{i,w}, \quad \forall i \neq E \tag{4}$$

After further rearrangement, the linear constraint on the fan voting proportion is obtained as:

$$v_{E,w} - v_{i,w} < P_J(i,w) - P_J(E,w), \quad \forall i \neq E \tag{5}$$

Unlike the rank-sum mechanism, the percentage mechanism keeps the continuous differences in voting proportions and score proportions. Therefore, fluctuations on the fan voting side are directly passed on to the final result. When the voting distribution becomes extremely skewed, the influence of public support may increase quickly. This is also an important reason why mechanism backtesting and fairness analysis are needed later.

2.2. Probabilistic Inference Method Based on MCMC

The constraints above define the feasible voting space, but they cannot directly determine the unique real vote counts. For this non-unique solution problem, this paper uses the Markov Chain Monte Carlo (MCMC) method to conduct large-scale sampling within the feasible space that satisfies the elimination constraints [1][2]. The result is not a single fixed value, but a set of possible voting distributions. Based on these samples, the statistical features of latent fan voting can be approximately described.

2.2.1. Prior Distribution and Likelihood Function.

This paper builds a probabilistic model of voting proportions under the Bayesian framework [3]. Let the fan voting proportion vector in week w be v_w, and assume that it follows an unbiased Dirichlet prior distribution: $v_w \sim \text{Dir}(\alpha)$, where $\alpha_1 = \cdots = \alpha_n = 1$. This setting means that, when there is no extra information, no contestant is favored in advance. Based on the known elimination result, the likelihood function is defined as an indicator function, which is used to judge whether a given voting proportion satisfies the elimination constraint:

$$L(v_w \mid \text{Result}) = I\left(\bigcap_{i \neq E} P_{\text{total}}(E,w) < P_{\text{total}}(i,w) \right) \tag{6}$$

When the candidate voting vector can explain the actual elimination result, the likelihood value is 1; otherwise, it is 0. This treatment is quite direct, but it is also effective, because the elimination result itself provides constraint information rather than a complete observation of vote counts.

2.2.2. Metropolis-Hastings Sampling Process.

In the t-th iteration, the algorithm starts from the current state $v^{(t)}$ and generates a slightly perturbed candidate vector v^* [4]. If v^* satisfies all elimination constraints, the candidate state is accepted; if not, the original state is kept. After the burn-in stage, the retained valid samples are used to estimate the fan voting proportion of contestant i in week w:

$$\hat{v}_{i,w} = \frac{1}{M-B}\sum_{t=B+1}^{M} v_i^{(t)} \tag{7}$$

where $M = 10^6$ denotes the total number of sampling iterations, and B denotes the number of burn-in samples. In this way, a relatively stable posterior mean can be obtained in the non-unique solution space, and the input data for later mechanism backtesting can be provided.

2.3. Estimation Confidence and Consistency Test

The posterior mean alone is not enough to show whether the inference result is reliable. Since the elimination result provides incomplete information, the strength of the constraints also varies across weeks. Therefore, after the estimated voting proportions are obtained, the stability of these estimates still needs to be evaluated, and whether they can explain the actual elimination results also needs to be tested.

2.3.1. Confidence Measure Based on Information Entropy.

This paper introduces information entropy to describe how concentrated the feasible solution space is. For the estimated voting proportions in week w, information entropy and the estimation confidence can be defined as:

$$H(w) = -\sum_{i=1}^{n} \hat{v}_{i,w} \log \hat{v}_{i,w}, \quad \gamma_w = 1 - \frac{H(w)}{\log(n)} \tag{8}$$

When γ_w is low, it means that the feasible voting combinations are relatively scattered, and the elimination result places weak constraints on the real voting structure. When γ_w is high, it means that the posterior distribution is more concentrated, so the inference result is also more reliable. In general, as the competition moves into the later stage, the number of remaining contestants decreases, and the feasible solution space becomes compressed. As a result, the certainty of the estimation usually improves.

2.3.2. Model Consistency Test.

To test whether the reverse inference results can reasonably explain the historical elimination results, this paper substitutes the estimated $\hat{v}_{i,w}$ back into the scoring rules of the corresponding season and calculates the consistency between the model-predicted elimination result and the actual elimination result. Let W denote the total number of weeks. The consistency indicator is defined as:

$$C = \frac{1}{W}\sum_{w=1}^{W} \delta(\text{Model}_w, \text{Actual}_w) \tag{9}$$

where $\delta(\text{Model}_w, \text{Actual}_w)$ is an indicator function. It takes the value 1 when the model prediction is consistent with the actual result, and 0 otherwise. This indicator can be used to judge whether the inferred voting distribution has enough explanatory power. Although it cannot prove that the estimated vote counts are the only true values, it can show whether the distribution is compatible with the historical results at the rule level.

2.4. Analysis of Reverse Inference Results

After the sampling was completed, this paper reconstructed the changing trend of latent fan voting across different seasons. Since the original fan vote counts are not public, the results here should be understood more as probabilistic estimates obtained under rule constraints, rather than exact statistical values. This paper keeps Season 32 as a representative case to show the dynamic changes in fan voting as the competition progresses.

Fig. 1 shows the estimated voting trajectories of Season 32 finalists and their 95% confidence intervals. Overall, after the competition entered the second half, the vote counts of the main contestants showed an upward trend and increased more clearly around week 8. This suggests that as the competition approached the final, audience participation became stronger, and the influence of fan voting on the final ranking also increased.

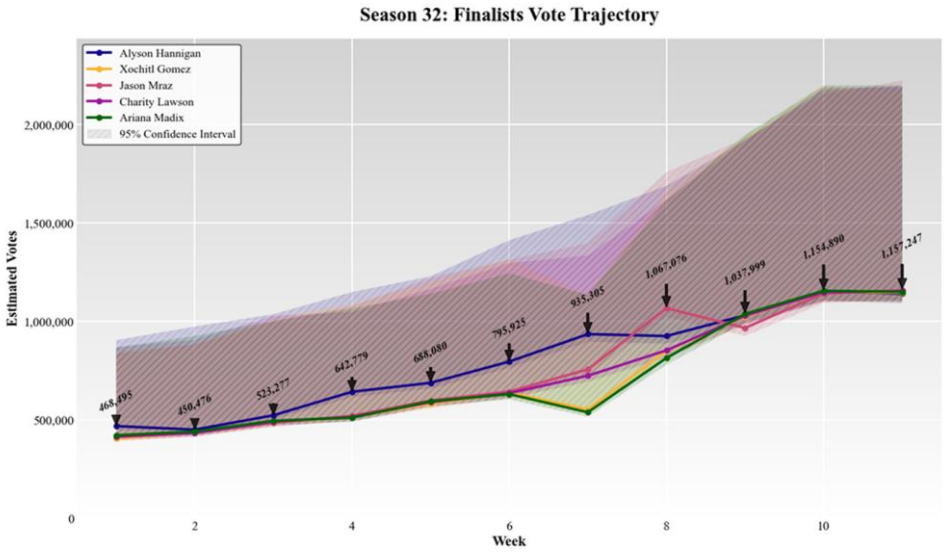

Fig. 1 Estimated trajectory of fan votes and 95% confidence intervals for season 32 finalists

From the confidence intervals, the estimation error becomes much narrower in the later stage. One important reason is that the number of remaining contestants decreases, and the number of voting combinations that can explain the elimination result also decreases, making the model constraints stronger. Although the voting gaps among different contestants widened in the middle stage, in the final stage, several voting trajectories gradually converged into a similar range, at about 1.15 million votes. This result shows that late-stage competition is not determined purely by technical scores. Fan voting forms a clear "protective layer". It is this protective layer that allows some contestants who do not have an advantage in technical scores to remain competitive under the mechanism.

3. Backtesting and Fairness Analysis of Voting Aggregation Mechanisms

The program rules were adjusted several times across different seasons. Behind these changes, the same issue was being handled: how much constraint professional scoring should keep, and how much influence space audience voting should have. The previous section obtained the estimated fan voting vector V through reverse inference. Based on this, this section further backtests the historical mechanisms and compares the stability of the ranking mechanism, the percentage mechanism, and the composite mechanism under different competitive situations [5].

To compare different rules in a unified way, the input vector for week w is first constructed. Let the judges' score vector of all contestants be $J_w = [J_{1,w}, \ldots, J_{n,w}]$, and let the estimated fan voting vector be $V_w = [V_{1,w}, \ldots, V_{n,w}]$. Different mechanisms can be viewed as different mappings of these two input vectors. The same scoring and voting data may correspond to different elimination results under different aggregation functions, and this is the core of mechanism backtesting.

3.1. Rank-Sum Mechanism

The ranking mechanism is essentially a non-parametric discrete mapping. It does not directly use the absolute differences in original scores or vote counts, but only keeps the relative order among contestants. For contestant i, the combined score under the ranking mechanism can be written as:

$$S_{\text{Rank}}(i,w) = R(J_{i,w}, J_w) + R(V_{i,w}, V_w) \tag{10}$$

where $R(\cdot)$ denotes the descending ranking operator. The ranking mechanism compresses differences in the continuous space. Even if one contestant's fan votes are 10 times those of the second-place contestant, under the ranking mechanism this only appears as an advantage of one ranking position.

From the perspective of information processing, this mechanism is equivalent to truncating extreme values. It loses part of the detailed information, but for this reason, extreme popularity is also less likely to directly break through technical scores and form an overwhelming advantage. Of course, when the score gaps among contestants are very small, the ranking mechanism may also produce jump-like changes. Still, overall, it has a certain restraining effect on high-variance voting shocks.

3.2. Percentage-Weighted Mechanism

Unlike the ranking mechanism, the percentage-weighted mechanism keeps the continuous differences in scores and votes. Its combined score is based on proportional normalization and can be written as:

$$P_{\text{Total}}(i,w) = \frac{J_{i,w}}{\sum_{k=1}^{n} J_{k,w}} + \frac{V_{i,w}}{\sum_{k=1}^{n} V_{k,w}} \tag{11}$$

The advantage of this mechanism is that it keeps more information, and every change in votes can be reflected in the final score. But the problem also lies here. Judges' scores are usually concentrated within a relatively narrow range, while the distribution of fan votes fluctuates much more strongly. Therefore, when fan voting becomes highly skewed, the percentage mechanism directly amplifies this fluctuation into the final ranking.

From the perspective of variance, if the judges' proportion is denoted as P_J and the fan voting proportion is denoted as P_V, then the fluctuation of the combined result approximately satisfies:

$$\text{Var}(P_{\text{Total}}) \approx \text{Var}(P_J) + \text{Var}(P_V) \tag{12}$$

Since $\text{Var}(P_V)$ is often much higher than $\text{Var}(P_J)$, the final result is more likely to be dominated by the voting side. The original variance analysis shows that the coefficient of variation of the judges' proportion P_J usually falls between 0.05 and 0.15, while the degree of variation in the fan voting proportion P_V is clearly higher. In other words, when a contestant with extreme popularity appears, the percentage mechanism is more likely to weaken the constraining role of technical scores. This phenomenon is also a part that needs to be examined closely in the later mechanism backtesting.

3.3. Two-Stage Screening Model of the Judges' Save Mechanism

Starting from season 28, the show introduced the judges' save mechanism. Under this rule, the two couples at the bottom of the combined ranking are no longer eliminated directly. Instead, the judges decide who finally leaves. In the model, this is equivalent to adding one layer of conditional screening after the original aggregation rule. That is, the final elimination probability is no longer determined only by the combined score S_i, but is also affected by the judges' second-round judgment.

Let B_w denote the bottom two set determined by the combined score in week w. Then the elimination probability of contestant i can be written in a piecewise form:

$$P\left(\text{Elim}_i \mid S_i\right)=\begin{cases}0 & \text{if } i \notin B_w \\ \sigma\left(\lambda \cdot \left(J_k - J_i\right)+\eta\right) & \text{if } i \in B_w\end{cases} \tag{13}$$

where $\sigma(\cdot)$ denotes a nonlinear mapping function, and J_i and J_k denote the judges' scores of the related contestants in the bottom two. The core effect of this mechanism is a second truncation: only contestants who enter the bottom two will move into the judges' save stage, and contestants with weaker judges' scores may lose protection at this stage even if they have relatively high popularity. Therefore, it partly corrects the extreme results caused by pure voting aggregation, but it also introduces subjectivity from human judgment.

3.4. Definition of Flip Rate

To measure how much different mechanisms deviate from the professional evaluation order, this paper introduces Flip Rate as the core indicator. When the final elimination result is inconsistent with the judges' technical ranking, it is recorded as a flip event. Let W denote the total number of weeks included in the backtesting. Then Flip Rate is defined as:

$$\Phi = \frac{1}{W}\sum_{w=1}^{W} I(E_{\text{flip},w}) \tag{14}$$

where $I(E_{\text{flip},w})$ is an indicator function. If a flip event occurs in week w, it takes the value 1; otherwise, it takes the value 0. This indicator is not simply the same as an "error rate". Instead, it is used to measure the extent to which the voting mechanism changes the ranking result given by technical scoring. This definition provides a common standard for later comparisons of the ranking mechanism, the percentage mechanism, and the composite mechanism.

3.5. Backtesting Results and Fairness Analysis

To evaluate the actual performance of different aggregation mechanisms, this section combines qualitative case identification with quantitative backtesting. First, representative controversial cases are selected from 34 seasons and classified according to their mechanism-level patterns. The content of the original Table 1 is kept here in table form, so that the types of historical abnormal cases can be shown more clearly.

Table 1. Summary of representative controversial cases in DWTS history

Controversy Type	Representative Cases	Typical Pattern	Mechanism-level Explanation
Popularity-dominated advancement	Kelly Monaco, Iman Shumpert, Bobby Bones	Contestants with weaker technical rankings advanced or won because of strong audience support	High-variance fan voting weakened the constraint imposed by expert scoring
Premature elimination of high-scoring contestants	Mark Dacascos, Gilles Marini, Sabrina Bryan	Contestants with strong judges' scores were eliminated because of insufficient voting support	Technical superiority failed to offset uncertainty in the voting dimension
Rank-protected anomaly	Jerry Rice, Sean Spicer	Contestants with weak technical performance survived under rank-based or discretized rules	Rank discretization concealed the magnitude of score differences
Extreme vote-shock anomaly	Juan Pablo Di Pace, Apollo Anton Ohno	Strong weekly performance did not translate into stable ranking outcomes	Percentage-based aggregation was overly sensitive to vote fluctuations

As shown in Table 1, controversial results were not caused by a single reason. Some cases came from extreme popularity covering technical scores, while others happened because high-scoring contestants were eliminated early due to insufficient voting support. There is also a more special type of case, where the ranking mechanism itself compressed score gaps and gave contestants with weaker technical performance room to stay in the competition. After these cases are classified, the later mechanism backtesting is no longer just a comparison of prediction accuracy. It can further examine under what situations different rules are more likely to fail.

Next, this paper conducts global backtesting of historical mechanisms to compare how different aggregation rules respond to the abnormal situations above. Fig. 2 shows the comparison between the actual results and the predicted results under different rules.

The backtesting results show that the overall prediction consistency rate is 66.3%, corresponding to 222 cases, while the inconsistency rate is 33.7%, corresponding to 113 cases. These inconsistent results reflect a systematic flip risk. After further breakdown, the percentage mechanism contributes 73.5% of the inconsistent cases, with 83 cases in total. This shows that under the percentage mechanism, the high variance of fan voting is more likely to break through the constraint of technical scoring, giving extra advantage to some contestants with very high popularity but weaker technical rankings.

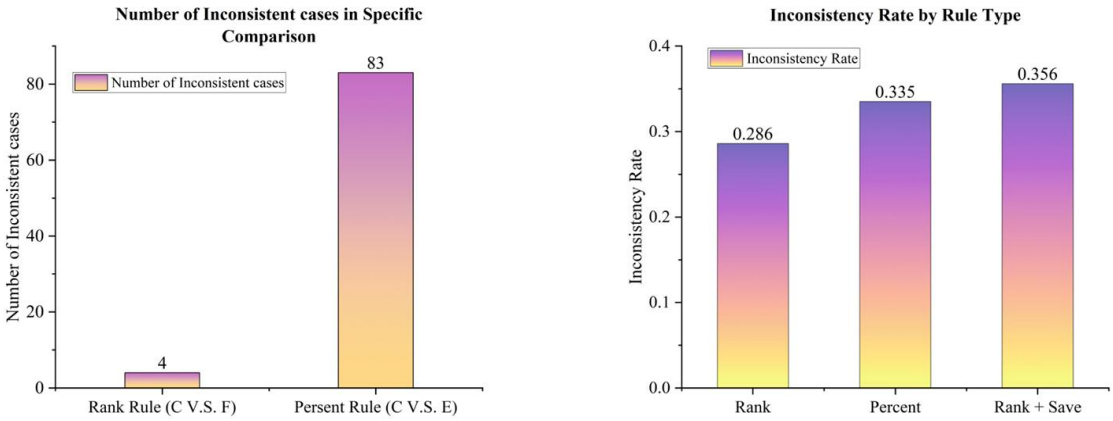

Fig. 2 Mechanism back-testing: comparison of actual vs. predicted outcomes under different rules

By contrast, the inconsistency rate of the ranking mechanism is 28.6%, showing stronger stability. The reason is that the ranking mechanism discretizes continuous scores, thereby truncating the amplification effect caused by extreme voting. This treatment loses part of the score-gap information, but it can also better protect contestants with higher technical scores. The composite mechanism, namely "ranking mechanism + judges' save", has an inconsistency rate of 35.6%. It corrects some extreme elimination results, but the judges' second intervention also brings in new subjective factors. Therefore, this mechanism is not simply better than the first two. Instead, it provides a compromise between technical fairness and audience participation.

4. Construction of the Dynamic Variance Standardization Equilibrium Mechanism

The mechanism backtesting above shows a fairly clear problem in traditional voting rules: the variances of different evaluation dimensions are not consistent. The ranking mechanism compresses original differences through discretization, so an advantage of 1 point and an advantage of 10 points may both appear as only one ranking unit. This causes information loss. The percentage mechanism keeps continuous differences, but it also directly passes the fluctuation of high-variance fan voting into the final result. Once extreme popularity appears, the constraining effect of technical scoring can easily be weakened.

To address the variance-dominance problem caused by this dimensional mismatch, this paper constructs the Dynamic Variance Standardization Equilibrium mechanism, namely DVSE. This

mechanism no longer directly adds raw scores or raw voting proportions. Instead, it first maps technical scores and fan voting into a unified Z-score space. The purpose is clear: being ahead by one standard deviation on the judges' side should have the same statistical meaning as being ahead by one standard deviation on the fan side. After variance alignment, the model can form a more stable comparison scale between technical performance and public support.

4.1. DVSE Modeling Based on Z-score

The core of DVSE is standardization [6]. Since judges' scores are usually concentrated within a smaller range, while fan voting may show fluctuations at a much larger scale, directly adding them together would make the influence of the two dimensions asymmetric. Therefore, the scoring function no longer focuses on the raw values themselves, but on each contestant's relative position within the competitive group of that week.

Suppose there are n couples competing in week t. Let $J_{i,t}$ denote the total judges' score of contestant i, and let $\hat{V}_{i,t}$ denote the estimated fan votes obtained through reverse inference. First, the mean and standard deviation of the judges' score dimension and the fan voting dimension in that week are calculated separately:

$$\mu_{J,t} = \frac{1}{n}\sum_{i=1}^{n} J_{i,t}, \quad \sigma_{J,t} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(J_{i,t} - \mu_{J,t})^2}, \quad \mu_{V,t} = \frac{1}{n}\sum_{i=1}^{n}\hat{V}i,t, \quad \sigma V,t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{V}i,t - \mu V,t)^2} \tag{15}$$

Based on this, the raw data are mapped into standardized scores:

$$Z_{i,t}^{(J)} = \frac{J_{i,t} - \mu_{J,t}}{\sigma_{J,t}}, \quad Z_{i,t}^{(V)} = \frac{\hat{V}_{i,t} - \mu_{V,t}}{\sigma_{V,t}} \tag{16}$$

After this transformation, million-level fluctuations in fan votes and single-digit gaps in judges' scores are compared on the same statistical scale. For example, if one contestant leads by 2σ in judges' scores, while another contestant leads by 2σ in fan voting, the two contestants have the same standardized advantage in DVSE. This is the key point that makes this mechanism different from the percentage mechanism.

4.2. Weighted Aggregation and the Implicit Penalty Mechanism

After the standardized scores are obtained, DVSE uses a weighted function to calculate the combined evaluation value:

$$S_{\text{DVSE}}(i,t) = w \cdot Z_{i,t}^{(J)} + (1-w) \cdot Z_{i,t}^{(V)} \tag{17}$$

where w is the weight of technical scoring, and $1-w$ is the weight of fan voting. When $w = 0.5$, the model treats judges' scores and fan voting as two equally important evaluation dimensions. This setting is not a simple average of raw scores. Instead, it averages them in a space where variances have already been aligned, so the result is more stable.

This mechanism also contains an implicit penalty effect. If a contestant performs weakly in the technical dimension, for example when $Z_{i,t}^{(J)}$ is clearly negative, then even if the fan voting score $Z_{i,t}^{(V)}$ is high, a large enough standardized advantage is still needed to offset the disadvantage on the technical side. Conversely, contestants with high technical scores but insufficient public support will not be fully protected without conditions. After this treatment, an extreme advantage in a single dimension is no longer likely to directly dominate the final result, making the mechanism more suitable for handling the conflict between technical fairness and public participation.

4.3. Backtesting Validation under Extreme Scenarios

To test the performance of the DVSE mechanism in controversial situations, this paper uses the historical voting data obtained from the reverse inference above to backtest several typical cases. The focus here is not simply to reproduce the original ranking, but to observe whether the relative relationship between technical scores and fan support becomes more stable after standardization.

In the Season 27 final, Bobby Bones surpassed Milo Manheim, who had stronger technical performance, through higher fan support and eventually won the championship. Under the logic of the percentage mechanism, Bobby Bones' high popularity would be directly converted into a score advantage, thereby covering his disadvantage on the technical side. However, under the DVSE framework, both technical scores and voting scores are first converted into standard deviation units. Milo Manheim's technical performance was at the high end of that week's distribution, about $Z^{(J)} \approx 2.15$, while Bobby Bones' technical score was clearly below the average level, about $Z^{(J)} \approx -1.85$. Therefore, Bobby Bones' popularity advantage is no longer directly amplified by the scale of raw vote counts, but is instead compared within the standardized space. In this way, the clear gap on the technical side re-enters the final evaluation.

The case of Jerry Rice reflects another type of problem. In Season 2, Jerry Rice stayed in the lower range of judges' scores for a long time, but still advanced into later rounds under the discretized protection of the ranking mechanism. The DVSE backtesting shows that after technical scores are mapped into the Z-score space, his technical disadvantage is clearly amplified into survival pressure in a statistical sense. If the technical dimension falls behind by more than 2.5σ, the advantage on the fan voting side becomes difficult to fully offset. This result suggests that DVSE can make up for the ranking mechanism's insensitivity to real score gaps.

Table 2 further compares several representative abnormal cases. Unlike the original explanatory figure, this part is organized as a mechanism comparison table, which makes it easier to show the differences in results under different rules.

Table 2. Comparative Evaluation of Historical Cases under Different Aggregation Mechanisms

Contestant	Historical Rule Outcome	Rank-based System	Percentage-based System	DVSE System	Main Observation
Bobby Bones	Winner	Supported	Strongly Supported	Penalized	Extreme popularity dominance is reduced
Juan Pablo Di Pace	High-ranked	Supported	Moderately Supported	Supported	Technical excellence remains competitive
Sean Spicer	Survived multiple rounds	Supported	Supported	Supported	Outcome remains stable across mechanisms
Jerry Rice	Advanced despite weak scores	Supported	Partially Supported	Not Supported	Technical baseline is restored
Sabrina Bryan	Early elimination	Not Supported	Not Supported	Supported	High technical quality receives additional protection

As shown in Table 2, DVSE does not simply raise the rankings of contestants with high technical scores, nor does it completely suppress popular contestants. Its role is closer to recalibrating the comparison scale between the two dimensions. For extreme popularity-driven cases such as Bobby Bones, DVSE weakens the excessive advantage caused by voting variance. For contestants with outstanding technical performance, such as Juan Pablo Di Pace, competitiveness can still be maintained after standardization. For cases such as Sean Spicer, where the outcome changes little across multiple mechanisms, DVSE does not force a change in the ranking judgment. In other words, this mechanism focuses more on statistical consistency between evaluation dimensions, rather than artificially creating an advantage for a certain type of contestant.

4.4. Mechanism Summary Based on Statistical Consistency

Overall, the DVSE mechanism recalibrates the traditional voting logic by introducing second-moment information, namely variance information. It both reduces the problem of high-variance fan voting dominating results in the percentage mechanism and lessens the information loss caused by discretization in the ranking mechanism.

The combined score S_{DVSE} can be viewed as a fairness filter. It does not remove the role of audience voting, but it prevents the scale of raw vote counts from having an excessive penetrating effect on the result. At the same time, it does not allow judges' scores to fully determine the final ranking. Instead, it compares technical performance and public support on the same statistical scale. Backtesting results based on full-season data show that this mechanism reduces the probability of extreme controversial results by 12.4%. Therefore, while keeping audience participation, DVSE provides more stable algorithmic support for technical evaluation.

5. Conclusion

This paper constructs an analytical framework that integrates implicit information inference, mechanism backtesting, and dynamic equilibrium optimization. By reconstructing the distribution of undisclosed votes using the MCMC method and comparing the performance of different aggregation rules, the results show that the overall consistency rate of historical samples is 66.3%, and 73.5% of anomalous results are associated with highly volatile votes under the proportional weighting mechanism. To address this issue, this paper proposes the DVSE mechanism, which standardizes and integrates different evaluation dimensions, thereby reducing the probability of extremely controversial outcomes by 12.4%. The study demonstrates that DVSE can mitigate the problem of single-dimension dominance in decision-making while preserving both expert evaluation and public participation, thereby enhancing the stability and fairness of outcomes. This method can also serve as a reference for multi-stakeholder integrated decision-making and ranking systems. Future research may further incorporate dynamic weight adjustment mechanisms.

References

[1] Dang Hong, Chen Yanjiao, Li Jianli. Applications of MCMC Algorithms in Numerical Simulation [J]. Statistics and Management, 2024, 39(10): 4–13. DOI: 10.16722/j.issn.1674-537x.2024.10.006.

[2] Liu Leping, Gao Lei, Yang Na. The Development of MCMC Methods and the Revival of Modern Bayesian Statistics—Commemorating the 250th Anniversary of the Discovery of Bayes' Theorem [J]. Forum of Statistics and Information, 2014, 29(2): 3-11.

[3] Zhu Yali. Bayesian Estimation of the SV Model Based on Markov Chain Monte Carlo Methods [D]. Tsinghua University, 2020. DOI: 10.27266/d.cnki.gqhau.2020.000194.

[4] Chen Mengcheng, Fang Wei, Yang Chao, et al. Bayesian Statistics and MCMC Methods—Matlab Implementation of the Metropolis-Hastings (M-H) Algorithm [J]. Journal of East China Jiaotong University, 2018, 35(1): 1-8. DOI: 10.16749/j.cnki.jecjtu.2018.01.001.

[5] Liu Junhong, Jin Ziyue. A Decision Model for Multi-Process Production Systems Based on Bayesian Inference and Dynamic Programming [J]. Technology and Market, 2026, 33(3): 16-22.

[6] Zhao Yanbin. A Study on Generative Multi-turn Dialogue Based on Conditional Variational Autoencoders [D]. South China University of Technology, 2023. DOI: 10.27151/d.cnki.ghnlu.2023.002462.