Sparse Online Principal Component for Skew Factor Model

Yu Jin, Di Chang, Guangbao Guo

Abstract—**This paper proposes a Skew Factor Model to handle data with skewness and heavy tails. It is assumed that the errors follow flexible skewed distributions. To enable efficient estimation in high-dimensional settings, we further develop a Sparse Online Principal Component method. This method enforces sparsity and works efficiently with streaming data. Simulation experiments show that it achieves higher accuracy and better sparsity than existing methods. The method is robust and scalable for large asymmetric data sets.**

***Index Terms*—Skew Factor Model; Sparse Online Principal Component; multivariate skew distributions; simulation experiments**

I. INTRODUCTION

SKEW Factor Model (SFM) is a statistical framework. It is specifically designed to handle data exhibiting skewness or asymmetry. SFM introduces skewness parameters to account for non-symmetric and heavy-tailed behavior. The error terms in this model follow multivariate skewed distributions. This structure improves robustness and enhances the accuracy of inference and prediction.

II. SKEW FACTOR MODEL

A. Multivariate Skew Distributions

In this section, we propose three types of multivariate skew distributions to model asymmetric and heavy-tailed data characteristics, which are commonly observed in practical applications.

1) Multivariate Skew Normal (MSN) Distribution: The MSN distribution extends the traditional multivariate normal distribution by incorporating skewness. Its probability density function (PDF) is

$$f_{\mathbf{X}}(\mathbf{x}) = 2\phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\alpha}^\top \mathbf{x}),$$

where $\boldsymbol{\mu}$ is the location vector, $\boldsymbol{\Sigma}$ is the scale matrix and $\boldsymbol{\alpha}$ is the shape parameter.

2) Multivariate Skew-Cauchy (MSC) Distribution: The MSC distribution extends the skew-cauchy distribution by incorporating a location parameter and a skewness shape. Its PDF is

$$f_{\mathbf{X}}(\mathbf{x}) = 2t_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, 1)\Phi(\boldsymbol{\alpha}^\top (\mathbf{x} - \boldsymbol{\mu})).$$

Manuscript received November 21, 2024; revised June 29, 2025.

This work was supported by a grant from the National Social Science Foundation Project under project ID 23BTJ059.

Yu Jin is a postgraduate student of Mathematics, Shandong University of Technology, Zibo, 255000, China (email: jinyu11172024@163.com).

Di Chang is a postgraduate student of Mathematics, Shandong University of Technology, Zibo, 255000, China (email: CD08210122@163.com).

Guangbao Guo is a professor of Mathematics and Statistics, Shandong University of Technology, Zibo, China (corresponding author to provide phone:15269366362; e-mail: ggb11111111@163.com).

3) Multivariate Skew-t (MST) Distribution: The MSC distribution is a skewed extension of the multivariate Cauchy distribution. It combines extremely heavy tails with skewness, making it suitable for modeling data with strong asymmetry and extreme values. Its PDF is

$$f_{\mathbf{X}}(\mathbf{x}) = 2t_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)T_1\left(\frac{\boldsymbol{\alpha}^\top \mathbf{x}}{\sqrt{(\nu + d)/(\nu + p)}}; \nu + p\right).$$

B. Factor Model

The traditional factor model posits that the observed data matrix $X \in \mathbb{R}^{n \times p}$ can be expressed as

$$X = FA^\top + \varepsilon,$$

where $A = (a_{ij})_{p \times m}$ denotes the factor loading matrix capturing the linear relationships between observed variables and latent factors; $F = (f_1, f_2, \dots, f_m) \in \mathbb{R}^{n \times m}$ represents the matrix of unobserved latent common factors that drive the shared variation among the observed variables; and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p) \in \mathbb{R}^{n \times p}$ is the matrix of idiosyncratic errors accounting for noise and unique variation specific to each observed variable.

C. Skew Factor Model

The SFM advances conventional factor models via skewed error distributions to address data asymmetry. The SFM is expressed as

$$X = FA^\top + \varepsilon, \quad \varepsilon \sim S(\mu, bI_{p \times p}),$$

where $\mathcal{S}$ represents a skew distribution, and μ is the location parameter, while $bI_{p\times p}$ scales the error terms.

III. SPARSE ONLINE PRINCIPAL COMPONENT ESTIMATION

A. Parameter Estimation

For traditional factor models, parameter estimation begins by centering the data matrix Y, yielding the centered matrix $\tilde{Y} = Y - \bar{Y}$.

The sample covariance matrix is then computed as

$$\Sigma_Y = \frac{1}{n-1}\tilde{Y}^\top \tilde{Y}.$$

An eigenvalue decomposition of Σ_Y is performed

$$\Sigma_Y = V\Lambda V^\top.$$

The latent factor scores are then estimated by projecting the centered data matrix onto the loading space

$$F = \tilde{Y}V_k,$$

where $\tilde{Y}$ denotes the centered data and V_k is the estimated loading matrix.

This procedure efficiently estimates parameters. It can also be extended to handle skewed or heavy-tailed errors by incorporating flexible distributional assumptions into the factor model.

B. Sparse Principal Component

The Sparse Principal Component (SPC) method extracts principal components from data. It applies sparsity to simplify the components. This helps remove redundant information. The optimization problem is expressed as

$$\min_{\Lambda} \|Y - \Lambda F\|_F^2 + \lambda \|\Lambda\|_1.$$

When the error term ϵ is assumed to follow a skew-normal distribution, the likelihood function L is expressed as

$$L(\Lambda, F, \gamma, \alpha, \omega) = \prod_{i=1}^{n} \prod_{j=1}^{p} f(y_{ij}; \alpha, \omega, \gamma).$$

Using the SPC approach on the SFM, the resulting optimization problem is formulated as

$$\min_{\Lambda, F} \left[\|Y - \Lambda F\|_F^2 + \lambda \|\Lambda\|_1 + \text{penalty}(\gamma) \right],$$

where $\|Y - \Lambda F\|_F^2$ represents the reconstruction error, $\|\Lambda\|_1$ is the sparsity term that enforces sparsity in the factor loading matrix Λ, and penalty(γ) is a regularization term associated with skewness.

C. Sparse Online Principal Component

The Sparse Online Principal Component (SOPC) method combines online updates with sparsity regularization.This method combines the strengths of both SPC and Online Principal Component (OPC) techniques. This method improves both efficiency and sparsity.

The OPC method incorporates a sparsity parameter θ into the online eigen-decomposition process, enabling the sequential estimation of sparse eigenvectors as new data arrives. Given the first $k < n$ observations, the data matrix is

$$X^k = \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_k^\top \end{pmatrix}.$$

The sample covariance matrix is given by

$$S^k = V^k \Lambda^k {V^k}^\top.$$

When the $(k+1)$-th observation X_{k+1} arrives, the updated sample covariance matrix is expressed as

$$S^{k+1} = \frac{k}{k+1} S^k + \frac{1}{k+1} X_{k+1}^\top X_{k+1}.$$

We perform eigendecomposition on S^{k+1} to obtain

$$S^{k+1} = V^{k+1} \Lambda^{k+1} {V^{k+1}}^\top.$$

Let $V_{SO}^{k+1} = V^{k+1}$ be the initial estimate of the sparse eigenvector matrix.

We update V_{SO}^{k+1} by solving the following sparse optimization problem, subject to the orthonormality constraint $W_{SO} W_{SO}^\top = I_{m \times m}$

$$V_{SO}^{k+1} = \arg \min_{V_S^{k+1}} \Big\{ \left\| X_{k+1} - W_{SO} {V_S^{k+1}}^\top X_{k+1} \right\|_F^2 + \rho \sum_{j=1}^{m} \|V_{S,j}^{k+1}\|_2^2 + \theta \sum_{j=1}^{m} \|V_{S,j}^{k+1}\|_1 \Big\}.$$

The associated loss function to be minimized is

$$L\left(V_{so}^{k+1}\right) = \text{trace}\left(X_{k+1}^\top X_{k+1}\right) + \sum_{j=1}^{m} \left[{V_{so,j}^{k+1}}^\top \left(X_{k+1}^\top X_{k+1} + \rho\right) V_{so,j}^{k+1} - 2 w_{so,j}^\top X_{k+1}^\top X_{k+1} V_{so,j}^{k+1} + \theta \|V_{so,j}\|_1 \right].$$

Based on the optimized sparse eigenvectors, the loading matrix $\hat{A}_{SO}^{k+1}$ and the specific variance matrix $\hat{D}_{SO}^{k+1}$ are computed as

$$\hat{A}_{SO}^{k+1} = \left(\sqrt{\lambda_1^{k+1}} v_{SO_1}^{k+1}, \sqrt{\lambda_2^{k+1}} v_{SO_2}^{k+1}, \cdots, \sqrt{\lambda_m^{k+1}} v_{SO_m}^{k+1} \right) = \left(\hat{a}_{SO_j}^{k+1} \right)_{p \times m},$$

$$\hat{D}_{SO}^{k+1} = \text{diag}\left(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_p^2\right), \quad \hat{\sigma}_i^2 = X_{k+1}^\top X_{k+1} \quad (i = 1, 2, \cdots, p).$$

In summary, SOPC produces an updated sparse loading matrix and a specific variance matrix. This allows for real-time updates and enhances model interpretability. Sparsity plays a key role in this improvement.

IV. SIMULATION EXPERIMENTS

This section presents two numerical studies. We choose the Mean Squared Error (MSE) of the factor loading matrix A as an indicator. The MSE is defined as

$$\text{MSE}_{\hat{A}} = \frac{1}{p^2} \|A - \hat{A}\|_F^2.$$

A. Case 1: the impact of sample size n

In Case 1, with fixed dimensions $(p, m) = (10, 5)$, the data matrix X is generated as follows. Before standardization, the parameters include a mean vector $\mu \sim U[0, 1000]$, entries $a_{ij} \sim U(-1, 1)$, factor matrix $F \sim N_m(0, I_{m \times m})$, and skewed errors $\varepsilon \sim \text{Skew}(0, D)$ where $D \in (0, 1)$. The sample size n varies from 2000 to 6000 in increments of 1000. The methods SOPC, SPC, Perturbation Principal Component (PPC), Stochastic Approximation Principal Component (SAPC), and Incremental Principal Component (IPC) are then evaluated under three skew distributions.

a) MSN distribution: As shown in Fig.1, the SOPC method consistently achieves the lowest estimation error, improving steadily from 0.40 at sample size 2000 to 0.34 at 6000, demonstrating strong robustness. IPC also shows a decreasing trend, dropping from 0.48 to 0.36, indicating stable improvement despite a higher initial error. SAPC declines monotonically from 0.49 to 0.36 but remains slightly less accurate than SOPC and IPC. PPC reaches its lowest error of 0.38 at sample size 4000 but fluctuates at other sizes, reflecting less stability. SPC shows a variable pattern, with a slight decrease followed by an increase at 4000, suggesting weaker performance overall. Overall, SOPC demonstrates the most favorable error profile across all sample sizes, followed by IPC and SAPC. PPC and SPC exhibit weaker performance in terms of both accuracy and stability.

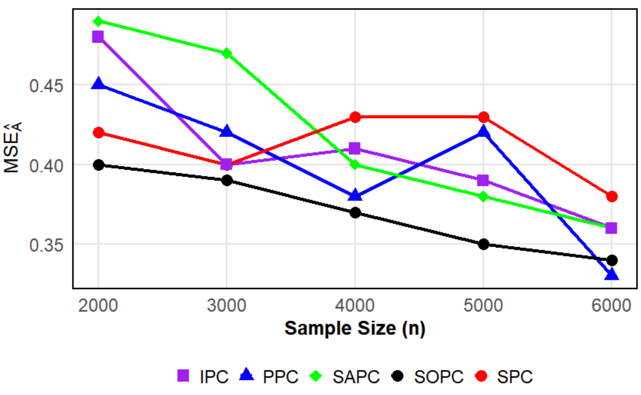

Fig. 1. $\mathrm{MSE}_{\hat{A}}$ results with the MSN distribution in Case 1

b) MSC distribution: As shown in Fig.2, the SOPC method consistently achieves the lowest error, falling from 0.43 at 2000 samples to 0.37 at 6000 samples, indicating robust and effective estimation. PPC closely follows, with errors dropping from 0.44 to 0.38, reflecting similar accuracy and stability. SPC and SAPC exhibit more irregular trends. SPC's error decreases from 0.46 to 0.39 but fluctuates slightly around sample size 4000. SAPC's error initially declines from 0.45 to 0.39 but then slightly increases to 0.41 at the largest sample size, suggesting minor instability. In contrast, IPC maintains relatively higher errors, starting at 0.48 and only slightly decreasing to 0.44, showing limited improvement as sample size grows. This implies less effective error reduction compared to the other methods.

In summary, SOPC and PPC demonstrate superior performance with clear decreasing error trends. SPC and SAPC show moderate performance with some fluctuations, while IPC exhibits the highest and most stable error levels.

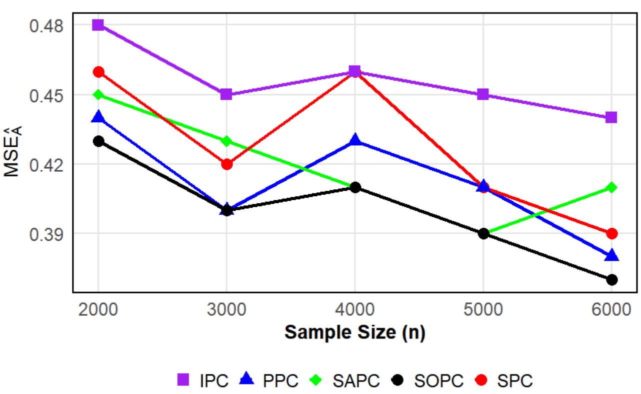

Fig. 2. $\mathrm{MSE}_{\hat{A}}$ results with the MSC distribution in Case 1

c) MST distribution: As shown in Fig.3, the SOPC method consistently attains the lowest or near-lowest error values, starting at 0.41 for sample size 2000 and decreasing to 0.37 at sample size 6000. This steady, slight decline suggests stable and effective estimation performance. The PPC and SAPC methods exhibit very similar patterns, with $\mathrm{MSE}_{\hat{A}}$ values gradually decreasing from around 0.43 and 0.43 respectively at the smallest sample size to 0.37 at the largest. Their curves closely track each other, indicating comparable accuracy and robustness. The SPC method shows a relatively stable trend with minor fluctuations; its error ranges between 0.42 and 0.38 over the sample sizes. Although its performance is close to that of PPC and SAPC, it is generally slightly worse than SOPC. The IPC method fluctuates somewhat more, starting with an error of 0.44, briefly decreasing to 0.41, then rising slightly again before ending at 0.39. This suggests less consistency in estimation accuracy compared to other methods.

In summary, all methods demonstrate improvements as sample size grows. SOPC slightly outperforms others in error reduction and stability, while PPC and SAPC show similar, strong performance. SPC remains competitive but slightly behind, and IPC exhibits the least stable trend.

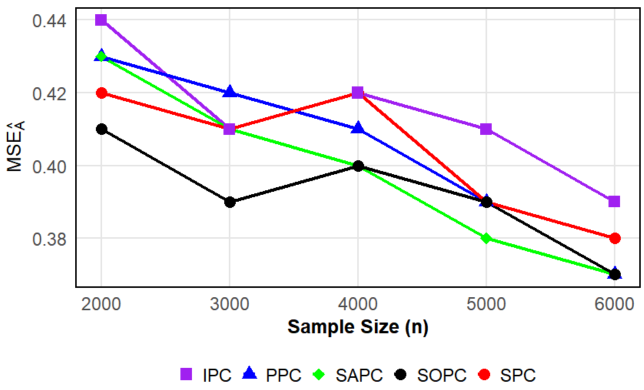

Fig. 3. $\mathrm{MSE}_{\hat{A}}$ results with the MST distribution in Case 1

B. Case 2: the impact of sample size p

In Case 2, with fixed parameters $(n, m) = (2000, 5)$, the data matrix X is generated as follows. The mean vector μ is uniformly sampled from $[0, 1000]$ before standardization, entries a_{ij} follow $U(-1, 1)$, the factor matrix F is drawn from $N_m(0, I_{m\times m})$, and skewed errors $\varepsilon \sim \mathrm{Skew}(0, D)$ where $D \in (0, 1)$. The dimension p varies from 10 to 14, evaluated under three skew distributions.

a) MSN distribution: As shown in Fig.4, the SOPC method consistently achieves the lowest estimation error across all dimensions. Its $\mathrm{MSE}_{\hat{A}}$ declines from 0.49 at dimension 10 to 0.40 at dimension 14, reflecting strong robustness and improved accuracy in higher dimensions. The IPC method also performs well, with relatively stable $\mathrm{MSE}_{\hat{A}}$ values between 0.48 and 0.44, though consistently higher than those of the SOPC method. The SAPC method shows a clear downward trend, with its error decreasing from 0.57 to 0.47, suggesting improved estimation as dimensionality increases. The SPC method exhibits moderate fluctuation, with $\mathrm{MSE}_{\hat{A}}$ values ranging from 0.53 to 0.46 and a local peak at dimension 13, indicating some instability. In contrast, the PPC method consistently shows the highest error, decreasing only from 0.64 to 0.54, and thus demonstrates the weakest performance among all methods.

In summary, the SOPC method demonstrates the most effective and stable estimation performance across all dimensions, followed by the IPC and SAPC methods. The SPC method shows moderate variability, whereas the PPC method performs the worst in terms of both accuracy and stability.

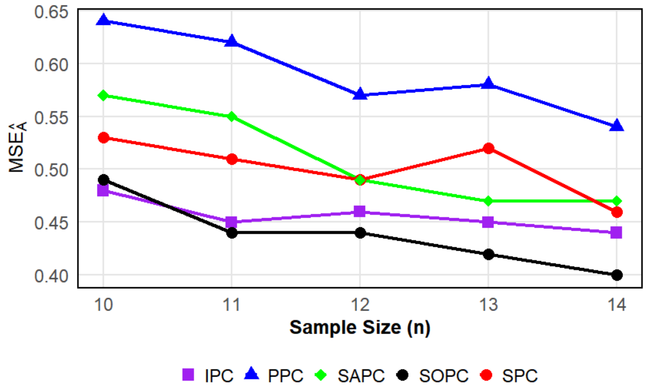

Fig. 4. $\text{MSE}_{\hat{A}}$ results with the MSN distribution in Case 2

b) MSC distribution: As shown in Fig.5, the SOPC method consistently achieves the lowest estimation error across all dimensions. Its $\text{MSE}_{\hat{A}}$ decreases from 0.46 at dimension 10 to 0.42 at dimension 13, followed by a slight increase to 0.43 at dimension 14. Despite this minor fluctuation, the overall trend remains favorable, indicating robust and accurate performance under increasing dimensionality. The IPC method exhibits a similar improvement, with its $\text{MSE}_{\hat{A}}$ decreasing steadily from 0.54 to 0.44, matching the SOPC method at the highest dimension. The PPC method also shows a downward trend, with $\text{MSE}_{\hat{A}}$ dropping from 0.54 to 0.44, although it fluctuates slightly around dimensions 12 and 13. The SAPC method starts with the highest error of 0.57 at dimension 10 and declines to 0.47 at dimension 14, demonstrating consistent but comparatively less accurate performance. The SPC method follows a similar trajectory, with $\text{MSE}_{\hat{A}}$ decreasing from 0.53 to 0.44, though it remains less stable than SOPC or IPC.

In summary, the SOPC method outperforms the others overall, particularly in lower dimensions. The IPC and PPC methods also perform well, while the SAPC and SPC methods show improvement but remain less competitive in terms of accuracy and stability.

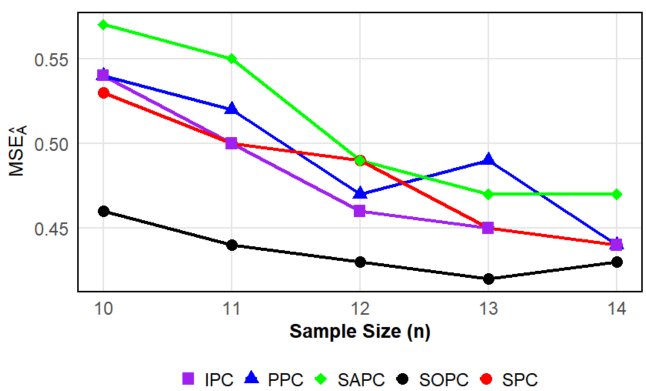

Fig. 5. $\text{MSE}_{\hat{A}}$ results with the MSC distribution in Case 2

c) MST distribution: As shown in Fig.6, the SOPC method achieves the lowest overall estimation error across increasing dimensions. Its $\text{MSE}_{\hat{A}}$ decreases steadily from 0.52 at dimension 10 to 0.42 at dimension 14, demonstrating strong robustness and stable accuracy in high-dimensional settings. The IPC method performs similarly well, with its $\text{MSE}_{\hat{A}}$ dropping from 0.48 to 0.40. This consistent downward trend highlights its reliable performance as dimensionality increases. The PPC method shows a non-monotonic pattern: its $\text{MSE}_{\hat{A}}$ first decreases from 0.48 to 0.43, then slightly increases to 0.46, indicating moderate accuracy but some fluctuation. The SAPC method starts with the highest error of 0.56 and declines to 0.45 at dimension 14, suggesting improved but relatively less accurate estimation. The SPC method follows a similar pattern, with $\text{MSE}_{\hat{A}}$ decreasing from 0.54 to 0.46 but peaking at dimension 13, which implies some instability in response to dimensional changes.

In summary, the SOPC and IPC methods exhibit the most favorable trends in error reduction, with the SOPC method slightly outperforming the others. The PPC method shows moderate competitiveness, while the SAPC and SPC methods demonstrate comparatively higher and less stable errors.

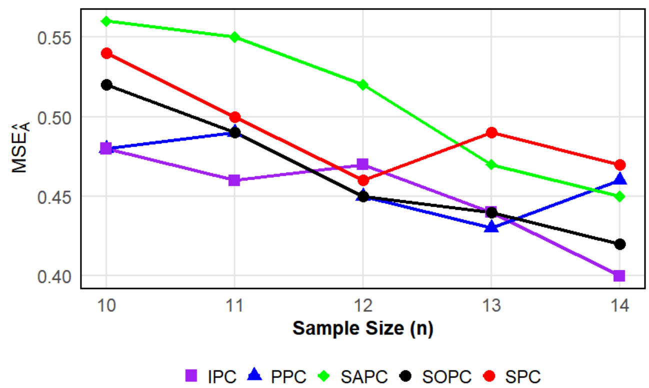

Fig. 6. $\text{MSE}_{\hat{A}}$results with the MST distribution in Case 2

V. CONCLUSION

We have developed a SFM that incorporates flexible skewed distributions to better capture asymmetry and heavy tails in the data. Building on this model, the proposed SOPC method enables efficient and scalable estimation in high-dimensional settings. Simulation results across multiple skewed distributions demonstrate that SOPC demonstrates superior performance over existing methods regarding both accuracy and sparsity. These findings confirm the robustness and practical value of SOPC for analyzing large, asymmetric data sets in practical applications.

REFERENCES

[1] Azzalini, A. (2014). *The Skew-Normal and Related Families*. Cambridge University Press.
[2] Arellano-Valle, R. B., Gómez, H. W., & Quintana, F. A. (2004). A new class of skew-normal distributions. *Communications in Statistics - Theory and Methods*, 58, 111–121.
[3] Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83, 715–726.
[4] Wang, J., Boyer, J., & Genton, M. G. (2004). A skew-symmetric representation of multivariate distributions. *Statistica Sinica*, 14, 1259–1270.
[5] Wei, C. (2022). Research on Parameter Estimation Methods of Factor Models [D]. Shandong University of Technology.
[6] DiCiccio, T. J., & Monti, A. C. (2011). Inferential aspects of the skew t-distribution. *Quaderni di Statistica*, 13, 1–21.

[7] Lee, S. X., & McLachlan, G. J. (2021). On formulations of skew factor models: Skew factors and/or skew errors. *Statistics & Probability Letters*, 168, 108935.
[8] Miao, Q., Cao, Y., Zhu, W., et al. (2024). Segregation flow behavior of polydisperse particle mixture with skewed distribution in a rotating drum. *Powder Technology*, 444, 120041.
[9] Zaka, A., Jabeen, R., Ahmad, M., et al. (2024). Control theory for skewed distribution under operation side of the telecommunication industry. *Measurement and Control*, 57(6), 703–723.
[10] Jondeep, D., Dimpal, P., Jyoti, P. H., et al. (2023). A new flexible alpha skew normal distribution. *Journal of the Indian Society for Probability and Statistics*, 24(2), 485–507.
[11] Lederer, J. (2020). *Fundamentals of High-Dimension Statistics*. Springer Nature Switzerland AG.
[12] Guo, G., Qian, G., & Zhu, L. (2022). A scalable quasi-Newton estimation algorithm for dynamic generalised linear models. *Journal of Nonparametric Statistics*, 34(4), 917–939.
[13] Guo, G., Qian, G., Lin, L., et al. (2021). Parallel inference for big data with the group Bayesian method. *Metrika*, 84, 225–243.
[14] Guo, G., Song, H., Zhu, L. (2024). The COR criterion for optimal subset selection in distributed estimation. *Statistics and Computing*, 34(5), 163.
[15] Guo, G., Sun, Y., & Jiang, X. (2020). A partitioned quasi-likelihood for distributed statistical inference. *Computational Statistics*, 35, 1577–1596.
[16] Guo, G. (2012). Parallel statistical computing for statistical inference. *Journal of Statistical Theory and Practice*, 6, 536–565.
[17] Guo, G., You, W., Qian, G., & Shao, W. (2015). Parallel maximum likelihood estimator for multiple linear regression models. *Journal of Computational and Applied Mathematics*, 273, 251–263.
[18] Guo, G., Sun, Y., Qian, G., & Wang, Q. (2022). LIC criterion for optimal subset selection in distributed interval estimation. *Journal of Applied Statistics*.
[19] Guo, G., Qian, G. (2025). Optimal subset selection for distributed local principal component analysis. *Physica A: Statistical Mechanics and its Applications*, 658, 130308.
[20] Chang, D., & Guo, G. (2024). LIC: An R package for optimal subset selection for distributed data. *SoftwareX*, 28, 101909.
[21] Wang, Q., Guo, G. B., Qian, G. Q., & Jiang, X. J. (2023). Distributed online expectation-maximization algorithm for Poisson mixture model. *Applied Mathematical Modelling*, 124, 734–748.
[22] Guo, G., Qian, G., & Zhu, L. (2023). Sparse online principal component analysis for parameter estimation in factor model. *Computational Statistics*, 38(2), 1095–1116.
[23] Guo, G., Yu, M., & Qian, G. (2024). ORKM: Online regularized K-means clustering for online multi-view data. *Information Sciences*, 680, 121133.
[24] Guo, G., Niu, R., Qian, G., et al. (2024). Trimmed scores regression for k-means clustering data with high-missing ratio. *Communications in Statistics - Simulation and Computation*, 53(6), 2805–2821.
[25] Guo, G., You, W., Lin, L., et al. (2016). Covariance matrix and transfer function of dynamic generalized linear models. *Journal of Computational and Applied Mathematics*, 296, 613–624.
[26] Guo, G. (2021). Taylor quasi-likelihood for limited generalized linear models. *Journal of Applied Statistics*, 48(4), 669–692.
[27] Guo, G., Song, H., & Zhu, L. (2025). The iterated score regression estimation algorithm for PCA-based missing data with high correlation. *Scientific Reports*, 15, 9067.
[28] Guo, G., Sun, Y., Qian, G., & Wang, Q. (2022). LIC: The LIC Criterion for Optimal Subset Selection. Available at: https://CRAN.R-project.org/package=LIC.
[29] Guo, G. (2025). TLIC: An R package for the LIC for T distribution regression analysis.
[30] Jing, G., & Guo, G. (2025). TLIC: An R package for the LIC for T distribution regression analysis. *SoftwareX*, 30, 102132.
[31] Guo, G., Wang, Q., Allison, J., et al. (2025). Accelerated distributed expectation-maximization algorithms for the parameter estimation in multivariate Gaussian mixture models. *Applied Mathematical Modelling*, 137, 115709.
[32] Zhang, C., & Guo, G. (2025). The Optimal Subset Estimation of Distributed Redundant Data. *IAENG International Journal of Applied Mathematics*, 55(2), 270–277.
[33] Liu, Q., & Guo, G. (2025). Distributed Estimation of Redundant Data. *IAENG International Journal of Applied Mathematics*, 55(2), 332–337.
[34] Li, J., & Guo, G. (2024). An Optimal Subset Selection Algorithm for Distributed Hypothesis Test. *IAENG International Journal of Applied Mathematics*, 54(12), 2811–2815.
[35] Jing, G., & Guo, G. (2025). Student LIC for Distributed Estimation. *IAENG International Journal of Applied Mathematics*, 55(3), 575–581.
[36] Song, L., & Guo, G. (2024). Full Information Multiple Imputation for Linear Regression Model with Missing Response Variable. *IAENG International Journal of Applied Mathematics*, 54(1), 77–81.
[37] Li, Y., & Guo, G. (2024). General Unilateral Loading Estimation. *Engineering Letters*, 32(1), 72–76.
[38] Shao, W., & Guo, G. (2018). Multiple-Try Simulated Annealing Algorithm for Global Optimization. *Mathematical Problems in Engineering*, 2018(1), 9248318.
[39] Guo, G., Shao, W., Lin, L., et al. (2016). Parallel tempering for dynamic generalized linear models. *Communications in Statistics - Theory and Methods*, 45(21), 6299–6310.
[40] Guo, G., & Lin, S. (2010). Schwarz method for penalized quasi-likelihood in generalized additive models. *Communications in Statistics—Theory and Methods*, 39(10), 1847–1854.
[41] Guo, G. (2020). A block bootstrap for quasi-likelihood in sparse functional data. *Statistics*, 54(5), 909–925.
[42] Guo, G., & Lin, L. (2016). Parallel bootstrap and optimal subsample lengths in smooth function models. *Communications in Statistics - Simulation and Computation*, 45, 2208–2231.
[43] Guo, G. (2018). Finite difference methods for the BSDEs in finance. *International Journal of Financial Studies*, 6(1), 26.
[44] Shao, W., Guo, G., Zhao, G., et al. (2014). Simulated annealing for the bounds of Kendall's τ and Spearman's ρ. *Journal of Statistical Computation and Simulation*, 84(12), 2688–2695.
[45] Shao, W., Guo, G., Meng, F., et al. (2013). An efficient proposal distribution for Metropolis–Hastings using a B-splines technique. *Computational Statistics & Data Analysis*, 57(1), 465–478.
[46] Guo, G., & Zhao, W. (2012). Schwarz methods for quasi stationary distributions of Markov chains. *Calcolo*, 49, 21–39.
[47] Guo, G., Allison, J., Zhu, L. (2019). Bootstrap maximum likelihood for quasi-stationary distributions. *Journal of Nonparametric Statistics*, 31(1), 64–87.
[48] You, W., Yang, Z., Guo, G., et al. (2019). Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble. *Knowledge-Based Systems*, 163, 598–610.
[49] Guangbao Guo, Yue Sun, Guoqi Qian, and Qian Wang. *LIC: The LIC Criterion for Optimal Subset Selection*, 2022. https://CRAN.R-project.org/package=LIC.
[50] Guangbao Guo, Haoyue Song, and Lixing Zhu. *COR: The COR for Optimal Subset Selection in Distributed Estimation*, 2021. https://CRAN.R-project.org/package=COR.
[51] Guangbao Guo, Guoqi Qian, Yixiao Liu, and Haoyue Song. *DLPCA: The Distributed Local PCA Algorithm*, 2022. https://CRAN.R-project.org/package=DLPCA.
[52] Guangbao Guo, Haoyue Song, and Lixing Zhu. *ISR: The Iterated Score Regression-Based Estimation Algorithm*, 2022. https://CRAN.R-project.org/package=ISR.
[53] Qian Wang, Guangbao Guo, and Guoqi Qian. *DEM: The Distributed EM Algorithms in Multivariate Gaussian Mixture Models*, 2022. https://CRAN.R-project.org/package=DEM.
[54] Qian Wang, Guangbao Guo, and Guoqi Qian. *DOEM: The Distributed Online Expectation Maximization Algorithms to Solve Parameters of Poisson Mixture Models*, 2022. https://CRAN.R-project.org/package=DOEM.
[55] Guangbao Guo, Chunjie Wei, and Guoqi Qian. *SOPC: The Sparse Online Principal Component Estimation Algorithm*, 2022. https://CRAN.R-project.org/package=SOPC.
[56] Guangbao Guo, Miao Yu, Haoyue Song, and Ruiling Niu. *ORKM: The Online Regularized K-Means Clustering Algorithm*, 2022. https://CRAN.R-project.org/package=ORKM.
[57] Chunjie Wei and Guangbao Guo. *OPC: The Online Principal Component Estimation Method*, 2022. https://CRAN.R-project.org/package=OPC.
[58] Guangbao Guo and Yaping Li. *DLEGFM: Distributed Loading Estimation for General Factor Model*, 2024. https://CRAN.R-project.org/package=DLEGFM.
[59] Guangbao Guo and Yu Li. *DIRMR: Distributed Imputation for Random Eeffects Models with Missing Responses*, 2024. https://CRAN.R-project.org/package=DIRMR.
[60] Guangbao Guo and Liming Song. *DLMRMV: Distributed Linear Models with Response Missing Variables*, 2024. https://CRAN.R-project.org/package=DLMRMV.
[61] Guangbao Guo and Jiarui Li. *pql: A Partitioned Quasi-likelihood for Distributed Statistical Inference*, 2024. https://CRAN.R-project.org/package=pql.
[62] Guangbao Guo and Jiarui Li. *FPCdpca: The FPCdpca Criterion on Distributed Principal Component Analysis*, 2024. https://CRAN.R-project.org/package=FPC.
[63] Guangbao Guo and Jiarui Li. *PPCDT: An Optimal Subset Selection for Distributed Hypothesis Testing*, 2024. https://CRAN.R-project.org/package=PPCDT.

[64] Guangbao Guo and Ruiling Niu. *DTSR: Distributed Trimmed Scores Regression for Handling Missing Data*, 2024. https://CRAN.R-project.org/package=DTSR.

[65] Guangbao Guo and Yaxuan Wang. *LLIC: Likelihood Criterion (LIC) Analysis for Laplace Regression Model*, 2024. https://CRAN.R-project.org/package=LLIC.

[66] Guangbao Guo and Di Chang. *Dogoftest: Distributed Online Goodness-of-Fit Tests for Distributed Datasets*, 2025. https://CRAN.R-project.org/package=Dogoftest.

[67] Guangbao Guo and Yu Jin. *SFM: A Package for Analyzing Skew Factor Models*, 2024. https://CRAN.R-project.org/package=SFM.

[68] Beibei Wu and Guangbao Guo. *TFM: Sparse Online Principal Component for TFM*, 2024. https://CRAN.R-project.org/package=TFM.

[69] Guangbao Guo and Guofu Jing. *TLIC: The LIC for T Distribution Regression Analysis*, 2024. https://CRAN.R-project.org/package=TLIC.

[70] Guangbao Guo and Siqi Liu. *LFM: Laplace Factor Model Analysis and Evaluation*, 2024. https://CRAN.R-project.org/package=LFM.