LEARNING FEATURE–STATE DEPENDENCE WITH SPARSE VARIATIONAL MIXTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional anomaly detection is often jointly affected by latent multimodality, variable dependence, and heavy-tailed contamination, making a single global distribution inadequate for characterizing the complex probabilistic structure of normal data. Mixture models can represent latent heterogeneous states through multiple local components, but existing methods typically overlook that variables may depend on the latent mixture state to different degrees, leading to parameter redundancy and unstable estimation in high-dimensional, finite-sample settings. To address this issue, we propose the Sparse Variational Student-t Mixture Factor Model (SVStMFM). The method uses Bernoulli structural variables to adaptively learn component-dependent and component-invariant parameters, while a low-rank Student-t factor structure captures both high-dimensional correlations and heavy-tailed behavior. We further develop structured variational Bayesian inference to jointly learn latent states, factor representations, local scales, and variable structures, and construct a locally adaptive anomaly score through a marginal–conditional decomposition of the Student-t distribution. Experiments on the Tennessee Eastman process and real industrial data show that SVStMFM consistently improves anomaly-detection performance while reducing model parameters by 83.1%. These results demonstrate that learning feature–state dependence provides a compact and robust solution for high-dimensional, multimodal, and heavy-tailed data.

1 INTRODUCTION

Anomaly detection aims to identify anomalous observations that deviate from the normal data distribution and is a fundamental problem in reliable machine learning and state awareness for complex systems. In high-dimensional tabular data such as industrial process data, latent operating-state changes, variable coupling, measurement noise, and outlier contamination often coexist, causing normal data to exhibit complex statistical properties such as multimodality, strong correlations, and heavy tails Zhang et al. (2023); Wang et al. (2024b); Chen et al. (2025); Yu et al. (2025a). Reliable anomaly detection therefore requires not only characterizing multiple latent normal states, but also stably estimating their local probabilistic structure in high-dimensional and contaminated data.

Both representation learning and probabilistic modeling have been widely used for high-dimensional anomaly detection Yin et al. (2024); Thimonier et al. (2024); Ye et al. (2025); Tsai et al. (2025); Peng et al. (2024). Mixture models explicitly represent latent heterogeneous structure through multiple local probability distributions, while Student-t mixture models further improve robustness to anomalous observations through heavy-tailed distributions Dai et al. (2024a); Li et al. (2025). However, conventional mixture models typically assume that all variables depend on the latent mixture state and repeatedly assign component-dependent parameters to every variable. In high-dimensional settings, this not only increases the number of parameters and the estimation variance, but may also mistake finite-sample fluctuations for genuine component differences.

Existing variable-selection methods alleviate high-dimensional redundancy by identifying variables relevant to clustering or anomaly detection Sun et al. (2018); Yan et al. (2019). However, *whether a variable is important* and *whether its distributional parameters depend on the latent mixture state* are two distinct questions. Some variables may have similar marginal statistical characteristics across states yet still contain useful information through their correlations with other variables. Directly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

removing these variables loses the joint dependency structure, whereas repeatedly estimating independent parameters for them in all components creates structural redundancy. We therefore consider a more fine-grained question: *Can we learn whether each variable genuinely depends on the latent mixture state while retaining the complete observation space and variable correlations?*

To this end, we propose the Sparse Variational Student-t Mixture Factor Model (SVStMFM). The model introduces a Bernoulli structural indicator for each variable to classify it as either a component-dependent variable (CDV) or a component-invariant variable (CIV). Both types of variables are retained in the complete observation space, and their correlations are jointly characterized through a low-rank Student-t latent-factor structure. We further develop structured variational Bayesian inference to jointly learn the variable structure, latent factors, mixture labels, local scales, and model parameters, and construct a locally adaptive anomaly score based on a marginal–conditional decomposition of the Student-t distribution.

The main contributions of this work are as follows:

(1) We propose a variable-level component-dependence structure-learning mechanism that adaptively determines whether a variable's parameters depend on the latent mixture state without removing the observed variable, thereby reducing unnecessary parameter replication across components.

(2) We propose a robust Student-t mixture factor model and a structured variational inference method. The low-rank latent representation reduces the complexity of modeling high-dimensional correlations and unifies heavy-tailed robust modeling and variable structure learning within a single probabilistic framework.

(3) We construct a probabilistic anomaly score consistent with the learned variable structure. The marginal distribution of the CIVs and the conditional distribution of the CDVs jointly characterize common deviations and local state-dependent deviations.

2 Related Work

2.1 Anomaly Detection and Probabilistic Modeling

Recent work on high-dimensional anomaly detection has primarily proceeded along the directions of representation learning, dependency-structure modeling, and probability density estimation. MCM learns the normal correlation structure among variables through masked cell modeling Yin et al. (2024); Wu et al. (2025); Thimonier et al. further jointly model feature–feature and sample–sample dependencies Thimonier et al. (2024); Dai et al. (2024b); Cho et al. (2025); and DRL alleviates the entanglement of normal and anomalous patterns in the latent space through decomposed representation learning Ye et al. (2025); Shentu et al. (2025). From a probabilistic modeling perspective, AnoLLM directly uses the negative log-likelihood to construct anomaly scores Tsai et al. (2025), while Variance-Stabilized Density Estimation and ConjNorm further investigate more stable and reliable density-scoring mechanisms Rozner et al. (2024); Peng et al. (2024). These studies indicate that effective anomaly detection depends not only on the capacity of latent representations, but also on the accurate characterization of variable dependencies and local probabilistic structure.

For data with multiple latent states, mixture models can explicitly characterize heterogeneous structure through multiple local probability distributions. IncVBGMM uses incremental variational Bayesian learning to adapt to distributional changes Dai et al. (2023), while MM-PPCA reduces the complexity of high-dimensional multistate modeling through a low-dimensional latent structure Lyu et al. (2024). In addition, incremental variational graph representation learning and probabilistic inference have also been used for adaptive process monitoring Lv et al. (2025); Peng et al. (2025). Other studies further combine dimensionality-reduced representations or local probabilistic models to characterize complex multimodal data Cui et al. (2024); Yu et al. (2025b). Fang et al. recently established a connection between clustering and anomaly detection from the perspective of probabilistic mixture models Fang et al. (2025). To handle outliers and heavy-tailed noise, VBSMM employs a variational Bayesian Student-t mixture model to improve robustness Dai et al. (2024a), and VBStMM-LLM further extends it to continuously evolving latent states Li et al. (2025). This work follows this robust probabilistic modeling direction but further focuses on variable-structure redundancy in high-dimensional mixture models.

2.2 Sparse Mixtures and Variable-Structure Learning

High-dimensional mixture models commonly face parameter redundancy and unstable covariance estimation. Existing studies reduce model complexity through sparse representations, latent dimensionality reduction, and variable selection, as in Sparse PCA and MM-PPCA Zhang et al. (2023); Lyu et al. (2024). HVMMM further shows that differentiated probabilistic modeling tailored to distinct variable statistical structures facilitates anomaly detection in complex data Wang et al. (2024a).

More closely related to this work is variable selection in mixture models. Yan et al. introduced variable selection into a Gaussian mixture model to mitigate the influence of irrelevant variables on high-dimensional multistate modeling Yan et al. (2019); Sun et al. introduced variable-saliency latent variables into a Student-t mixture model to jointly perform Bayesian clustering and feature selection Sun et al. (2018). These methods primarily determine whether variables are related to the latent clustering structure.

This work considers a different question: *whether a variable is important* is not equivalent to *whether its distributional parameters depend on the latent mixture state*. Therefore, SVStMFM learns not a conventional "retain–remove" structure, but a "component-dependent–component-invariant" structure. Component-invariant variables remain in the complete observation space and merely share parameters across mixture components, thereby reducing redundant parameters across components while preserving correlations among variables.

3 Sparse Variational Mixture Factor Model

3.1 Problem Formulation

Given the normal training set $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{x}_n = (x_{n1}, \ldots, x_{nD})^{\mathrm{T}} \in \mathbb{R}^D$. where N and D denote the numbers of samples and features, respectively, we assume that the data are jointly generated by K latent states. Let $z_n \in \{1, \ldots, K\}$ denote the latent state of sample $\boldsymbol{x}_n$.

High-dimensional data commonly exhibit multimodality, correlation, and heavy-tailed behavior simultaneously. As shown in Figure 1, the Student-t distribution is less sensitive than the Gaussian distribution to the influence of anomalous observations on parameter estimation. Conventional mixture models typically assume state-dependent parameters for every feature, although features may differ in their dependence on the latent states. As shown in Figure 2, assigning separate component parameters to all features may introduce redundancy and amplify finite-sample fluctuations.

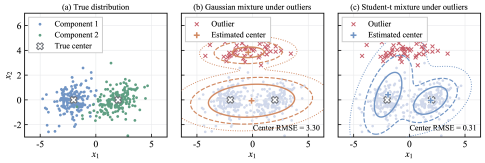


Figure 1: Robustness to outlier contamination.

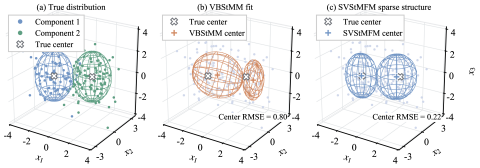


Figure 2: Feature–latent-state dependence.

Accordingly, we introduce a structural variable $\gamma_j \in \{0, 1\}$ for each feature j: $\gamma_j = 1$ indicates state-dependent parameters, whereas $\gamma_j = 0$ indicates parameters shared across states. Both types of features remain in the joint model. Thus, the purpose of this structural learning is to determine feature dependence on latent states rather than to remove features.

3.2 Model Structure of SVStMFM

1) Sparse robust factor-mixture modeling: Suppose the normal process data contain K latent components, with mixture weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^{\mathrm{T}}$, where $\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$. For sample $\boldsymbol{x}_n$, introduce the one-hot label $\boldsymbol{z}_n = (z_{n1}, \ldots, z_{nK})^{\mathrm{T}}$, where $z_{nk} \in \{0, 1\}$ and $\sum_{k=1}^K z_{nk} = 1$. Given the mixture weights, we have

$$p(\boldsymbol{Z} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}. \tag{1}$$

The mixture weights follow a Dirichlet prior,

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0), \tag{2}$$

where $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0K})^{\mathrm{T}}$.

A key assumption in conventional mixture models is that every variable changes with the latent component. SVStMFM relaxes this assumption by learning, at the distribution level, which variables are truly state dependent and which can share parameters across states. This induces a sparse feature–state dependence structure without removing variables from the observation space. Low-rank latent factors model cross-variable correlations, while local scale variables yield a robust Student-t representation. Figure 3 summarizes the resulting probabilistic model.

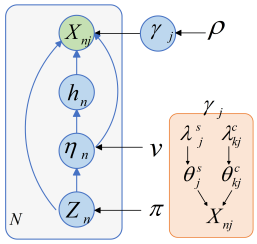


Figure 3: SVStMFM model.

Conventional Student-t mixture models generally assume that all process variables have component-dependent parameters, which requires repeated estimation of numerous mean and covariance parameters in high dimensions. In fact, variables can differ substantially in their dependence on the latent operating state. We therefore introduce a binary structural indicator $\gamma_j \in \{0, 1\}$ for the jth process variable. When $\gamma_j = 1$, the distributional parameters of x_j depend on the latent mixture component, and we call it a component-dependent variable (CDV). When $\gamma_j = 0$, the distributional parameters of the variable itself remain identical across mixture components, and we call it a component-invariant variable (CIV). Given the sparsity rate ρ,

$$p(\gamma_j \mid \rho) = \rho^{\gamma_j}(1-\rho)^{1-\gamma_j}, \qquad \gamma_j \in \{0, 1\}. \tag{3}$$

We further place a Beta prior on ρ,

$$p(\rho) = \mathrm{Beta}(\rho \mid a_{\rho 0}, b_{\rho 0}). \tag{4}$$

Thus, $a_{\rho 0}$ and $b_{\rho 0}$ can control the prior sparsity of the variable structure.

To reduce the number of high-dimensional covariance parameters while retaining correlations among process variables, we further introduce an r-dimensional latent factor $\boldsymbol{h}_n \in \mathbb{R}^r$, where $r \ll D$. To improve robustness against anomalous samples and heavy-tailed noise, we introduce a local scale variable η_n:

$$p(\eta_n \mid \nu) = \mathrm{Gam}\left(\eta_n \,\middle|\, \frac{\nu}{2}, \frac{\nu}{2}\right), \tag{5}$$

where the Gamma distribution uses the shape–rate parameterization and ν is the common degrees-of-freedom parameter. Given η_n, the latent factor satisfies

$$p(\boldsymbol{h}_n \mid \eta_n) = \mathcal{N}\left(\boldsymbol{h}_n \,\middle|\, \mathbf{0}, \eta_n^{-1}\boldsymbol{I}_r\right). \tag{6}$$

Define the augmented latent vector $\widetilde{\boldsymbol{h}}_n = [1, \boldsymbol{h}_n^{\mathrm{T}}]^{\mathrm{T}}$. For a component-dependent variable, let $\boldsymbol{\theta}_{kj}^c = [(\mu_{kj}^c), (\boldsymbol{b}_{kj}^c)^{\mathrm{T}}]^{\mathrm{T}}$ and $\lambda_{kj}^c = (\psi_{kj}^c)^{-1}$, where μ_{kj}^c, $\boldsymbol{b}_{kj}^c$, and ψ_{kj}^c denote the location parameter, factor loading, and residual scale, respectively, of the jth variable in the kth component. For a component-invariant variable, correspondingly define $\boldsymbol{\theta}_j^s = [\mu_j^s, (\boldsymbol{b}_j^s)^{\mathrm{T}}]^{\mathrm{T}}$ and $\lambda_j^s = (\psi_j^s)^{-1}$.

Given $z_{nk} = 1$, η_n, and $\boldsymbol{h}_n$, the conditional probability of the jth process variable can be written uniformly as

$$\begin{aligned} p(x_{nj} \mid z_{nk} = 1, \eta_n, \boldsymbol{h}_n, \gamma_j, \Theta) = &\left[\mathcal{N}\left(x_{nj} \,\middle|\, (\boldsymbol{\theta}_{kj}^c)^{\mathrm{T}}\widetilde{\boldsymbol{h}}_n, (\eta_n\lambda_{kj}^c)^{-1}\right)\right]^{\gamma_j} \\ &\times \left[\mathcal{N}\left(x_{nj} \,\middle|\, (\boldsymbol{\theta}_j^s)^{\mathrm{T}}\widetilde{\boldsymbol{h}}_n, (\eta_n\lambda_j^s)^{-1}\right)\right]^{1-\gamma_j}. \end{aligned} \tag{7}$$

As shown above, when $\gamma_j = 1$, μ_{kj}^c, $\boldsymbol{b}_{kj}^c$, and ψ_{kj}^c may all vary with the mixture component; hence, the location, marginal variation, and correlation structure with other variables are all component dependent. When $\gamma_j = 0$, these parameters are shared across all mixture components, reducing unnecessary parameter duplication.

Given the structural variables $\boldsymbol{\gamma}$, combine the means, factor loadings, and residual scales associated with all variables into $\boldsymbol{\mu}_k$, $\boldsymbol{B}_k$, and $\boldsymbol{\Psi}_k$, respectively, where $\boldsymbol{\Psi}_k$ is diagonal. Integrating out the

latent factor $\boldsymbol{h}_n$ gives

$$p(\boldsymbol{x}_n \mid z_{nk}=1,\eta_n)=\mathcal{N}\left(\boldsymbol{x}_n \left| \boldsymbol{\mu}_k, \eta_n^{-1}\boldsymbol{\Omega}_k\right.\right), \tag{8}$$

where

$$\boldsymbol{\Omega}_k=\boldsymbol{B}_k\boldsymbol{B}_k^{\mathrm{T}}+\boldsymbol{\Psi}_k. \tag{9}$$

Further integrating out the local scale η_n yields

$$p(\boldsymbol{x}_n \mid z_{nk}=1)=\mathrm{St}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k, \nu\right). \tag{10}$$

Thus, the low-rank term $\boldsymbol{B}_k\boldsymbol{B}_k^{\mathrm{T}}$ describes coupling among variables, whereas the diagonal term $\boldsymbol{\Psi}_k$ describes independent variable fluctuations that cannot be explained by the shared latent factors. Compared with direct estimation of a full covariance matrix, this parameterization uses latent factors with $r \ll D$ to compactly represent the correlation structure.

2) Priors and joint distribution: Given the latent factors and local scales, the observation model takes the form of Bayesian linear regression. We therefore use conjugate Normal–Gamma priors to jointly model the location parameters, factor loadings, and residual precisions. For the component-dependent parameters of the jth variable in the kth component,

$$p(\lambda_{kj}^c)=\mathrm{Gam}(\lambda_{kj}^c \mid a_0^c, b_0^c), \tag{11}$$

$$p(\boldsymbol{\theta}_{kj}^c \mid \lambda_{kj}^c)=\mathcal{N}\left(\boldsymbol{\theta}_{kj}^c \left| \boldsymbol{m}_0^c, (\lambda_{kj}^c \boldsymbol{R}_0^c)^{-1}\right.\right). \tag{12}$$

Similarly, the component-invariant parameters satisfy

$$p(\lambda_j^s)=\mathrm{Gam}(\lambda_j^s \mid a_0^s, b_0^s), \tag{13}$$

$$p(\boldsymbol{\theta}_j^s \mid \lambda_j^s)=\mathcal{N}\left(\boldsymbol{\theta}_j^s \left| \boldsymbol{m}_0^s, (\lambda_j^s \boldsymbol{R}_0^s)^{-1}\right.\right), \tag{14}$$

where $\boldsymbol{R}_0^c$ and $\boldsymbol{R}_0^s$ are both $(r+1)\times(r+1)$ positive-definite prior precision matrices.

Let $\Theta^c=\{\boldsymbol{\theta}_{kj}^c, \lambda_{kj}^c\}_{k=1,j=1}^{K,D}$ and $\Theta^s=\{\boldsymbol{\theta}_j^s, \lambda_j^s\}_{j=1}^{D}$. The joint probability distribution of the proposed model can then be written compactly as

$$\begin{aligned} p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{H},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\pi},\rho,\Theta^c,\Theta^s)&=p(\boldsymbol{\pi})p(\rho)p(\boldsymbol{\gamma}\mid\rho)p(\Theta^c)p(\Theta^s)\\ &\times p(\boldsymbol{Z}\mid\boldsymbol{\pi})p(\boldsymbol{\eta}\mid\nu)p(\boldsymbol{H}\mid\boldsymbol{\eta})p(\boldsymbol{X}\mid\boldsymbol{Z},\boldsymbol{H},\boldsymbol{\eta},\boldsymbol{\gamma},\Theta^c,\Theta^s). \end{aligned} \tag{15}$$

In summary, in the proposed SVStMFM, $\boldsymbol{Z}$ describes the latent multimodal structure of normal process data; η_n reduces the influence of anomalous observations on model-parameter estimation; $\boldsymbol{h}_n$ and the factor loadings compactly represent correlations among high-dimensional process variables; and the structural indicator γ_j adaptively determines whether the statistical parameters of a variable depend on the latent mixture component, thereby sparsifying the mixture structure and compressing the parameters while retaining the complete observation space.

3.3 Structured Variational Inference

Exact inference is intractable because the posterior couples mixture assignments, latent factors, local scales, structural indicators, and model parameters. We adopt the structured variational factorization

$$q=q(\boldsymbol{\pi})q(\rho)\prod_{j=1}^{D}q(\gamma_j)\prod_{k=1}^{K}\prod_{j=1}^{D}q(\boldsymbol{\theta}_{kj}^c,\lambda_{kj}^c)\prod_{j=1}^{D}q(\boldsymbol{\theta}_j^s,\lambda_j^s)\prod_{n=1}^{N}q(\boldsymbol{z}_n,\boldsymbol{h}_n,\eta_n), \tag{16}$$

with the structured local factor

$$q(\boldsymbol{z}_n,\boldsymbol{h}_n,\eta_n)=\prod_{k=1}^{K}\left[r_{nk}q(\boldsymbol{h}_n,\eta_n\mid z_{nk}=1)\right]^{z_{nk}}. \tag{17}$$

This form preserves the local dependence among the mixture assignment, latent factor, and local scale, while retaining tractable variational factors for the global parameters.

The variational factors are optimized by maximizing the evidence lower bound (ELBO),

$$\mathcal{L}(q)=\mathbb{E}_q[\ln p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{H},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\pi},\rho,\Theta^c,\Theta^s)]-\mathbb{E}_q[\ln q(\boldsymbol{Z},\boldsymbol{H},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\pi},\rho,\Theta^c,\Theta^s)]. \tag{18}$$

Each variational factor is updated by coordinate ascent while holding the remaining factors fixed.

The structural posterior is defined before it enters the local updates:

$$q(\gamma_j)=\mathrm{Bern}(\gamma_j\mid\phi_j),\qquad \phi_j=q(\gamma_j=1)=\mathbb{E}[\gamma_j]. \tag{19}$$

Using ϕ_j as a soft weight, the three quantities governing local inference can be written directly in terms of the CDV and CIV model parameters:

$$\begin{aligned}
\boldsymbol{A}_k &= \boldsymbol{I}_r + \sum_{j=1}^{D}\Big\{\phi_j \mathbb{E}\big[\lambda_{kj}^c \boldsymbol{b}_{kj}^c (\boldsymbol{b}_{kj}^c)^{\mathrm{T}}\big] + (1-\phi_j)\mathbb{E}\big[\lambda_j^s \boldsymbol{b}_j^s (\boldsymbol{b}_j^s)^{\mathrm{T}}\big]\Big\}, \\
\boldsymbol{g}_{nk} &= \sum_{j=1}^{D}\Big\{\phi_j \mathbb{E}\big[\lambda_{kj}^c \boldsymbol{b}_{kj}^c (x_{nj}-\mu_{kj}^c)\big] + (1-\phi_j)\mathbb{E}\big[\lambda_j^s \boldsymbol{b}_j^s (x_{nj}-\mu_j^s)\big]\Big\}, \\
c_{nk} &= \sum_{j=1}^{D}\Big\{\phi_j \mathbb{E}\big[\lambda_{kj}^c (x_{nj}-\mu_{kj}^c)^2\big] + (1-\phi_j)\mathbb{E}\big[\lambda_j^s (x_{nj}-\mu_j^s)^2\big]\Big\}.
\end{aligned} \tag{20}$$

The corresponding local-scale posterior and its parameters are

$$q(\eta_n \mid z_{nk}=1) = \mathrm{Gam}(\eta_n \mid a_{nk}^{\eta}, b_{nk}^{\eta}), \tag{21}$$

where, $a_{nk}^{\eta} = \frac{\nu+D}{2}$ and $b_{nk}^{\eta} = \frac{1}{2}\left[\nu + c_{nk} - \boldsymbol{g}_{nk}^{\mathrm{T}} \boldsymbol{A}_k^{-1} \boldsymbol{g}_{nk}\right]$. Integrating out $\boldsymbol{h}_n$ contributes a determinant factor proportional to $\eta_n^{-r/2}$, which cancels the apparent $r/2$ contribution to the Gamma shape. Hence the final shape is $(\nu+D)/2$, not $(\nu+D+r)/2$.

The mixture responsibility remains

$$r_{nk} = \frac{\exp(\ell_{nk})}{\sum_{l=1}^{K}\exp(\ell_{nl})}, \tag{22}$$

where ℓ_{nk} is the expected component-wise log joint probability. Its full expression is provided in Appendix C.

The responsibilities provide the effective component counts, yielding the conjugate posterior update

$$q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}), \tag{23}$$

where,$\alpha_k = \alpha_{0k} + \sum_{n=1}^{N} r_{nk}$. Thus, the inferred latent-state assignments directly update the mixture proportions.

The structural probability is then updated by Bayesian evidence comparison:

$$\mathrm{logit}(\phi_j) = \mathbb{E}[\ln \rho] - \mathbb{E}[\ln(1-\rho)] + \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\mathcal{E}_{nkj}^c - \mathcal{E}_{nkj}^s\right). \tag{24}$$

Here, $\mathcal{E}_{nkj}^c$ and $\mathcal{E}_{nkj}^s$ are the expected log-likelihoods under the CDV and CIV parameterizations, respectively; their closed-form expressions are given in Appendix C. Thus, ϕ_j combines the sparsity prior with the expected-likelihood difference between the two structures. It increases when the CDV branch provides a larger gain; otherwise, the model favors parameters shared across latent states. This is not conventional feature selection: every variable remains in the joint observation model.

The structural posterior probabilities further update the sparsity rate through Beta–Bernoulli conjugacy:

$$q(\rho) = \mathrm{Beta}(\rho \mid a_\rho, b_\rho), \tag{25}$$

where $a_\rho = a_{\rho 0} + \sum_{j=1}^{D}\phi_j$ and $b_\rho = b_{\rho 0} + D - \sum_{j=1}^{D}\phi_j$. This closes the feedback between feature-structure inference and the global sparsity prior.

Given the local posterior moments, both CDV and CIV parameter blocks retain Normal–Gamma conjugacy. The CDV branch is updated with effective weights $\phi_j r_{nk}$, whereas the CIV branch uses $(1-\phi_j)r_{nk}$ and pools information across mixture components. The common degrees of freedom ν is optimized by one-dimensional numerical root finding. These coordinate updates are iterated until ELBO convergence. Complete parameter updates, the numerical update of ν, and the expanded ELBO are provided in Appendix C.

The complete complexity analysis and variational-inference pseudocode are provided in Appendix B.

3.4 Anomaly Detection Based on SVStMFM

Based on the learned SVStMFM, we further construct a probabilistic anomaly score consistent with the variable structure. Rather than independently modeling different types of variables, SVStMFM

preserves correlations among all variables through latent factors. We therefore use a marginal–conditional decomposition of the multivariate Student-t distribution to characterize common cross-component deviations and local state-dependent deviations separately.

After the model converges, the final structure of each variable is determined from its structural posterior probability ϕ_j as $\hat{\gamma}_j = \mathbb{I}(\phi_j \geq 0.5)$. Let $\mathcal{C}$ and $\mathcal{I}$ denote the index sets of component-dependent and component-invariant variables, respectively, and write the observation as $\boldsymbol{x} = (\boldsymbol{x}^{\mathcal{C}}, \boldsymbol{x}^{\mathcal{I}})$. For the kth Student-t component, its location parameter and scale matrix can be written as $\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{Ck} \\ \boldsymbol{\mu}_I \end{bmatrix}$, $\boldsymbol{\Omega}_k = \begin{bmatrix} \boldsymbol{\Omega}_{CC,k} & \boldsymbol{\Omega}_{CI,k} \\ \boldsymbol{\Omega}_{IC,k} & \boldsymbol{\Omega}_{II} \end{bmatrix}$, where $\boldsymbol{\mu}_I$ and $\boldsymbol{\Omega}_{II}$ are shared across mixture components, whereas the cross terms retain the correlation structure between the two types of variables.

Because the multivariate Student-t distribution remains Student-t under both marginalization and conditioning, the marginal distribution of the component-invariant variables is $p(\boldsymbol{x}^{\mathcal{I}}) = \mathrm{St}\left(\boldsymbol{x}^{\mathcal{I}} \mid \boldsymbol{\mu}_I, \boldsymbol{\Omega}_{II}, \nu\right)$. Given $\boldsymbol{x}^{\mathcal{I}}$, the conditional distribution of $\boldsymbol{x}^{\mathcal{C}}$ under the kth component is

$$p_k(\boldsymbol{x}^{\mathcal{C}} \mid \boldsymbol{x}^{\mathcal{I}}) = \mathrm{St}\left(\boldsymbol{x}^{\mathcal{C}} \mid \boldsymbol{m}_k, s\,\boldsymbol{S}_k, \nu + d_I\right), \tag{26}$$

where $d_I = |\mathcal{I}|$ and $\boldsymbol{m}_k = \boldsymbol{\mu}_{Ck} + \boldsymbol{\Omega}_{CI,k}\boldsymbol{\Omega}_{II}^{-1}(\boldsymbol{x}^{\mathcal{I}} - \boldsymbol{\mu}_I)$, $\boldsymbol{S}_k = \boldsymbol{\Omega}_{CC,k} - \boldsymbol{\Omega}_{CI,k}\boldsymbol{\Omega}_{II}^{-1}\boldsymbol{\Omega}_{IC,k}$, $s = \frac{\nu+\delta}{\nu+d_I}$, $\delta = (\boldsymbol{x}^{\mathcal{I}} - \boldsymbol{\mu}_I)^{\mathrm{T}}\boldsymbol{\Omega}_{II}^{-1}(\boldsymbol{x}^{\mathcal{I}} - \boldsymbol{\mu}_I)$.

Because $p(\boldsymbol{x}^{\mathcal{I}})$ is identical across mixture components, it cancels during posterior normalization. Thus, for a test sample $\boldsymbol{x}_t$, the posterior responsibility is

$$r_{tk} = \frac{\hat{\pi}_k p_k(\boldsymbol{x}_t^{\mathcal{C}} \mid \boldsymbol{x}_t^{\mathcal{I}})}{\sum_{l=1}^{K} \hat{\pi}_l p_l(\boldsymbol{x}_t^{\mathcal{C}} \mid \boldsymbol{x}_t^{\mathcal{I}})}. \tag{27}$$

The instantaneous anomaly score is then defined as

$$G_t = -\log p(\boldsymbol{x}_t^{\mathcal{I}}) + \sum_{k=1}^{K} r_{tk}\left[-\log p_k(\boldsymbol{x}_t^{\mathcal{C}} \mid \boldsymbol{x}_t^{\mathcal{I}})\right]. \tag{28}$$

The first term measures the sample's deviation from the structure shared across components, while the second uses the posterior responsibilities to evaluate its conditional deviation from local latent states. This decomposition preserves the correlations between the two types of variables while avoiding the use of a single global mixture likelihood that could mask local anomalies. Equivalently, G_t can be viewed as a responsibility-weighted negative log-likelihood over the complete Student-t components.

For process data with temporal continuity, we further use an exponential moving average to accumulate anomaly information:

$$H_t = \lambda G_t + (1-\lambda)H_{t-1}, \tag{29}$$

where $\lambda \in (0, 1]$ is the smoothing coefficient. Using the normal training data, we obtain $\{H_n\}_{n=1}^{N}$ and take its $(1-\alpha)$ empirical quantile as the detection threshold, $H_{\lim} = Q_{1-\alpha}(\{H_n\}_{n=1}^{N})$. The current sample is identified as anomalous when $H_t \geq H_{\lim}$.

4 Experiments

4.1 Experimental Setup

The Tennessee Eastman process (TEP) dataset contains 52 process variables under six operating conditions and 21 fault types, including 41 measured and 11 manipulated variables. All sequences are sampled every 3 min. For each operating condition, 48 h of normal data are used for training and an additional 72 h sequence is used for testing, with faults injected at 60 h. We compare PCA, CVA, MPPCA, GMM, VBGMM, DAGMM, VBStMM, and SVStMFM using FAR, DR, Accuracy, F1, AUC, and training time. The process flow, complete dataset description, hyperparameter settings, and metric definitions are provided in Appendix D.1.

4.2 Main Results on the Tennessee Eastman Process

Table 1: Performance comparison of fault-monitoring methods on the TEP dataset

Method	FAR↓	DR↑	Accuracy↑	F1↑	AUC↑	Time (s)↓
PCAYin et al. (2014)	0.1113	0.7586	0.7803	0.8151	0.8868	**0.0175**
CVAXu et al. (2025)	0.1449	0.8537	0.8540	0.8774	0.9248	0.0198
MPPCALyu et al. (2024)	0.1080	0.7402	0.7655	0.8003	0.8781	0.2202
GMMYu & Qin (2008)	0.2777	0.8882	0.8605	0.9003	0.9252	0.0822
VBGMMDai et al. (2023)	0.2506	0.8845	0.8620	0.8998	0.9262	0.3779
DAGMMKim et al. (2026)	0.0938	0.5569	0.6151	0.6498	0.7968	4.7974
VBStMMDai et al. (2024a)	0.1015	0.8370	0.8472	0.8657	0.9218	0.7165
SVStMFM	**0.0646**	**0.9387**	**0.9301**	**0.9529**	**0.9731**	0.1851

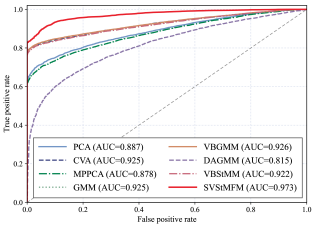
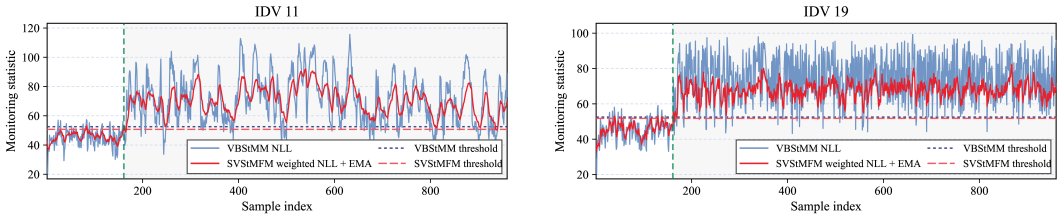


Figure 4: ROC curves on the TEP dataset

Table 1 summarizes the overall performance. SVStMFM achieves an FAR of 0.0646 and DR, Accuracy, F1, and AUC values of 0.9387, 0.9301, 0.9529, and 0.9731, respectively, outperforming all comparison methods. Relative to VBStMM, SVStMFM improves DR by 0.1017 and reduces FAR by 0.0369, showing that robust Student-t modeling and variable-level component dependence jointly suppress the influence of anomalous observations and redundant parameters. The ROC curves in Figure 4 provide a consistent comparison across decision thresholds. Figure 5 further presents the monitoring curves for IDV 11 and IDV 19. For both faults, the SVStMFM statistics rapidly cross their control limits after fault injection and remain at high levels, attaining DRs of 0.986 and 0.989, respectively, with zero FAR. The remaining per-fault DR/FAR results are reported in Table 2, while the remaining monitoring curves are provided in Appendix D.1.



(a) IDV 11 (b) IDV 19

Figure 5: Monitoring results of VBStMM and SVStMFM for representative TEP faults

4.3 Ablation Experiments and Structural Analysis

Table 2: Per-fault DR/FAR comparison of four models on the TEP dataset

Method	d01	d02	d03	d04	d05	d06	d07	d08	d09	d10	d11
VBStMM	1.000/0.106	0.989/0.081	0.169/0.150	1.000/0.075	1.000/0.075	1.000/0.038	1.000/0.038	0.983/0.106	0.145/0.206	0.920/0.050	0.874/0.088
S-VBGMM	0.995/0.006	0.986/0.019	0.156/0.119	0.900/0.019	0.368/0.019	1.000/0.006	1.000/0.000	0.980/**0.013**	0.135/0.313	0.611/0.000	0.700/0.006
VBStMM-NS	0.998/0.031	0.988/0.069	0.135/**0.075**	0.099/0.088	0.343/0.088	1.000/0.025	0.504/0.019	0.985/0.094	0.124/0.188	0.890/0.025	0.193/0.025
SVStMFM	**1.000/0.003**	**0.989/0.000**	**0.883**/0.263	**1.000/0.000**	**1.000/0.000**	**1.000/0.000**	**1.000/0.000**	**1.000**/0.213	**0.830/0.156**	**0.978/0.000**	**0.986/0.000**

Method	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	Mean
VBStMM	0.999/0.113	0.956/0.044	1.000/0.113	0.216/0.081	0.949/**0.206**	**0.975**/0.088	0.921/0.100	0.958/0.094	0.893/0.056	0.631/0.225	0.837/0.102
S-VBGMM	0.990/0.113	**0.958**/0.019	1.000/**0.019**	0.184/0.019	0.503/0.363	0.914/0.006	0.910/0.031	0.191/0.006	0.635/0.006	0.546/**0.113**	0.698/**0.058**
VBStMM-NS	0.999/0.050	0.955/0.006	0.079/0.069	0.161/0.075	0.946/0.219	0.344/0.081	0.913/0.056	0.876/0.050	0.854/0.025	0.644/0.231	0.620/0.076
SVStMFM	**1.000/0.000**	0.950/**0.000**	**1.000**/0.044	**0.485/0.000**	**1.000**/0.281	0.969/**0.000**	**0.979/0.000**	**0.989/0.000**	**0.911/0.000**	**0.769**/0.374	**0.939**/0.064

Table 2 isolates the effects of heavy-tailed modeling and variable-level structure learning. SVStMFM attains a mean DR of 0.939, exceeding S-VBGMM and VBStMM by 0.241 and 0.102, respectively, while reducing the mean FAR from 0.102 for VBStMM to 0.064. Figure 6 shows that SVStMFM uses only 1213 parameters, approximately 16.9%–21.2% of those required by the other models, while achieving a training time of 0.185 s and a mean F1 score of 0.953. Thus, the low-rank factor representation and cross-component parameter sharing substantially compress the model without sacrificing anomaly-detection performance. Figure 7 further shows that most structural posterior probabilities concentrate near 0 or 1. After thresholding, 29 variables are identified as

component-dependent variables and 23 as component-invariant variables, confirming that the model learns a clear feature/latent-state dependence structure from data.

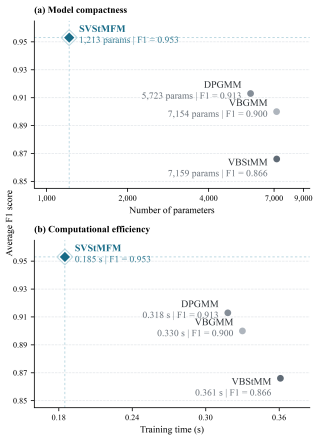


Figure 6: Comparison of parameter counts, training times, and mean F1 scores across models

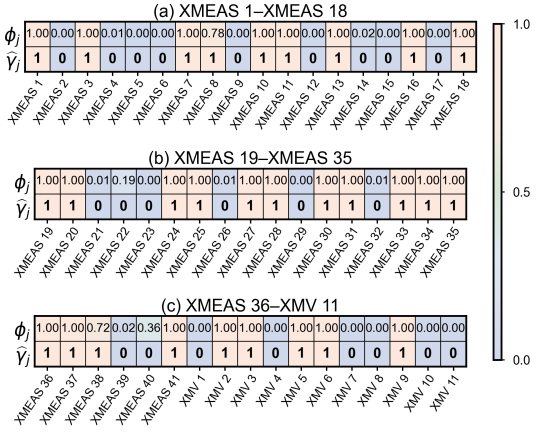


Figure 7: Variable-selection probabilities and variable identification results of SVStMFM

4.4 Real-World Industrial Validation

The SRCC dataset was collected from five ventilation exhaust fans (VEFs) in a steelmaking process from 2010 to 2020. It contains vibration, temperature, and related monitoring variables and covers four major fault types: unbalance, misalignment, looseness, and bearing faults Li et al. (2025).

Table 3: Performance comparison of fault-monitoring methods on the SRCC dataset

Method	FAR↓	DR↑	Accuracy↑	F1↑	AUC↑	Time (s)↓
PCAYin et al. (2014)	0.0600	0.3421	0.4901	0.5024	0.8163	1.0143
CVAXu et al. (2025)	0.1600	0.4342	0.5347	0.5841	0.7814	1.1167
MPPCALyu et al. (2024)	0.1200	0.4539	0.5594	0.6079	0.8167	2.0535
GMMYu & Qin (2008)	0.2600	0.5000	0.5594	0.6307	0.7529	**0.9136**
VBGMMDai et al. (2023)	0.2400	0.4342	0.5149	0.5739	0.7543	1.0161
DAGMMKim et al. (2026)	0.1800	0.2763	0.4109	0.4138	0.5630	3.0867
VBStMMDai et al. (2024a)	0.1000	0.2829	0.4356	0.4300	0.7693	1.1963
SVStMFM	**0.0000**	**0.7961**	**0.8465**	**0.8864**	**0.9443**	1.2114

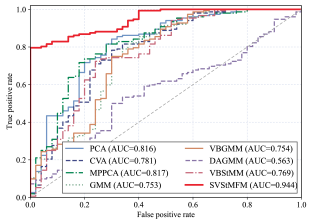


Figure 8: ROC curves on the SRCC dataset

As shown in Table 3, SVStMFM achieves an FAR of 0 and DR, F1, and AUC values of 0.7961, 0.8864, and 0.9443, respectively, outperforming all comparison methods. These results indicate that the sparse variable structure retains key fault information, while Student-t modeling mitigates noise and anomalous observations in long-term industrial measurements, enabling the model to generalize to complex real-world processes. Figure 8 further confirms the advantage across decision thresholds. Detailed process and equipment descriptions, fault definitions, and confusion matrices are provided in Appendix D.2.

5 Conclusion

We propose SVStMFM for learning feature–state dependence in high-dimensional anomaly detection. The model reduces redundant component-specific parameters through adaptive parameter sharing while preserving correlated observations. Experiments on benchmark and real industrial data show strong detection performance with substantial model compression, demonstrating the value of feature–state dependence learning for compact and robust probabilistic modeling.

AI USE STATEMENT

Generative AI tools were used only for language polishing. All AI-assisted content was reviewed by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work involves no human subjects, personal privacy, or sensitive personal data, and raises no additional ethical concerns.

REPRODUCIBILITY STATEMENT

The main paper describes the model, inference procedure, and experimental setup, while the appendix provides detailed derivations and implementation information for reproducibility.

REFERENCES

Wei Chen, Wenjie Guo, and Weijie Mao. A simultaneous multi-mode identification and outlier filtering method for imbalanced nonlinear process monitoring. *Chemical Engineering Science*, 313:121714, 2025. doi: 10.1016/j.ces.2025.121714.

Dongchan Cho, Jiho Han, Keumyeong Kang, Minsang Kim, Honggyu Ryu, and Namsoon Jung. Structured temporal causality for interpretable multivariate time series anomaly detection. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 59257–59294. Curran Associates, Inc., 2025. doi: 10.52202/085713-1983.

Yanfeng Cui, Wei Fan, and Yongzan Zhou. Dimensionality reducing gaussian mixture-based reconstruction for fault detection in multimode processes. *The Canadian Journal of Chemical Engineering*, 102(12):4267–4280, 2024. doi: 10.1002/cjce.25308.

Qingyang Dai, Chunhui Zhao, and Biao Huang. Incremental variational bayesian gaussian mixture model with decremental optimization for distribution accommodation and fine-scale adaptive process monitoring. *IEEE Transactions on Cybernetics*, 53(8):5094–5107, 2023. doi: 10.1109/TCYB.2022.3172790.

Qingyang Dai, Chunhui Zhao, and Shunyi Zhao. Variational bayesian student's-t mixture model with closed-form missing value imputation for robust process monitoring of low-quality data. *IEEE Transactions on Cybernetics*, 54(1):373–386, 2024a. doi: 10.1109/TCYB.2022.3230048.

Zhihao Dai, Ligang He, Shuang-Hua Yang, and Matthew Leeke. SARAD: Spatial association-aware anomaly detection and diagnosis for multivariate time series. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 48371–48410. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1533.

Zeyu Fang, Ming Gu, Sheng Zhou, Jiawei Chen, Qiaoyu Tan, Haishuai Wang, and Jiajun Bu. Towards a unified framework of clustering-based anomaly detection. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 15943–15959, 2025.

Bum Jin Kim, Yeonsoo Kim, and Jong Woo Kim. Explainable multimode process monitoring framework using deep autoencoding gaussian mixture model and kernel shapley additive explanations. *Journal of Process Control*, 165:103784, 2026. doi: 10.1016/j.jprocont.2026.103784.

Jingyan Li, Jiaju Man, Jinping Liu, Di Luo, Haidong Shao, and Yong-Fang Xie. VBStMM-LLM: Robust fault identification leveraging variational bayesian student's t-mixture modeling with lifelong learning mechanism. *IEEE Transactions on Reliability*, 74(4):4775–4790, 2025. doi: 10.1109/TR.2025.3601737.

M. Lv, Y. Li, H. Gao, B. Sun, C. Yang, and W. Gui. Toward adaptive and interpretable process monitoring: Incremental variational graph attention autoencoder with probabilistic inference. *IEEE Transactions on Cybernetics*, 55(9):4114–4127, 2025. doi: 10.1109/TCYB.2025.3583035.

Yuting Lyu, Le Zhou, Ya Cong, Hongbo Zheng, and Zhihuan Song. Multirate mixture probability principal component analysis for process monitoring in multimode processes. *IEEE Transactions on Automation Science and Engineering*, 21(2):2027–2038, 2024. doi: 10.1109/TASE.2023.3253285.

Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. ConjNorm: Tractable density estimation for out-of-distribution detection. In *International Conference on Learning Representations*, 2024.

Haotian Peng, Jiawei Liu, Jinsong Du, Jie Gao, and Wei Wang. BearLLM: A prior knowledge-enhanced bearing health management framework with unified vibration signal representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):19866–19874, 2025. doi: 10.1609/aaai.v39i19.34188.

Amit Rozner, Barak Battash, Henry Li, Lior Wolf, and Ofir Lindenbaum. Anomaly detection with variance stabilized density estimation. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pp. 3121–3137, 2024.

Qichao Shentu, Beibu Li, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 81358–81381, 2025.

Jianyong Sun, Aimin Zhou, Simeon Keates, and Shengbin Liao. Simultaneous bayesian clustering and feature selection through student's t mixtures model. *IEEE Transactions on Neural Networks and Learning Systems*, 29(4):1187–1199, 2018. doi: 10.1109/TNNLS.2016.2619061.

Hugo Thimonier, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. Beyond individual input for deep anomaly detection on tabular data. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 48097–48123, 2024.

Che-Ping Tsai, Ganyu Teng, Phillip Wallis, and Wei Ding. AnoLLM: Large language models for tabular anomaly detection. In *International Conference on Learning Representations*, 2025.

Min Wang, Donghua Zhou, and Maoyin Chen. Hybrid variable monitoring mixture model for anomaly detection in industrial processes. *IEEE Transactions on Cybernetics*, 54(1):319–331, 2024a. doi: 10.1109/TCYB.2022.3228524.

Zhiqiong Wang, Renping Gong, Lisha Song, Shuguang He, and Yuan Gao. A data-driven monitoring scheme for multivariate multimodal data. *Computers & Industrial Engineering*, 192:110186, 2024b. doi: 10.1016/j.cie.2024.110186.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 17017–17045, 2025.

Chong Xu, Daoping Huang, Guangping Yu, and Yiqi Liu. Distributed process monitoring of the large-scale system using spatio-temporal-causality and wasserstein-distance-based canonical variate analysis. *Journal of Process Control*, 146:103367, 2025. doi: 10.1016/j.jprocont.2024.103367.

Dandan Yan, Shuai Zhang, and Uk Jung. A variable-selection control chart via penalized likelihood and gaussian mixture model for multimodal and high-dimensional processes. *Quality and Reliability Engineering International*, 35(4):1263–1275, 2019. doi: 10.1002/qre.2458.

Hangting Ye, He Zhao, Wei Fan, Mingyuan Zhou, Dandan Guo, and Yi Chang. DRL: Decomposed representation learning for tabular anomaly detection. In *International Conference on Learning Representations*, 2025.

Jiaxin Yin, Yuanyuan Qiao, Zitang Zhou, Xiangchao Wang, and Jie Yang. MCM: Masked cell modeling for anomaly detection in tabular data. In *International Conference on Learning Representations*, 2024.

S. Yin, S. X. Ding, X. Xie, and H. Luo. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics*, 61(11):6418–6428, 2014.

Jie Yu and S. Joe Qin. Multimode process monitoring with bayesian inference-based finite gaussian mixture models. *AIChE Journal*, 54(7):1811–1829, 2008. doi: 10.1002/aic.11515.

W. Yu, B. Huang, G. Xiao, and C.-K. Zhang. A variational bayesian inference-based robust dissimilarity analytics model for industrial fault detection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55(5):3275–3286, 2025a. doi: 10.1109/TSMC.2025.3538854.

Wanke Yu, Chunhui Zhao, Biao Huang, and Hui Yang. A probabilistic quality-relevant monitoring method with gaussian mixture model. *IEEE Transactions on Automation Science and Engineering*, 22:4790–4801, 2025b. doi: 10.1109/TASE.2024.3411471.

J. Zhang, D. Zhou, and M. Chen. Self-learning sparse PCA for multimode process monitoring. *IEEE Transactions on Industrial Informatics*, 19(1):29–39, 2023.

A PRELIMINARIES

Complex industrial process data are often simultaneously characterized by multimodality, strong noise, and anomalous observations. Although conventional Gaussian mixture models can represent multimodal distributions, their Gaussian component assumption makes them sensitive to outliers. When the training data are contaminated by sensor noise, operating condition disturbances, or local anomalies, the estimated means and covariance matrices of Gaussian components may shift substantially, thereby degrading subsequent fault identification. In contrast, the Student-t distribution has heavier tails and controls its robustness to anomalous observations through the degrees-of-freedom parameter. Student-t mixture models can therefore characterize both the multimodal structure and the non-Gaussian heavy-tailed behavior of industrial process data.

A.1 STUDENT-t MIXTURE MODEL

Let the observed dataset be $X = \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{x}_n \in \mathbb{R}^D$. A multivariate Student-t distribution can be represented as a Gaussian–Gamma scale mixture:

$$\mathrm{St}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}\big(\boldsymbol{x} \mid \boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\big)\, \mathrm{Gam}\Big(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\Big)\, \mathrm{d}\eta. \tag{30}$$

Here, $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Lambda}$ is the precision matrix, ν is the degrees-of-freedom parameter, and η is a local scale variable. Under the shape–rate parameterization, $\mathrm{Gam}(\eta \mid a, b) = \frac{b^a}{\Gamma(a)}\eta^{a-1}\exp(-b\eta)$.

Integrating out the scale variable gives the explicit density of the Student-t distribution:

$$\mathrm{St}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma((\nu+D)/2)\,|\boldsymbol{\Lambda}|^{1/2}}{\Gamma(\nu/2)(\pi\nu)^{D/2}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-(\nu+D)/2}. \tag{31}$$

Here, $\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})$ is the Mahalanobis distance. Because ν controls the tail thickness, a smaller ν assigns greater probability density to anomalous observations far from the distribution center, reducing their influence on parameter estimation. As $\nu \to \infty$, the Student-t distribution approaches a Gaussian distribution.

To describe latent local heterogeneity and heavy-tailed behavior in industrial process data, we introduce K Student-t components to construct a Student-t mixture model with density

$$p(\boldsymbol{x} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\nu}) = \sum_{k=1}^K \pi_k\, \mathrm{St}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \nu_k), \tag{32}$$

where π_k is the weight of the kth mixture component and satisfies $\sum_{k=1}^K \pi_k = 1$ and $\pi_k > 0$. The quantities $\boldsymbol{\mu}_k$, $\boldsymbol{\Lambda}_k$, and ν_k denote the mean, precision matrix, and degrees-of-freedom parameter of the kth component, respectively.

To represent the association between samples and mixture components, we introduce the latent variable $\boldsymbol{z}_n = (z_{n1}, z_{n2}, \ldots, z_{nK})^\mathsf{T}$, where $z_{nk} \in \{0, 1\}$ and $\sum_{k=1}^K z_{nk} = 1$. When $z_{nk} = 1$,

sample $\boldsymbol{x}_n$ belongs to the kth Student-t component. Given the mixture weights $\boldsymbol{\pi}$, the latent variables follow a categorical distribution:

$$p(Z \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}. \tag{33}$$

With the scale variable η_{nk}, the conditional distribution of a sample assigned to a component is

$$p(\boldsymbol{x}_n \mid z_{nk} = 1, \eta_{nk}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, (\eta_{nk}\boldsymbol{\Lambda}_k)^{-1}\right), \tag{34}$$

$$p(\eta_{nk} \mid z_{nk} = 1, \nu_k) = \mathrm{Gam}\Big(\eta_{nk} \mid \frac{\nu_k}{2}, \frac{\nu_k}{2}\Big). \tag{35}$$

The Student-t mixture model thus uses multiple components to describe latent operating substates induced by local load fluctuations, equipment-state changes, and nonlinear variable coupling. Its local scale variables and degrees-of-freedom parameters additionally improve robustness to outliers and heavy-tailed noise.

A.2 Parameter Estimation for Mixture Models

Parameter estimation for mixture models involves not only estimating component parameters and mixture weights but also inferring the latent component assignment of each sample. Because latent variables make it generally impossible to obtain a closed-form solution to the observed-data likelihood by directly setting its gradient to zero, direct optimization becomes difficult. The expectation-maximization (EM) algorithm and variational inference (VI) are two common approaches to parameter estimation in such latent-variable models.

Let the observed dataset be $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, the set of latent variables be $\boldsymbol{Z}$, and the set of model parameters be $\boldsymbol{\theta}$. Because the component assignments are unobserved, the observed-data likelihood is obtained by marginalizing the latent variables, and the corresponding log-likelihood is

$$\ell(\boldsymbol{\theta}) = \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}). \tag{36}$$

The EM algorithm solves this problem by alternating between an expectation step (E-step) and a maximization step (M-step). At iteration t, the E-step computes the conditional posterior distribution of the latent variables under the current parameters $\boldsymbol{\theta}^{(t)}$ and constructs the conditional expectation of the complete-data log-likelihood:

$$Q\Big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}\Big) = \mathbb{E}_{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)})}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})\right]. \tag{37}$$

The M-step updates the model parameters by maximizing this expectation:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q\Big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}\Big). \tag{38}$$

These two steps alternate until the log-likelihood or model parameters satisfy a given convergence criterion.

For probabilistic models with multiple layers of latent variables and complex parameter dependencies, however, the exact posterior distribution of the latent variables and parameters is generally analytically intractable. Variational inference approximates the exact posterior with a tractable family of probability distributions, thereby transforming Bayesian posterior inference into deterministic optimization over probability distributions.

In a Bayesian modeling framework, collect the latent variables and model parameters as $\boldsymbol{\Phi} = \{\boldsymbol{Z}, \boldsymbol{\theta}\}$. Bayes' rule gives the exact posterior

$$p(\boldsymbol{\Phi} \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X}, \boldsymbol{\Phi})}{p(\boldsymbol{X})}, \tag{39}$$

where $p(\boldsymbol{X}, \boldsymbol{\Phi})$ is the joint distribution of the observed and unknown variables, and $p(\boldsymbol{X}) = \int p(\boldsymbol{X}, \boldsymbol{\Phi})\, \mathrm{d}\boldsymbol{\Phi}$ is the model evidence.

When this integral is analytically intractable, a tractable variational distribution $q(\boldsymbol{\Phi})$ is introduced to approximate the exact posterior $p(\boldsymbol{\Phi} \mid \boldsymbol{X})$. The log evidence decomposes as

$$\begin{aligned}\ln p(\boldsymbol{X}) &= \int q(\boldsymbol{\Phi}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{\Phi})}{q(\boldsymbol{\Phi})} \, \mathrm{d}\boldsymbol{\Phi} \\ &\quad + \int q(\boldsymbol{\Phi}) \ln \frac{q(\boldsymbol{\Phi})}{p(\boldsymbol{\Phi} \mid \boldsymbol{X})} \, \mathrm{d}\boldsymbol{\Phi} \\ &= \mathcal{L}(q) + \mathrm{KL}[q(\boldsymbol{\Phi}) \| p(\boldsymbol{\Phi} \mid \boldsymbol{X})] \,, \end{aligned} \tag{40}$$

where

$$\mathcal{L}(q) = \int q(\boldsymbol{\Phi}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{\Phi})}{q(\boldsymbol{\Phi})} \, \mathrm{d}\boldsymbol{\Phi} \tag{41}$$

is the evidence lower bound (ELBO). Equivalently,

$$\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{\Phi})} \left[\ln p(\boldsymbol{X}, \boldsymbol{\Phi}) \right] - \mathbb{E}_{q(\boldsymbol{\Phi})} \left[\ln q(\boldsymbol{\Phi}) \right] . \tag{42}$$

The first term is the expectation of the joint log probability under the variational distribution, whereas the second term is the entropy of the variational distribution. Because the KL divergence is non-negative, $\mathcal{L}(q) \leq \ln p(\boldsymbol{X})$. Maximizing the ELBO is therefore equivalent to minimizing the KL divergence between the variational distribution and the exact posterior:

$$q^*(\boldsymbol{\Phi}) = \arg\min_q \mathrm{KL}[q(\boldsymbol{\Phi}) \| p(\boldsymbol{\Phi} \mid \boldsymbol{X})] = \arg\max_q \mathcal{L}(q). \tag{43}$$

To obtain a tractable variational distribution, the mean-field assumption typically factorizes it into mutually independent factors:

$$q(\boldsymbol{\Phi}) = \prod_{j=1}^{M} q_j(\boldsymbol{\Phi}_j). \tag{44}$$

The mean-field factorization does not imply that the unknown variables in the original probabilistic model are mutually independent; rather, it imposes a factorization constraint on the approximate posterior to reduce the computational complexity of posterior inference.

Fixing the remaining variational factors, optimizing the ELBO with respect to $q_j(\boldsymbol{\Phi}_j)$ under the normalization constraint $\int q_j(\boldsymbol{\Phi}_j) \, \mathrm{d}\boldsymbol{\Phi}_j = 1$ gives

$$\ln q_j^*(\boldsymbol{\Phi}_j) = \mathbb{E}_{q_{-j}} \left[\ln p(\boldsymbol{X}, \boldsymbol{\Phi}) \right] + \mathrm{const}, \tag{45}$$

where $q_{-j} = \prod_{i \neq j} q_i(\boldsymbol{\Phi}_i)$ and $\mathbb{E}_{q_{-j}}[\cdot]$ denotes the expectation over the remaining unknown variables. The variational factors can thus be updated alternately by coordinate ascent until the ELBO converges.

Compared with the point estimates obtained by EM, variational inference provides approximate posterior distributions for both the model parameters and latent variables, thereby quantifying parameter-estimation uncertainty and using prior distributions to regularize model complexity. Variational inference consequently provides a parameter learning framework for Bayesian mixture models with complex latent-variable structure.

B Complexity Analysis and Algorithm Pseudocode

B.1 Complexity Analysis

Let the numbers of training samples, variables, mixture components, and latent factors be N, D, K, and r, respectively, where $r \ll D$. In each variational iteration, the dominant computation for updating the local latent variables and responsibilities arises from second-order moment operations between the D observed variables and the r-dimensional latent factors, with complexity $O(NKDr^2)$. Updating the component-wise factor posterior parameters additionally requires inversions of $(r+1) \times (r+1)$ matrices, with complexity $O(KDr^3)$. Therefore, the overall time complexity per iteration of SVStMFM is

$$O\big(NKDr^2 + KDr^3\big) \,. \tag{46}$$

After T iterations, the total time complexity is $O\big(T(NKDr^2 + KDr^3)\big)$. In the typical case $N \gg r$, the computation is dominated by $O(NKDr^2T)$. In comparison, mixture models using

full covariance matrices require $O(NKD^2T)$ complexity. Thus, when $r^2 \ll D$, the low-rank factor representation can substantially reduce the computational cost of high-dimensional probabilistic modeling.

In terms of parameter count, a mixture model using full covariance matrices must estimate a D-dimensional mean vector and a $D \times D$ symmetric covariance matrix for each component, giving a dominant distributional parameter count of $KD(D+3)/2$. For SVStMFM, let D_c and D_s denote the numbers of component-dependent and component-invariant variables ultimately identified, respectively, where $D_c + D_s = D$. Since the former have independent parameters across components whereas the latter share a single parameter set, the final effective number of distributional parameters is

$$P_{\mathrm{SVStMFM}} = (r+2)(KD_c + D_s). \tag{47}$$

Accounting further for the mixture weights, common degrees of freedom, and sparsity-rate parameter adds only $K+1$ scalar parameters. Therefore, compared with the $O(KD^2)$ parameters required by a full-covariance mixture model, SVStMFM uses a low-rank factor representation and parameter sharing across components to reduce the model size to depend approximately linearly on the observed dimension D and latent dimension r.

B.2 Variational Inference Algorithm

C Complete Variational Inference Derivations

This appendix retains the complete coordinate updates and derivations underlying the inference loop summarized in Section 3.3, including the computational details omitted from Eq. (16). The Gamma distribution is parameterized by shape and rate. Throughout, $\gamma_j = 1$ denotes a component-dependent variable (CDV), whereas $\gamma_j = 0$ denotes a component-invariant variable (CIV). Both types remain in the joint observation model.

C.1 Variational Factorization and Useful Moments

We use the structured family

$$q = q(\boldsymbol{\pi})q(\rho)\prod_{j=1}^{D} q(\gamma_j)\prod_{k=1}^{K}\prod_{j=1}^{D} q(\boldsymbol{\theta}_{kj}^c, \lambda_{kj}^c)\prod_{j=1}^{D} q(\boldsymbol{\theta}_j^s, \lambda_j^s)\prod_{n=1}^{N} q(\boldsymbol{z}_n, \boldsymbol{h}_n, \eta_n), \tag{48}$$

$$q(\boldsymbol{z}_n, \boldsymbol{h}_n, \eta_n) = \prod_{k=1}^{K} \left[r_{nk} q(\boldsymbol{h}_n, \eta_n \mid z_{nk} = 1)\right]^{z_{nk}}. \tag{49}$$

For a generic Normal–Gamma factor,

$$q(\boldsymbol{\theta}, \lambda) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}, (\lambda \boldsymbol{R})^{-1})\mathrm{Gam}(\lambda \mid a, b), \tag{50}$$

the moments required below are

$$\bar{\lambda} = \mathbb{E}[\lambda] = \frac{a}{b}, \qquad \ell_\lambda = \mathbb{E}[\ln \lambda] = \psi(a) - \ln b, \qquad \boldsymbol{v} = \mathbb{E}[\lambda \boldsymbol{\theta}] = \bar{\lambda}\boldsymbol{m}, \tag{51}$$

and

$$\boldsymbol{M} = \mathbb{E}[\lambda \boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}] = \boldsymbol{R}^{-1} + \bar{\lambda}\boldsymbol{m}\boldsymbol{m}^{\mathrm{T}}. \tag{52}$$

Writing $\boldsymbol{\theta} = [\mu, \boldsymbol{b}^{\mathrm{T}}]^{\mathrm{T}}$, partition these moments as

$$\boldsymbol{v} = \begin{bmatrix} v_0 \\ \boldsymbol{v}_b \end{bmatrix}, \qquad \boldsymbol{M} = \begin{bmatrix} M_{00} & \boldsymbol{M}_{0b} \\ \boldsymbol{M}_{b0} & \boldsymbol{M}_{bb} \end{bmatrix}. \tag{53}$$

Let $\phi_j = \mathbb{E}[\gamma_j] = q(\gamma_j = 1)$. The local statistics of the CDV branch are

$$\begin{aligned} \boldsymbol{A}_{kj}^c &= \boldsymbol{M}_{bb,kj}^c, \\ \boldsymbol{g}_{nkj}^c &= x_{nj}\boldsymbol{v}_{b,kj}^c - \boldsymbol{M}_{b0,kj}^c, \\ c_{nkj}^c &= \bar{\lambda}_{kj}^c x_{nj}^2 - 2x_{nj}v_{0,kj}^c + M_{00,kj}^c, \end{aligned} \tag{54}$$

Algorithm 1 SVStMFM Parameter Learning

Require: Normal training set $\boldsymbol{X}$, number of components K, latent-factor dimension r, prior hyperparameters, convergence threshold ε, and maximum number of iterations $T_{\max}$

Ensure: Variational posterior parameters, variable-structure posterior probabilities $\{\phi_j\}_{j=1}^{D}$, and evidence lower bound $\mathcal{L}(q)$

Initialize the local posterior $q(\boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{\eta})$, the mixture-weight posterior $q(\boldsymbol{\pi})$, the structural posteriors $q(\boldsymbol{\gamma})$ and $q(\rho)$, the component-dependent parameter posteriors $q(\boldsymbol{\theta}_{kj}^{c}, \lambda_{kj}^{c})$, the component-invariant parameter posteriors $q(\boldsymbol{\theta}_{j}^{s}, \lambda_{j}^{s})$, and the degrees of freedom ν
$t \leftarrow 1$
while $t < T_{\max}$ **do**
  1) E-step
  Compute $\mathbb{E}[\lambda]$, $\mathbb{E}[\ln \lambda]$, $\mathbb{E}[\lambda\boldsymbol{\theta}]$, and $\mathbb{E}[\lambda\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}]$ from the current Normal–Gamma posteriors
  for $k = 1, \ldots, K$ **do**
    Compute $\boldsymbol{A}_k$ and its inverse $\boldsymbol{V}_k = \boldsymbol{A}_k^{-1}$ from the current $\{\phi_j\}_{j=1}^{D}$ and model parameters
    for $n = 1, \ldots, N$ **do**
      Compute $\boldsymbol{g}_{nk}$ and c_{nk}, and update the conditional latent-factor posterior $q(\boldsymbol{h}_n \mid \eta_n, z_{nk} = 1)$
      Update the local-scale posterior $q(\eta_n \mid z_{nk} = 1)$ and compute $u_{nk} = \mathbb{E}[\eta_n \mid z_{nk} = 1]$ and $\ell_{nk}^{\eta} = \mathbb{E}[\ln \eta_n \mid z_{nk} = 1]$
      Compute the local posterior moments $\widetilde{\boldsymbol{s}}_{nk}$ and $\widetilde{\boldsymbol{H}}_{nk}$
      Compute the unnormalized log-responsibility ℓ_{nk}
  for $n = 1, \ldots, N$ **do**
    Update the responsibilities $r_{nk} = \exp(\ell_{nk}) / \sum_{l=1}^{K} \exp(\ell_{nl})$
  for $j = 1, \ldots, D$ **do**
    Compute the expected log-likelihoods $\mathcal{E}_{nkj}^{c}$ and $\mathcal{E}_{nkj}^{s}$ for the component-dependent and component-invariant branches
    Update the structural posterior $q(\gamma_j) = \mathrm{Bern}(\gamma_j \mid \phi_j)$, where $\phi_j = [1 + \exp(-\mathrm{logit}(\phi_j))]^{-1}$
  2) M-step
  Compute $N_k = \sum_{n=1}^{N} r_{nk}$ and update the mixture-weight posterior $q(\boldsymbol{\pi})$
  for $k = 1, \ldots, K$ **do**
    for $j = 1, \ldots, D$ **do**
      Update the component-dependent parameter posterior $q(\boldsymbol{\theta}_{kj}^{c}, \lambda_{kj}^{c})$ using r_{nk}, ϕ_j, $\widetilde{\boldsymbol{s}}_{nk}$, and $\widetilde{\boldsymbol{H}}_{nk}$
  for $j = 1, \ldots, D$ **do**
    Update the component-invariant parameter posterior $q(\boldsymbol{\theta}_{j}^{s}, \lambda_{j}^{s})$ using r_{nk}, $1 - \phi_j$, $\widetilde{\boldsymbol{s}}_{nk}$, and $\widetilde{\boldsymbol{H}}_{nk}$
  Update the sparsity-rate posterior $q(\rho)$ using $\{\phi_j\}_{j=1}^{D}$
  Update the common degrees of freedom ν by one-dimensional numerical root finding
  Compute the evidence lower bound $\mathcal{L}^{(t)}(q)$
  if $t > 1$ and $|\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}| / |\mathcal{L}^{(t-1)}| \le \varepsilon$ **then**
    Terminate the iterations
  $t \leftarrow t + 1$
Return the converged variational posterior parameters, variable-structure posterior probabilities $\{\phi_j\}_{j=1}^{D}$, and evidence lower bound $\mathcal{L}(q)$

and those of the CIV branch are

$$\begin{aligned} \boldsymbol{A}_{j}^{s} &= \boldsymbol{M}_{bb,j}^{s}, \\ \boldsymbol{g}_{nj}^{s} &= x_{nj}\boldsymbol{v}_{b,j}^{s} - \boldsymbol{M}_{b0,j}^{s}, \\ c_{nj}^{s} &= \bar{\lambda}_{j}^{s} x_{nj}^{2} - 2x_{nj} v_{0,j}^{s} + M_{00,j}^{s}. \end{aligned} \tag{55}$$

Their structure-posterior-weighted combination is

$$\begin{aligned} \boldsymbol{A}_k &= \boldsymbol{I}_r + \sum_{j=1}^{D} \left[\phi_j \boldsymbol{A}_{kj}^{c} + (1 - \phi_j)\boldsymbol{A}_{j}^{s}\right], \\ \boldsymbol{g}_{nk} &= \sum_{j=1}^{D} \left[\phi_j \boldsymbol{g}_{nkj}^{c} + (1 - \phi_j)\boldsymbol{g}_{nj}^{s}\right], \\ c_{nk} &= \sum_{j=1}^{D} \left[\phi_j c_{nkj}^{c} + (1 - \phi_j) c_{nj}^{s}\right]. \end{aligned} \tag{56}$$

C.2 Posterior of Latent Factors and Local Scales

The variational optimality condition gives

$$\ln q^*(\boldsymbol{h}_n, \eta_n \mid z_{nk} = 1) = \mathbb{E}_{-\{\boldsymbol{h}_n, \eta_n\}} \left[\ln p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\pi}, \rho, \Theta^c, \Theta^s)\right] + \text{const.} \tag{57}$$

After retaining the terms that depend on $\boldsymbol{h}_n$ and η_n, this becomes

$$\begin{aligned} &\ln q^*(\boldsymbol{h}_n, \eta_n \mid z_{nk} = 1) \\ &= \left(\frac{\nu + D + r}{2} - 1\right) \ln \eta_n - \frac{\eta_n}{2} \left[\nu + c_{nk} + \boldsymbol{h}_n^{\mathrm{T}} \boldsymbol{A}_k \boldsymbol{h}_n - 2\boldsymbol{g}_{nk}^{\mathrm{T}} \boldsymbol{h}_n\right] + \text{const.} \end{aligned} \tag{58}$$

Completing the square yields

$$\begin{aligned} &\boldsymbol{h}_n^{\mathrm{T}} \boldsymbol{A}_k \boldsymbol{h}_n - 2\boldsymbol{g}_{nk}^{\mathrm{T}} \boldsymbol{h}_n \\ &\quad = (\boldsymbol{h}_n - \boldsymbol{m}_{nk}^h)^{\mathrm{T}} \boldsymbol{A}_k (\boldsymbol{h}_n - \boldsymbol{m}_{nk}^h) - \boldsymbol{g}_{nk}^{\mathrm{T}} \boldsymbol{A}_k^{-1} \boldsymbol{g}_{nk}, \end{aligned} \tag{59}$$

where

$$\boldsymbol{V}_k = \boldsymbol{A}_k^{-1}, \qquad \boldsymbol{m}_{nk}^h = \boldsymbol{V}_k \boldsymbol{g}_{nk}. \tag{60}$$

Consequently,

$$\begin{aligned} q(\boldsymbol{h}_n \mid \eta_n, z_{nk} = 1) &= \mathcal{N}(\boldsymbol{h}_n \mid \boldsymbol{m}_{nk}^h, (\eta_n \boldsymbol{A}_k)^{-1}), \\ q(\eta_n \mid z_{nk} = 1) &= \mathrm{Gam}(\eta_n \mid a_{nk}^\eta, b_{nk}^\eta), \end{aligned} \tag{61}$$

with

$$a_{nk}^\eta = \frac{\nu + D}{2}, \qquad b_{nk}^\eta = \frac{1}{2} \left[\nu + c_{nk} - \boldsymbol{g}_{nk}^{\mathrm{T}} \boldsymbol{V}_k \boldsymbol{g}_{nk}\right]. \tag{62}$$

Although the joint quadratic form in Eq. (58) contains the latent dimension r, integration over $\boldsymbol{h}_n$ contributes a determinant factor proportional to $\eta_n^{-r/2}$. It cancels the $r/2$ term, so the final Gamma shape is $(\nu + D)/2$, not $(\nu + D + r)/2$.

The required local moments are

$$\begin{aligned} u_{nk} &= \mathbb{E}[\eta_n \mid z_{nk} = 1] = \frac{a_{nk}^\eta}{b_{nk}^\eta}, \\ \ell_{nk}^\eta &= \mathbb{E}[\ln \eta_n \mid z_{nk} = 1] = \psi(a_{nk}^\eta) - \ln b_{nk}^\eta, \\ \mathbb{E}[\eta_n \boldsymbol{h}_n \mid z_{nk} = 1] &= u_{nk} \boldsymbol{m}_{nk}^h, \\ \mathbb{E}[\eta_n \boldsymbol{h}_n \boldsymbol{h}_n^{\mathrm{T}} \mid z_{nk} = 1] &= \boldsymbol{V}_k + u_{nk} \boldsymbol{m}_{nk}^h (\boldsymbol{m}_{nk}^h)^{\mathrm{T}}. \end{aligned} \tag{63}$$

For the categorical factor,

$$\ln r_{nk} \propto \mathbb{E}[\ln \pi_k] + \mathbb{E}[\ln p(\boldsymbol{x}_n, \boldsymbol{h}_n, \eta_n \mid z_{nk} = 1)]. \tag{64}$$

Collecting all normalized local terms gives

$$\begin{aligned} \ell_{nk} = \mathbb{E}[\ln \pi_k] &+ \frac{1}{2} \sum_{j=1}^{D} \left[\phi_j \ell_{\lambda,kj}^c + (1 - \phi_j) \ell_{\lambda,j}^s\right] - \frac{D}{2} \ln(2\pi) \\ &+ \frac{\nu}{2} \ln \frac{\nu}{2} - \ln \Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2} \ln |\boldsymbol{A}_k| + \ln \Gamma(a_{nk}^\eta) - a_{nk}^\eta \ln b_{nk}^\eta, \end{aligned} \tag{65}$$

$$r_{nk} = \frac{\exp(\ell_{nk})}{\sum_{l=1}^{K} \exp(\ell_{nl})}. \tag{66}$$

C.3 Posterior Updates of Model Parameters

Dirichlet–Multinomial conjugacy gives

$$q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}), \quad \alpha_k = \alpha_{0k} + N_k, \quad N_k = \sum_{n=1}^{N} r_{nk}, \tag{67}$$

and $\mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\sum_l \alpha_l)$. For $\widetilde{\boldsymbol{h}}_n = [1, \boldsymbol{h}_n^{\mathrm{T}}]^{\mathrm{T}}$, define

$$\begin{aligned} \widetilde{\boldsymbol{s}}_{nk} &= \mathbb{E}[\eta_n \widetilde{\boldsymbol{h}}_n \mid z_{nk} = 1] = \begin{bmatrix} u_{nk} \\ u_{nk} \boldsymbol{m}_{nk}^h \end{bmatrix}, \\ \widetilde{\boldsymbol{H}}_{nk} &= \mathbb{E}[\eta_n \widetilde{\boldsymbol{h}}_n \widetilde{\boldsymbol{h}}_n^{\mathrm{T}} \mid z_{nk} = 1] = \begin{bmatrix} u_{nk} & u_{nk} (\boldsymbol{m}_{nk}^h)^{\mathrm{T}} \\ u_{nk} \boldsymbol{m}_{nk}^h & \boldsymbol{V}_k + u_{nk} \boldsymbol{m}_{nk}^h (\boldsymbol{m}_{nk}^h)^{\mathrm{T}} \end{bmatrix}. \end{aligned} \tag{68}$$

For either parameter block, the coordinate optimum follows from

$$\ln q^*(\boldsymbol{\theta}, \lambda) = \mathbb{E}_{-\{\boldsymbol{\theta},\lambda\}}[\ln p(\cdot)] + \text{const.} \tag{69}$$

Bayesian linear regression with the Normal–Gamma prior is conjugate. For the CDV branch,

$$q(\boldsymbol{\theta}^c_{kj}, \lambda^c_{kj}) = \mathcal{N}(\boldsymbol{\theta}^c_{kj} \mid \boldsymbol{m}^c_{kj}, (\lambda^c_{kj}\boldsymbol{R}^c_{kj})^{-1})\mathrm{Gam}(\lambda^c_{kj} \mid a^c_{kj}, b^c_{kj}), \tag{70}$$

where

$$\begin{aligned}
\boldsymbol{R}^c_{kj} &= \boldsymbol{R}^c_0 + \phi_j \sum_{n=1}^{N} r_{nk} \widetilde{\boldsymbol{H}}_{nk}, \\
\boldsymbol{m}^c_{kj} &= (\boldsymbol{R}^c_{kj})^{-1} \left[\boldsymbol{R}^c_0 \boldsymbol{m}^c_0 + \phi_j \sum_{n=1}^{N} r_{nk} x_{nj} \widetilde{\boldsymbol{s}}_{nk} \right], \\
a^c_{kj} &= a^c_0 + \frac{\phi_j N_k}{2}, \\
b^c_{kj} &= b^c_0 + \frac{1}{2} \left[\phi_j \sum_{n=1}^{N} r_{nk} u_{nk} x^2_{nj} + (\boldsymbol{m}^c_0)^{\mathrm{T}} \boldsymbol{R}^c_0 \boldsymbol{m}^c_0 - (\boldsymbol{m}^c_{kj})^{\mathrm{T}} \boldsymbol{R}^c_{kj} \boldsymbol{m}^c_{kj} \right].
\end{aligned} \tag{71}$$

For the CIV branch,

$$q(\boldsymbol{\theta}^s_j, \lambda^s_j) = \mathcal{N}(\boldsymbol{\theta}^s_j \mid \boldsymbol{m}^s_j, (\lambda^s_j \boldsymbol{R}^s_j)^{-1})\mathrm{Gam}(\lambda^s_j \mid a^s_j, b^s_j), \tag{72}$$

where

$$\begin{aligned}
\boldsymbol{R}^s_j &= \boldsymbol{R}^s_0 + (1-\phi_j) \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \widetilde{\boldsymbol{H}}_{nk}, \\
\boldsymbol{m}^s_j &= (\boldsymbol{R}^s_j)^{-1} \left[\boldsymbol{R}^s_0 \boldsymbol{m}^s_0 + (1-\phi_j) \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} x_{nj} \widetilde{\boldsymbol{s}}_{nk} \right], \\
a^s_j &= a^s_0 + \frac{N(1-\phi_j)}{2}, \\
b^s_j &= b^s_0 + \frac{1}{2} \left[(1-\phi_j) \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} u_{nk} x^2_{nj} + (\boldsymbol{m}^s_0)^{\mathrm{T}} \boldsymbol{R}^s_0 \boldsymbol{m}^s_0 - (\boldsymbol{m}^s_j)^{\mathrm{T}} \boldsymbol{R}^s_j \boldsymbol{m}^s_j \right].
\end{aligned} \tag{73}$$

Thus the CDV update carries weight ϕ_j, whereas the CIV update carries $1 - \phi_j$; a CIV is shared across components but is not removed.

C.4 Posterior of Feature–Latent-State Dependence

The coordinate optimum for the structural indicator is

$$\ln q^*(\gamma_j) = \mathbb{E}_{-\gamma_j}\left[\ln p(\gamma_j \mid \rho) + \ln p(\boldsymbol{X} \mid \gamma_j, \cdots)\right] + \text{const.} \tag{74}$$

For branch $b \in \{c, s\}$, define the expected weighted residual

$$Q^b_{nkj} = \bar{\lambda}^b x^2_{nj} u_{nk} - 2 x_{nj} (\boldsymbol{v}^b)^{\mathrm{T}} \widetilde{\boldsymbol{s}}_{nk} + \mathrm{tr}(\boldsymbol{M}^b \widetilde{\boldsymbol{H}}_{nk}), \tag{75}$$

where the moments carry indices (k, j) for $b = c$ and index j for $b = s$. Its expected log-likelihood is

$$\mathcal{E}^b_{nkj} = -\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ell^{\eta}_{nk} + \frac{1}{2}\ell^b_{\lambda} - \frac{1}{2} Q^b_{nkj}. \tag{76}$$

Evaluating the optimum at $\gamma_j = 1$ and $\gamma_j = 0$ gives the posterior odds

$$\begin{aligned}
\ln \frac{q(\gamma_j = 1)}{q(\gamma_j = 0)} = \mathrm{logit}(\phi_j) &= \mathbb{E}[\ln \rho] - \mathbb{E}[\ln(1-\rho)] \\
&\quad + \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left(\mathcal{E}^c_{nkj} - \mathcal{E}^s_{nkj} \right),
\end{aligned} \tag{77}$$

$$\phi_j = \left[1 + \exp\{-\mathrm{logit}(\phi_j)\}\right]^{-1}. \tag{78}$$

This update performs a Bayesian evidence comparison between the component-dependent and component-invariant parameterizations; it does not delete the feature.

Beta–Bernoulli conjugacy gives

$$q(\rho) = \mathrm{Beta}(\rho \mid a_\rho, b_\rho),$$
$$a_\rho = a_{\rho 0} + \sum_{j=1}^{D} \phi_j, \qquad b_\rho = b_{\rho 0} + D - \sum_{j=1}^{D} \phi_j, \tag{79}$$

with

$$\mathbb{E}[\ln \rho] = \psi(a_\rho) - \psi(a_\rho + b_\rho), \qquad \mathbb{E}[\ln(1-\rho)] = \psi(b_\rho) - \psi(a_\rho + b_\rho). \tag{80}$$

C.5 Degrees of Freedom, ELBO, and Convergence

The model uses one common degrees-of-freedom parameter ν. Maximizing the ELBO with respect to ν gives

$$\ln \frac{\nu}{2} - \psi\Big(\frac{\nu}{2}\Big) + 1 + \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left(\ell_{nk}^{\eta} - u_{nk}\right) = 0. \tag{81}$$

There is no closed-form solution; we use Brent's method for one-dimensional root finding. We do not introduce component-specific degrees of freedom.

The evidence lower bound is

$$\begin{aligned} \mathcal{L}(q) = {} & \mathbb{E}_q[\ln p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\pi}, \rho, \Theta^c, \Theta^s)] \\ & - \mathbb{E}_q[\ln q(\boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\pi}, \rho, \Theta^c, \Theta^s)]\,. \end{aligned} \tag{82}$$

Using the joint distribution and the structured variational factorization, this bound decomposes as

$$\begin{aligned} \mathcal{L}(q) = {} & \mathbb{E}_q[\ln p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \Theta^c, \Theta^s)] + \mathbb{E}_q[\ln p(\boldsymbol{Z} \mid \boldsymbol{\pi})] \\ & + \mathbb{E}_q[\ln p(\boldsymbol{H} \mid \boldsymbol{\eta})] + \mathbb{E}_q[\ln p(\boldsymbol{\eta} \mid \nu)] + \mathbb{E}_q[\ln p(\boldsymbol{\gamma} \mid \rho)] \\ & + \mathbb{E}_q[\ln p(\boldsymbol{\pi})] + \mathbb{E}_q[\ln p(\rho)] + \mathbb{E}_q[\ln p(\Theta^c)] + \mathbb{E}_q[\ln p(\Theta^s)] \\ & - \mathbb{E}_q[\ln q(\boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{\eta})] - \mathbb{E}_q[\ln q(\boldsymbol{\pi})] - \mathbb{E}_q[\ln q(\rho)] \\ & - \sum_{j=1}^{D} \mathbb{E}_q[\ln q(\gamma_j)] - \sum_{k=1}^{K} \sum_{j=1}^{D} \mathbb{E}_q[\ln q(\boldsymbol{\theta}_{kj}^c, \lambda_{kj}^c)] \\ & - \sum_{j=1}^{D} \mathbb{E}_q[\ln q(\boldsymbol{\theta}_j^s, \lambda_j^s)]. \end{aligned} \tag{83}$$

The first group contains the expected observation likelihood and prior terms, whereas the remaining terms correspond to the entropies of the structured local posterior and global variational factors. Because $q(\boldsymbol{z}_n, \boldsymbol{h}_n, \eta_n)$ preserves local dependence, its entropy is evaluated jointly rather than factorized into separate terms. All quantities admit closed-form evaluation from the posterior moments derived above.

Each coordinate update above maximizes this bound with all other factors fixed. The iterations terminate when

$$\frac{|\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}|}{|\mathcal{L}^{(t-1)}|} \le \varepsilon, \tag{84}$$

or when the prescribed maximum number of iterations is reached.

D Additional Experimental Details

D.1 Tennessee Eastman Process

This section uses the Tennessee Eastman process (TEP) to validate the effectiveness of the proposed SVStMFM fault-monitoring method for non-Gaussian, strongly coupled, and nonlinear industrial processes.

D.1.1 Process and Dataset

Proposed by Downs and Vogel, the TEP is a classical simulation benchmark developed from a practical chemical production process. It consists of five operating units: a reactor, condenser, vapor–liquid separator, recycle compressor, and stripper. Significant material and energy coupling exists

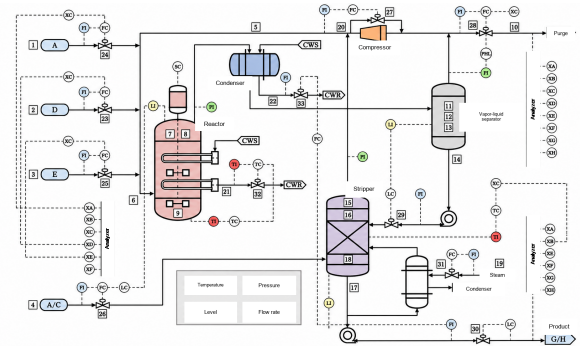


Figure 9: Flow diagram of the TEP

among these units. Consequently, changes in local variables may propagate throughout the production process through closed-loop control and material recirculation. The process flow is shown in Figure 9.

The dataset used in this study contains 52 process variables under six operating conditions and 21 fault types. The process variables comprise 41 measured variables and 11 manipulated variables. The fault types include step changes, random fluctuations, slow drifts, and valve faults. In the experiments, the sampling interval of every sequence is set to $3\,\mathrm{min}$. For each operating mode, $48\,\mathrm{h}$ of normal operation is used as training data. In addition, each state is run for $72\,\mathrm{h}$ to obtain test data, with a fault injected at $60\,\mathrm{h}$.

D.1.2 Hyperparameters, Baselines, and Evaluation Metrics

All experiments use the same hyperparameter settings. The initial number of mixture components is $K_{\mathrm{init}} = 10$, and the convergence threshold for the evidence lower bound is $\delta = 10^{-5}$. The model uses weakly informative priors: $\{\alpha_{0k} = 10^{-3},\ a_{\rho 0} = b_{\rho 0} = 1,\ \beta_0^c = \beta_0^s = 10^{-3},\ \boldsymbol{m}_0^c = \mathbf{0},\ \boldsymbol{m}_0^s = \mathbf{0},\ u_0^c = D_c + 2,\ u_0^s = D_s + 2,\ \boldsymbol{W}_0^c = \boldsymbol{I}/u_0^c,\ \boldsymbol{W}_0^s = \boldsymbol{I}/u_0^s\}$. For online monitoring, the EMA smoothing coefficient is set to $\lambda = 0.2$, and the control limit is determined by the 95% empirical quantile of the normal training statistics.

To validate the fault-monitoring performance of SVStMFM, we select GMM, VBGMM, VBStMM, and DPGMM as comparison methods. GMM uses a Gaussian mixture distribution to describe normal data; VBGMM estimates model parameters through variational Bayesian inference; VBStMM introduces the Student-t distribution to improve robustness to heavy-tailed noise and outlying samples; and DPGMM uses a Dirichlet process prior to infer the effective number of components. All methods model the normal data and determine control limits from empirical quantiles of the training statistics at the same confidence level.

We use the false alarm rate (FAR), fault detection rate (DR), and F1 score to evaluate monitoring performance. Let TP, TN, FP, and FN denote the numbers of correctly detected faulty samples, correctly identified normal samples, false alarms, and missed faulty samples, respectively. Then

$$\mathrm{FAR} = \frac{FP}{FP + TN},\ \mathrm{DR} = \frac{TP}{TP + FN},\ \mathrm{Precision} = \frac{TP}{TP + FP},\ \mathrm{F1} = \frac{2\,\mathrm{Precision} \times \mathrm{DR}}{\mathrm{Precision} + \mathrm{DR}}. \tag{85}$$

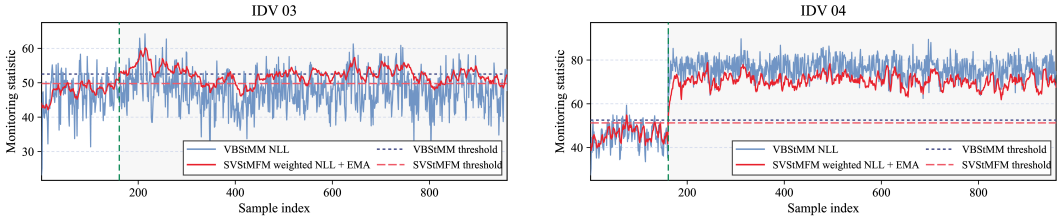
Here, FAR and DR measure the degree of false alarms and the fault detection capability, respectively; Precision measures the accuracy of fault alarms, and F1 jointly reflects the fault detection rate and alarm precision.

D.1.3 Per-Fault Results and Additional Analysis

To separately validate the effectiveness of the heavy-tailed distribution and sparse structure, we analyze comparison methods with similar structures. Both S-VBGMM and SVStMFM include a variable-selection mechanism. As shown in Table 2, the mean DR of S-VBGMM is 0.698, whereas that of SVStMFM reaches 0.939, an increase of 0.241. The advantage of the Student-t distribution is more evident for several difficult-to-detect faults. For example, for faults d03, d09, and d15, the DRs of S-VBGMM are 0.156, 0.135, and 0.184, respectively, whereas those of SVStMFM increase to 0.883, 0.830, and 0.485. The Gaussian distribution is relatively sensitive to observations far from the

center. A small number of outlying samples may shift the estimated component covariance, thereby enlarging the coverage of the normal distribution and causing some weak faults to be absorbed by normal components. The Student-t distribution uses local scale variables to reduce the influence of highly deviating samples on parameter estimation and can therefore obtain a more stable normal distribution. Consequently, SVStMFM is more sensitive to faults with small deviation magnitudes.

The comparison between VBStMM and SVStMFM shows that variable partitioning improves both monitoring performance and model efficiency. VBStMM estimates independent parameters for each Student-t component over all variables, and its mean DR and FAR are 0.837 and 0.102, respectively. After SVStMFM partitions the variables into component-dependent and component-invariant variables, the mean DR increases to 0.939 and the mean FAR decreases to 0.064. Thus, the sparse structure increases the mean DR by 0.102 and decreases the mean FAR by 0.038. This indicates that component-dependent modeling of all variables not only produces redundant parameters but may also allow statistical features shared across operating conditions to interfere with component estimation. The experiments show that variable partitioning does not lose monitoring information by reducing component parameters; instead, it improves the balance between fault detection and false alarm control.



(a) IDV 03 (b) IDV 04

Figure 10: Monitoring results of VBStMM and SVStMFM for IDV 03 and IDV 04

To evaluate the gains in monitoring performance and computational efficiency provided by the sparse structure, we compare the parameter counts, training times, and mean F1 scores of the models, as shown in Figure 6. SVStMFM has 1213 parameters, only approximately 16.9%–21.2% of the numbers in the other models, and its training time decreases to 0.185; meanwhile, its mean F1 score is 0.953, higher than those of the comparison models. This result both preserves the component-dependent information that distinguishes different local data clusters and uses component-invariant variables to characterize stable relationships common to the components, thereby reducing redundant parameters and establishing a structural basis for subsequent fault monitoring based on local responsibilities.

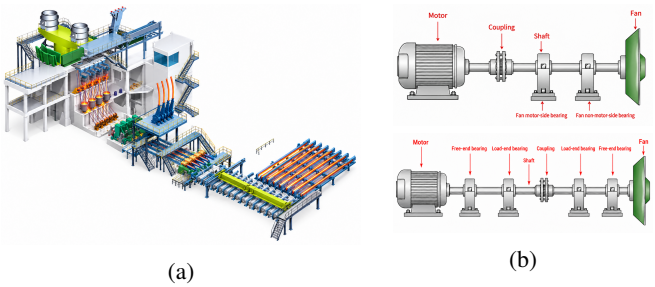
Furthermore, Figure 7 shows the posterior probability estimates of SVStMFM for the 52 process variables. Most probabilities are concentrated near 0 or 1, and only a few variables lie in the intermediate region, indicating that the model can form a relatively clear structural partition. After thresholding, 29 variables are identified as component-dependent variables, and the remaining 23 are treated as component-invariant variables. This result both preserves discriminative local information and uses the component-invariant variables to describe stable relationships across modes, thereby reducing redundant parameters and providing a structural basis for subsequent fault monitoring based on local responsibilities.

Figures 10a and 10b in Figure 10, together with Figures 5a and 5b in Figure 5, show the monitoring statistics and control-limit curves of SVStMFM for four representative faults: IDV 03, IDV 04, IDV 11, and IDV 19. Compared with VBStMM, SVStMFM attenuates instantaneous fluctuations in the statistics through responsibility weighting and EMA accumulation, yielding more stable monitoring curves. For IDV 04, IDV 11, and IDV 19, the SVStMFM statistics rapidly exceed the control limits after fault injection and remain at high levels, with DRs of 1.000, 0.986, and 0.989, respectively, and FARs of 0. For IDV 03, the SVStMFM statistic responds relatively weakly to the fault and has only a small gap from the control limit, but still increases the DR from 0.169 for VBStMM to 0.883. However, the FAR of SVStMFM for this fault reaches 0.263, indicating that its fault characteristics still overlap substantially with normal fluctuations; adaptive control limits may further reduce false alarms in future work.

We further compare SVStMFM with conventional monitoring models, mixture models, and deep models. The results are shown in Table 1 and Figure 4. The DR, Accuracy, F1, and AUC of SVStMFM reach 0.9387, 0.9301, 0.9529, and 0.9731, respectively, all outperforming the other methods. Its ROC curve lies above the other curves over most false alarm rates and maintains a high true positive rate particularly in the low-false-alarm region, indicating that the model is relatively stable under changes in the decision threshold. In addition, the training time of SVStMFM is 0.1851 s. Although this is higher than those of PCA, CVA, and GMM, it is lower than those of MPPCA, VBGMM, DAGMM, and VBStMM, indicating that the sparse variable structure improves monitoring performance while also reducing the computational overhead of model-parameter estimation.

D.2 Real-World SRCC Dataset

To further validate the applicability of the proposed method to real industrial processes, we conduct experiments on a fault-diagnosis dataset for ventilation exhaust fans (VEFs) in the secondary steelmaking, refining, and continuous casting process of a steel enterprise Li et al. (2025). The SRCC process and the two fan structures are illustrated in Figure 11.



(a) (b)

Figure 11: Schematic diagrams of the SRCC process and fan structures: (a) the SRCC process; (b) the cantilever and double-supported fan structures and bearing positions

Table 4: VEF fault types and empirical diagnostic rules in the SRCC process

Fault type	Fault cause	Empirical diagnostic rule
Unbalance	Dust accumulation on blades, assembly errors, blade wear, or detached parts	Identify anomalies by comparing the relationships among the 1×, 2×, and fundamental-frequency components.
Misalignment	Eccentric coupling installation and offset between the rotor and bearing centerlines	The 2× component exceeds 50% of the 1× component; the second harmonic reaches 25% of the threshold, with significant axial vibration; and the maximum axial-to-radial vibration ratio exceeds 50%.
Looseness	Loose foundation bolts, bearing housings, or thrust rings	The amplitude variation in the vertical direction at different end faces exceeds 10%; fractional and odd harmonics exceed 50% of the fundamental-frequency alarm level.
Bearing fault	Anomalies in rolling or sliding bearings	The amplitude of a rolling-bearing frequency component exceeds twice the normal level; sliding bearings exhibit characteristics similar to those of looseness faults.

D.2.1 Process, Equipment, and Data Collection

The secondary steelmaking, refining, and continuous casting (SRCC) process is a key stage in steel manufacturing. It mainly comprises electric-arc-furnace steelmaking, refining, and continuous slab casting, and imposes stringent control requirements on parameters such as temperature, humidity, and cooling rate. Ventilation exhaust fans regulate on-site temperature and humidity by extracting hot air from the casting workshop and are important for maintaining stable operation and product quality in the continuous-casting area. The on-site fans are mainly of two types: cantilever fans and double-supported fans. A cantilever fan has a single-ended support structure, with one end of the rotor shaft fixed to a bearing and the other end cantilevered. A double-supported fan has bearings at both ends of the rotor shaft, providing better balance and stability.

The dataset was collected from five fans in the steelmaking area of a steel enterprise from April 27, 2010, to September 7, 2020, using vibration and temperature sensors. Each measurement point contains monitoring variables including impact peak, mean impact, acceleration peak, acceleration

root mean square, velocity peak, velocity root mean square, displacement peak, and measurement-point temperature. The main fan faults comprise unbalance, misalignment, looseness, and bearing faults. Their causes and empirical diagnostic rules are shown in Table 4.

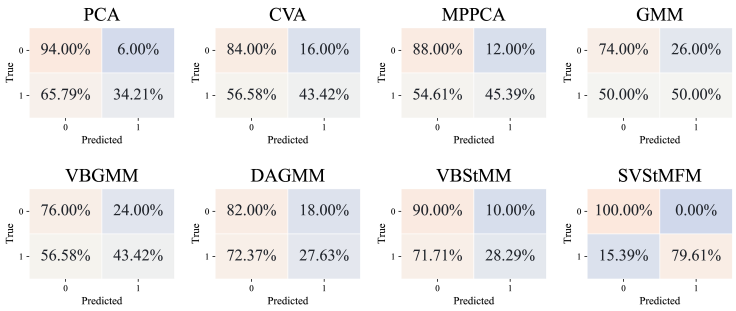


Figure 12: Confusion matrices of different fault-monitoring methods on the SRCC dataset

D.2.2 Additional Results and Analysis

We compare SVStMFM with conventional statistical monitoring methods, mixture models, and deep models on the SRCC fan dataset. The results are shown in Table 3, Figure 12, and Figure 8. On this dataset, SVStMFM has an FAR of 0, while its DR, Accuracy, F1, and AUC reach 0.7961, 0.8465, 0.8864, and 0.9443, respectively, all outperforming the other comparison methods. Compared with GMM and VBStMM, SVStMFM achieves clear improvements in both fault detection rate and overall identification performance, indicating that the sparse variable structure can retain key fault information while attenuating the interference of redundant variables with mixture-component estimation.

The confusion matrices show that the conventional methods miss many faulty samples, and most comparison methods identify fewer than 50% of the faulty samples. In contrast, SVStMFM correctly identifies all normal samples and increases the proportion of correctly identified faulty samples to 79.61%, indicating that its monitoring statistic has stronger discriminative ability for SRCC fan faults. This is because the characteristics of different fault types in the on-site fan vibration and temperature data are somewhat similar, while long-term data collection is susceptible to noise, load fluctuations, and local anomalous observations. The Student-t distribution can reduce the influence of anomalous observations on parameter estimation, while sparse variable selection further emphasizes the key variables related to changes in operating state. In terms of computational efficiency, the training time of SVStMFM is 1.2114 s, which is close to that of VBStMM and substantially lower than that of DAGMM. Therefore, in real SRCC fan fault-diagnosis scenarios, SVStMFM achieves higher detection accuracy and lower false alarm risk with modest computational overhead, validating the applicability of the proposed method to real industrial process monitoring.