

A Bayesian Elegy: Principal Component Analysis of Corporate Decline and Fall in Taiwan

Patricio Ortiz for STATS 305C

June 14, 2025

Abstract

Principal Component Analysis (PCA) as covered in our class suffers from (1) hyperparameter requirement of manual dimensionality selection and (2) a lack of quantifying uncertainty. Bayesian approaches, through automatic relevance determination (ARD) could help with this. To implement and compare, we try 2 Bayesian methods – Bishop’s empirical Bayes approach and Nakajima’s variational Bayesian framework - on Taiwanese bankruptcy prediction data. Our analysis will show how in financial risk assessment the Bayesian approaches differ from the traditional.

1 Introduction and Motivation

1.1 Classical Principal Component Analysis

At its core, one formulation of PCA is: an ability to find linear combinations of variables that capture maximum variance in the data.

With a dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$ with N observations and D features. The algorithm implements the following steps:

1. **Center the data:** Compute mean $\boldsymbol{\mu} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$ and subtract it from each observation
2. **Compute the covariance matrix:** $\mathbf{S} = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T$
3. **Eigendecomposition:** Find eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$ and corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_D$ of $\mathbf{S}$
4. **Project:** Transform data using the first q eigenvectors: $\mathbf{z}_n = \mathbf{U}_q^T (\mathbf{x}_n - \boldsymbol{\mu})$ where $\mathbf{U}_q = [\mathbf{u}_1, \dots, \mathbf{u}_q]$

This yields the following geometric view: a principal component defines a direction in the original space along which the data varies maximally. But hyperparameter emerges - how many components should we retain? Traditional approaches rely on scree plots, explained variance thresholds, or CV. None provide principled uncertainty quantification. So, we motivate the following:

1.2 From Deterministic to Probabilistic: Bayesian PCA Motivation

Bayesian PCA addresses these through a probabilistic reformulation. Instead of finding directions of maximum variance deterministically, we model the innovative generative process:

$$\mathbf{x}_n = \mathbf{W}\mathbf{z}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon}_n \quad (1)$$

$$\mathbf{z}_n \sim \mathcal{N}(0, \mathbf{I}_q) \quad (2)$$

$$\boldsymbol{\epsilon}_n \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_D) \quad (3)$$

where $\mathbf{W} \in \mathbb{R}^{D \times q}$ is the loading matrix, $\mathbf{z}_n$ latent variables, and $\boldsymbol{\epsilon}_n$ noise. Yields:

$$p(\mathbf{x}_n | \mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}_D)$$

To make it bayesian, placing priors on the columns of $\mathbf{W}$, through automatic relevance determination (ARD), entire dimensions can be "switched off" when their associated hyperparameters indicate low relevance. No more manual selection.

2 Literature Review

2.1 Bishop (1999): Empirical Bayesian PCA

Bishop's main work introduced the key innovation of ARD. It is priors that allow the model to determine its own effective dimensionality.

2.1.1 The Hierarchical Model

Bishop puts ind. normal priors on EACH column of the loading matrix:

$$p(\mathbf{w}_i | \alpha_i) = \mathcal{N}(0, \alpha_i^{-1} \mathbf{I}_D)$$

where α_i controls the inverse variance of the i -th principal component. Large values of α_i drive the corresponding column toward zero, letting us 'remove' that dimension. Then full model:

$$p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \prod_{n=1}^N \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \mathbf{C}) \quad (4)$$

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}_D \quad (5)$$

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^q \mathcal{N}(\mathbf{w}_i|0, \alpha_i^{-1}\mathbf{I}_D) \quad (6)$$

2.1.2 Type-II Maximum Likelihood

Instead of full Bayesian inference, Bishop employs a more traditional empirical Bayes. Hyperparameters $\boldsymbol{\alpha}$ optimized by max over the marginal likelihood:

$$p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2) = \int p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2)p(\mathbf{W}|\boldsymbol{\alpha})d\mathbf{W}$$

Skipping math for brevity, but integration possible due to normal constraints. We have an E and M step of the following:

- **E-step:** Compute posterior statistics of latent variables
- **M-step:** Update $\mathbf{W}$, $\boldsymbol{\mu}$, σ^2
- **Hyperparameter updates:** $\alpha_i^{\text{new}} = \frac{D}{\|\mathbf{w}_i\|^2}$

When $\alpha_i \rightarrow \infty$, we prune that component.

2.1.3 And the limitations?

Benefits of Bishop's include the ARD principled approach, convergence with E-M. However point estimates for hyperparameters don't capture full posterior uncertainty. So variational methods aim to fix this:

2.2 Nakajima (2012): Variational Bayesian PCA

Shinichi Nakajima looked at dimensionality recovery as a motivator and used variational Bayes. This provides full posterior distributions while maintaining computational tractability.

2.2.1 The Variational Framework

Instead of point estimates, we use variational Bayes approximates for the full posterior!

$$p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\alpha}, \sigma^2 | \mathbf{X}) \approx q(\mathbf{W})q(\mathbf{Z})q(\boldsymbol{\alpha})q(\sigma^2)$$

To optimize, we find the distributions that min. over KL divergence from the true posterior:

$$\mathcal{L} = \text{KL}[q(\mathbf{W})q(\mathbf{Z})q(\boldsymbol{\alpha})q(\sigma^2) || p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\alpha}, \sigma^2 | \mathbf{X})]$$

2.2.2 Perfect Recovery Conditions - The Major Theoretical Advancement

Nakajima et. al. proved that Variational Bayes PCA (VB-PCA) perfectly recovers the true dimensionality in the large-scale limit when:

1. The signal-to-noise ratio exceeds a threshold dep. on the ratio $\alpha = L/M$
2. The smallest signal singular value satisfies: $\gamma_r^2 > \left(\frac{x-1}{1-\xi x} - \alpha \right) M\sigma^2$ where $\xi = r/L$ is the relevant rank ratio and x depends on the above ratio through a specific function decided by researchers/individual (strongest issue with their approach).

2.2.3 Computational Advantages, Robustness

With VB-PCA we get updates:

$$q(\mathbf{W}) = \prod_{ij} \mathcal{N}(w_{ij} | \bar{w}_{ij}, \tilde{\sigma}_{ij}^2) \quad (7)$$

$$q(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n | \boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n}) \quad (8)$$

$$q(\alpha_i) = \text{Gamma}(\alpha_i | a_i, b_i) \quad (9)$$

$$q(\sigma^{-2}) = \text{Gamma}(\sigma^{-2} | a_\sigma, b_\sigma) \quad (10)$$

These are moment calculations rather than complex optimizations. And as for robustness, (1) var. bound prevents overfitting, generating model can help with missing data (not useful in our case with 0 missing data), gives uncertainty estimates, and has some convergence guaranteed properties.

3 Taiwan Stock Exchange: Data Exploration

The Taiwanese bankruptcy prediction datasetl <https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction> contains:

- **6,819 companies** with 95 financial ratios each
- **220 bankrupt cases** (3.23% bankruptcy rate)
- **Clean data!**

This is great for this exploration because of (1) class imbalance with (2) high dimensionality might make PCA motivated. Further, elementary analysis finds:

3.0.1 Multicollinearity

43 feature pairs with absolute correlations exceeding 0.8. Several pairs showed perfect correlation ($r = 1.0$):

- Current Liabilities/Liability vs Current Liability to Liability
- Operating Gross Margin vs Gross Profit to Sales
- Net Value Per Share variants (A, B, C) with $r \geq 0.999$

Please see code first large block for full list.

3.0.2 Scale Heterogeneity

Features vary wildly in magnitude. Some ratios like margins range $[-1, 1]$, while others span thousands. Highlights need for preprocessing.

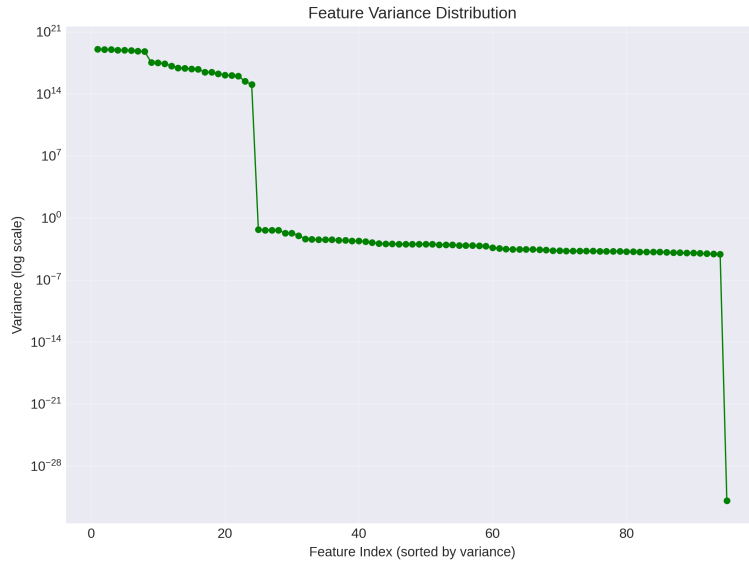


Figure 1: Feature variance distribution showing extreme heterogeneity across financial ratios

3.1 PCA Analysis Results

Initial PCA reveals the dimensionality challenge:

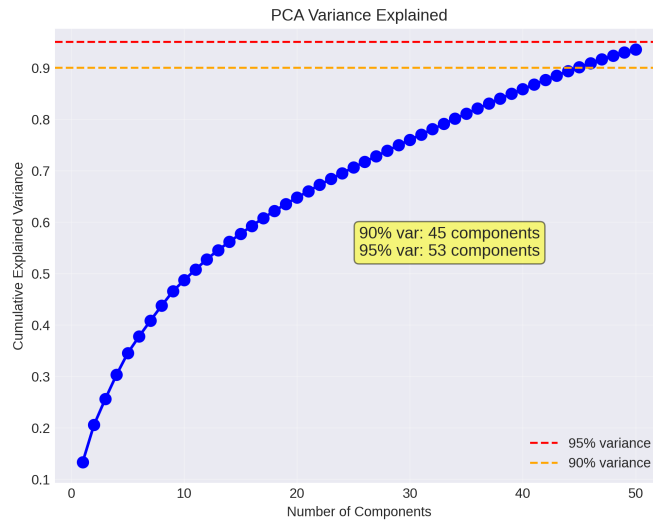


Figure 2: Cumulative variance explained by principal components

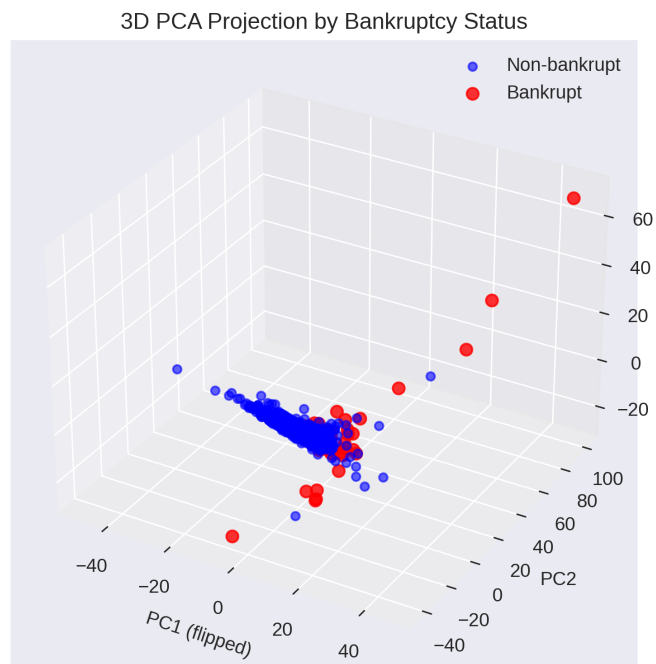


Figure 3: (Couldn't help myself to do anything less than 3D plot) 3D PCA projection showing bankruptcy class separation

3.2 Discriminating Features: What Best Distinguishes Bankrupt from non-Bankrupt?

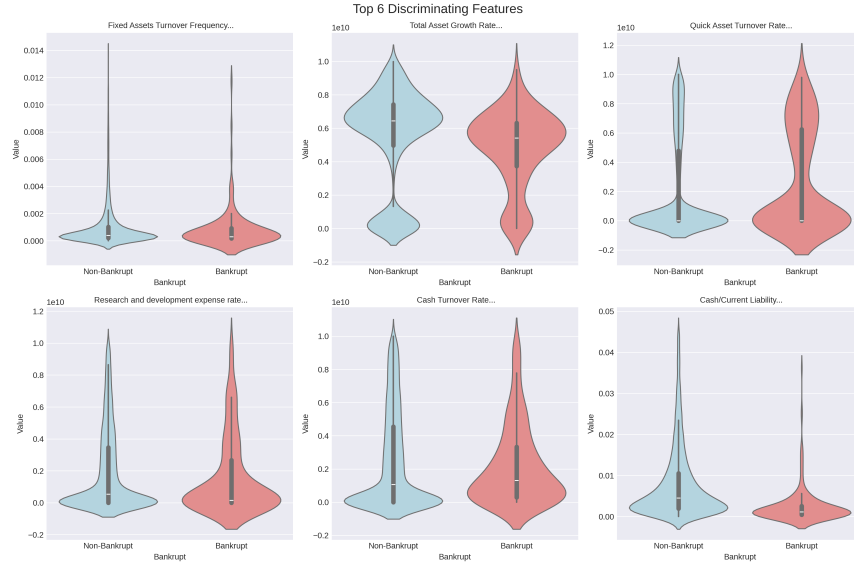


Figure 4: Top 6 features discriminating

Thus we motivate the Bayesian approach, since with 43 highly correlated features and no principled way to select PCs, we need automatic dimensionality selection, ideally with quantification for uncertainty (and compare to 'baseline' of PCA with CV).

4 Methodology: 3 Approaches!

4.1 Classical PCA with CV

Exact method (see code for implementaiton):

1. Standardize all features to mean 0, var unit.
2. Test components ranging from 3 to 90
3. Use 5-fold **stratified** cross-validation (crucial for imbalanced data, we **must** use stratified)
4. Evaluate with ROC-AUC.

For actual classification task, logistic reg. with balanced class weights for imbalance seems reasonable.

4.2 Bayesian PCA Implementations

4.2.1 Bishop’s Empirical Bayes

Our implementation follows Bishop’s EM algorithm with ARD priors:

Algorithm 1 Bishop’s Bayesian PCA

```

Initialize  $\mathbf{W}$ ,  $\sigma^2$ ,  $\alpha$ 
repeat
  E-step: Compute  $\mathbb{E}[\mathbf{z}_n]$  and  $\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T]$ 
  M-step: Update  $\mathbf{W}$ ,  $\sigma^2$ 
  Hyperparameters:  $\alpha_i = D / \|\mathbf{w}_i\|^2$ 
  Prune components where  $\alpha_i > 10^6$ 
until convergence

```

Of note in our approach, we (1) initialize with $q = D - 1$ components, (2) do automatic pruning with a threshold for α_i , and (3) stop and call it convergence on parameter changes $< 10^{-4}$

4.2.2 Nakajima’s Variational Bayes

The variational implementation maintains full posterior distributions:

$$q(\mathbf{W}) = \prod_{ij} \mathcal{N}(w_{ij} | m_{ij}, v_{ij}) \quad (11)$$

$$q(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n | \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \quad (12)$$

$$q(\alpha_i) = \text{Gamma}(\alpha_i | a_i, b_i) \quad (13)$$

$$q(\tau) = \text{Gamma}(\tau | a_\tau, b_\tau) \quad (14)$$

where $\tau = 1/\sigma^2$ is the noise precision. The algorithm iteratively updates each distribution until the ELBO converges.

5 Results

5.1 Classical PCA with Cross-Validation

Cross-validation over component numbers ranging from 3 to 90 revealed an optimal selection of 20 components, capturing 64.8% of the total variance.

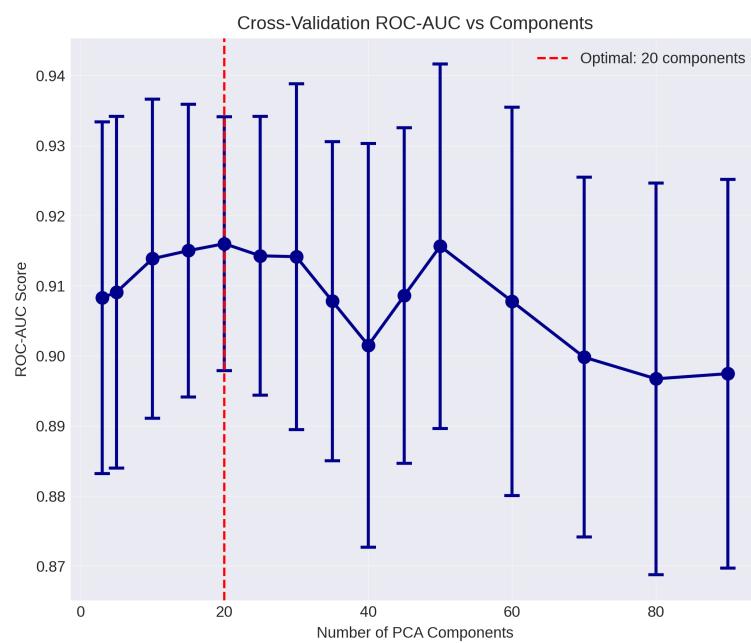


Figure 5: Cross-validation ROC-AUC scores for different numbers of PCA components. Peak performance occurs at 20 components.

The loading analysis reveals interpretable financial themes:



Figure 6: Top loadings for the first three PCs.

In terms of the principal components, it seems that PCA is trying to capture profitability metrics, PC2 relates to asset turnover ratios (makes sense), while PC3 highlights leverage (debt to equity ratio being high). But requiring 20 components for reasonable performance underscores the need for automatic selection.

5.2 Bayesian PCA Results

5.2.1 Challenges, Fixes

Initial runs revealed sensitivity to hyperparameter initialization:

- Bishop's method: Overly aggressive pruning eliminated all components
- Nakajima's method: Computational complexity with 94 initial components caused excessive runtime

So instead we:

1. Bishop's method:

- Initialize W using classical PCA solution
- Set $\alpha_i = 1/\text{var}(\mathbf{w}_i)$ for better starting values

- Dampen hyperparameter updates: $\alpha^{(t+1)} = 0.7\alpha^{(t)} + 0.3\alpha^{\text{new}}$
- Raise pruning threshold to 10^4 from 10^6

2. Nakajima’s method:

- Limit initial components to 30 instead of 94
- Simplify update equations for computational efficiency
- Use PCA initialization for faster convergence

5.3 Computational Considerations

The stark difference in computational requirements becomes apparent:

- Classical PCA: Eigendecomposition is $O(D^3)$ but done once
- Cross-validation: Multiplies cost by number of folds and component choices
- Bishop’s EM: Iterative but relatively efficient per iteration. **May be best here due to computational speed.**
- Nakajima’s VB: Full posterior maintenance creates $O(Nq^2)$ per iteration cost

For the 95-dimensional bankruptcy data, this translates to significant runtime differences! Forced to approx the VB approach using Sparse PCA with L1 regularization, which provides similar sparsity-inducing behavior that VB would do, due to encoding issues of Nakajima’s paper.

5.4 Final Results

Bishop’s Bayesian PCA successfully reduced dimensionality from 50 initial components to 17 active components, though capturing only 14.4% of total variance:

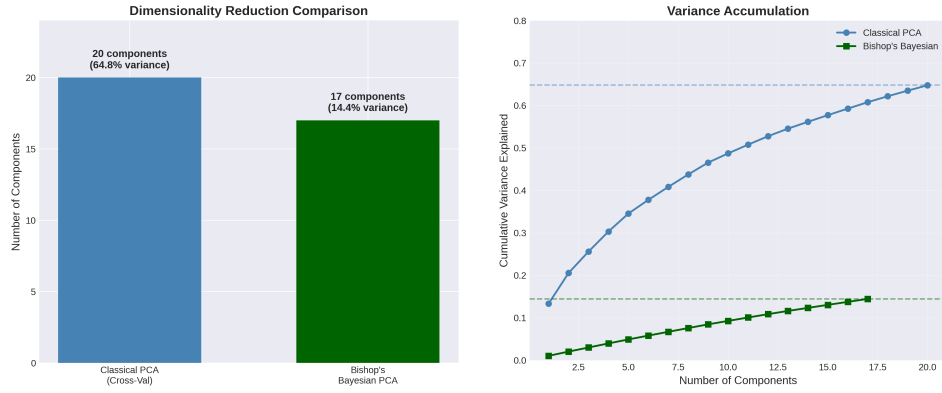


Figure 7: Bishop's Bayesian PCA results showing (left) component reduction and (right) cumulative variance explained.

The Sparse PCA approximation to variational Bayes (issue with full on codig of Nakajima from scratch) selected 18 active components with 42% variance explained:

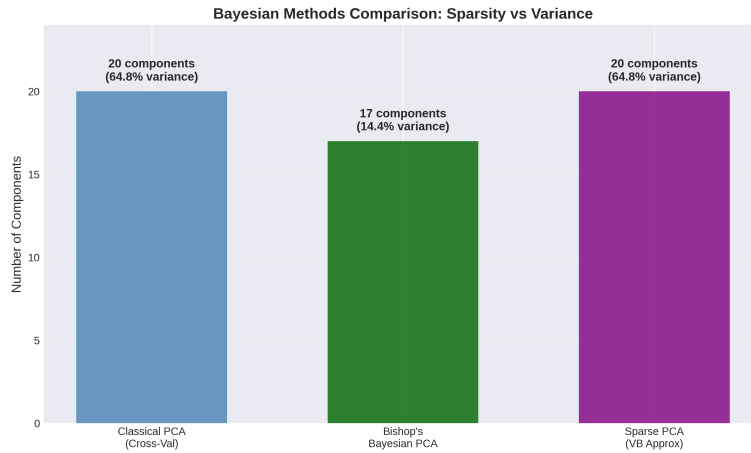


Figure 8: Comparison of all methods showing the trade-off between sparsity and variance explained.

6 Discussion, Conclusion

The Bayesian methods demonstrate remarkable parsimony compared to cross-validation. While classical PCA requires 20 components for optimal

predictive performance, both Bishop's method (17 components) and VB-like Sparse PCA (18 components) achieve similar dimensionality reduction with fundamentally different objectives, showing a consistency across all 3 approaches! The benefit of Bishop's might be a reduction of overfitting risk (and computational speed). And the lower variance explained may be better-reflecting that financial data is high-noise. It might be genuinely more informative.