Augmenting Bayesian Networks with Categorical Principal Components

Ashley Bojko BOJKO@SMOULASOUTEZ.CZ
Nový PORG High School
Prague, Czechia

Marie Melínová MELM13@VSE.CZ
Faculty of Informatics and Statistics
Prague University of Economics and Business
Prague, Czechia

Jiří Vomlel* VOMLEL@UTIA.CAS.CZ
Institute of Information Theory and Automation (ÚTIA)
Czech Academy of Sciences
Prague, Czechia
Institute of Information Studies and Librarianship
Faculty of Arts, Charles University
Prague, Czechia

Editors: Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

Abstract

We propose a method for learning interpretable Bayesian networks (BNs) over high-dimensional categorical data by augmenting the original variable set with discrete components derived from expert-defined groups of variables. Because the components are computed deterministically and entered as fully observed columns, the augmented dataset defines a BN in a curved exponential family, making standard BIC asymptotically valid for joint model scoring. To enable a fair comparison of models with different numbers of component nodes — which induce different model classes over the original variables — we introduce a modified BIC that evaluates marginal likelihood over the original variables while retaining the full parameter penalty. We show that this criterion is conservative relative to the true Geiger-Heckerman effective dimension and establish when the marginal and joint criteria agree asymptotically. We compare PCA-derived components (applied to ordinally coded variables) with Multiple Correspondence Analysis (MCA), the discrete analog of PCA, and show theoretically and empirically that MCA is the more natural choice. Experiments on the Czech Attitude Barometer dataset demonstrate that MCA models achieve the best marginal BIC as well as the best held-out marginal log-likelihood across all augmented and non-augmented models, yielding a sparse, stable, and sociologically interpretable network.

1. Introduction

Bayesian networks (BNs) are valued in applied sciences for combining probabilistic rigor with graphical interpretability. However, as the number of variables grows, two problems arise simultaneously: structure learning becomes computationally intractable – the space of DAGs is super-exponential in the number of nodes – and learned networks tend to be

* Corresponding author.

© 2026 A. Bojko, M. Melínová & J. Vomlel.

dense and difficult to interpret. We address this problem for *discrete BNs over categorical variables*, a setting common in social-science and epidemiological studies where assessments produce nominal or ordinal categories. Here all variables are discrete and all CPTs are multinomial.

The literature offers several families of remedies. *Regularized structure learning* (Friedman and Koller, 2003; Tsamardinos et al., 2006) scales better than exhaustive search but leaves the variable set intact, while *variable selection* (Aliferis et al., 2010) reduces the node set at the cost of discarding informative within-group structure. Latent variable models introduce a discrete latent class node $L^{(g)}$ as the sole representative of each group (Lauritzen, 1995; Zhang, 2004; Meilă and Jordan, 2001); in the continuous setting, *Factor Analysis* (Spearman, 1904; Thurstone, 1935; Tipping and Bishop, 1999), *Structural Equation Models* (SEM) (Jöreskog, 1970), and *BN structure learning over latent variables* (Silva et al., 2006; Elidan et al., 2001) pursue related goals. These approaches require EM (Dempster et al., 1977), face identifiability issues, and incur SE-family BIC complications (Section 4); SEM further learns the inter-factor structure separately from the measurement model. Section 4.2 argues these difficulties are theoretical rather than incidental, and why they preclude a direct empirical comparison with our approach.

We avoid these difficulties by computing group summaries *deterministically* and adding them as fully observed columns alongside the original variables. Because the augmented dataset is fully observed, the model belongs to a curved exponential family (Geiger et al., 2001), so standard BIC applies and no EM is required. The structure over both original and component nodes is learned simultaneously in one unified model, with no prescribed separation between measurement and structural levels; the structure-learning algorithm decides freely how each component node is used (Remark 3 gives a precise taxonomy).

The choice of extraction method matters, and the discrete setting constrains it. PCA treats variables as continuous: applied to ordinally coded categorical variables, it produces continuous scores that must be discretized by an arbitrary binning step unrelated to the categorical geometry of the data, introducing information loss and artificial structure. Multiple Correspondence Analysis (MCA) (Greenacre, 1984), the discrete analog of PCA, instead operates directly on the χ^2 metric native to categorical variables and yields discrete components without an intermediate continuous score, making it the principled choice here.

To compare models with different numbers of component nodes fairly, we introduce a modified BIC that evaluates marginal likelihood over the original variables while retaining the full parameter penalty. We show this criterion is conservative relative to the true Geiger-Heckerman effective dimension, characterize the source and direction of the bias, and establish when the marginal and joint criteria agree asymptotically. Experiments on the Czech Attitude Barometer dataset (Kudrnáč et al., 2024) show that MC-2 – two MCA components per expert group – achieves the best held-out predictive log-likelihood among all augmented and non-augmented models, yielding a sparse, stable, and sociologically interpretable network.

2. Background

2.1. Bayesian networks and BIC

A Bayesian network (BN) is a pair (G, θ) where G is a DAG over discrete variables, and $V = \{X_1, \ldots, X_p\}$ and θ are the conditional probability table (CPT) parameters. For node X_i with r_i states and q_i parent configurations, the CPT contributes $(r_i - 1)q_i$ free parameters; the total is $d(G) = \sum_i (r_i - 1)q_i$. For a fully observed i.i.d. sample D of size N, the log-likelihood $\ell(G, \theta; D)$ of data D given the BN model (specified by G and θ) factorizes over nodes, and the maximum likelihood estimate (MLE) is $\hat{\theta}_{ijk} = N_{ijk}/N_{ij}$, where N_{ijk} counts co-occurrences of $X_i = k$ with parent configuration j. The BIC score is

$$\mathrm{BIC}(G, D) \quad = \quad \ell(G, \hat{\theta}; D) - \frac{d(G)}{2} \log N \ . \tag{1}$$

Its theoretical justification rests on the Laplace approximation to the log marginal likelihood, which is valid when the model is a *curved exponential (CE) family* (Haughton, 1988). Geiger et al. (2001) established that discrete BNs over fully observed data form a CE family, making BIC valid for structure selection in that setting.

2.2. BIC for networks with hidden variables

When a BN contains genuinely hidden nodes, the map from all network parameters θ to the marginal distribution over observed variables need not be injective (e.g., label-switching in mixture models), causing the Jacobian rank to fall below $d(G)$. Geiger et al. (1996) showed that the correct BIC penalty in this case replaces $d(G)/2$ with $\dim(G, \hat{\theta})/2$, the Jacobian rank of the map from θ to observed variables. Rusakov and Geiger (2005) further showed that for models in a *stratified exponential (SE) family* — which includes all BNs with genuinely unobserved nodes — the correct asymptotic expansion of the log marginal likelihood involves a rational coefficient of $\log N$ and an additional $\log \log N$ term, so standard BIC is not even asymptotically correct up to $O(1)$. A key question for our proposal is whether augmented BNs fall into the CE or SE category.

3. The Augmented Bayesian Network

Definition 1 (Partition and components) *Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector of categorical variables partitioned by domain experts into G disjoint groups $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(G)}$. For each group g and each $l = 1, \ldots, m_g$, let $f_l^{(g)} : \mathcal{X}^{(g)} \to \{1, \ldots, k_l\}$ be a* deterministic *map from group observations to a finite set of discrete states, computed from the training data before any BN is fitted. The l-th* discrete component of group g *is*

$$Z_l^{(g)} = f_l^{(g)}(\mathbf{X}^{(g)}).$$

We write $\mathbf{Z} = (Z_l^{(g)})_{g,l}$ for the full vector of $M = \sum_g m_g$ components.

Definition 2 (Augmented dataset and augmented BN) *The* augmented dataset *is $D^+ = \{(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N$ where $\mathbf{z}^{(i)} = (f_l^{(g)}(\mathbf{x}^{(g,i)}))_{g,l}$ is computed deterministically from $\mathbf{x}^{(i)}$. The dataset is* fully observed. *An* augmented BN *is a BN (G^+, θ^+) with node set*

$V^+ = \{X_1, \ldots, X_p\} \cup \{Z_l^{(g)}\}_{g,l}$, *whose structure* G^+ *is learned by a score-based algorithm on* D^+. *The nodes* $\{Z_l^{(g)}\}_{g,l}$ *are called* component nodes *to distinguish them from the original observed nodes* $\{X_1, \ldots, X_p\}$. *The BIC score* $\mathrm{BIC}(G^+, D^+)$ *of the augmented BN with respect to the augmented dataset* D^+ *will be denoted as* $\mathrm{BIC}_{\mathrm{joint}}$.

Remark 3 (Structural roles of component nodes) *Because no edge constraint is imposed on component nodes, the learned* G^+ *can assign each* $Z_l^{(g)}$ *one of four structural roles, determined by whether it has children and by the origin of its parents:*

(1) leaf*:* $Z_l^{(g)}$ *has no children, regardless of its parent source; its CPT parameters cancel from* $p_{\hat{\theta}^+}(\mathbf{X})$ *through normalization, contributing zero to the effective dimension.*

(2) hub*:* $Z_l^{(g)}$ *has parents only in* $\mathbf{X}^{(g)}$ *and children outside the group; it compresses within-group dependence and routes it to the inter-group structure, which is the intended role.*

(3) source*:* $Z_l^{(g)}$ *has no parents but has children among observed variables; when marginalized out, it induces a mixture distribution over its children, which is the role most responsible for the large gap between* ℓ_{marg} *and* ℓ_{joint} *observed experimentally.*

(4) mediator*:* $Z_l^{(g)}$ *has at least one parent from outside* $\mathbf{X}^{(g)}$ *and has children elsewhere; it acts as an intermediate node in an inter-group pathway rather than as a group summary.*

These four roles partition all meaningful configurations: the only excluded case, no parents and no children, would yield a disconnected node. The method does not prescribe which role a component takes; the structure-learning algorithm decides based on the score.

3.1. PCA-derived components

For ordinally coded categorical variables, one may treat each $X_j^{(g)}$ as approximately continuous, run standard PCA on the group, and discretize the continuous scores. The first component $s_1^{(g)}(\mathbf{x}^{(g)}) = (\mathbf{w}_1^{(g)})^\top \mathbf{x}^{(g)}$ is discretized via equal-frequency binning into k bins with boundaries fixed on the training set:

$$Z_1^{(g)} = c_1^{(g)}(s_1^{(g)}(\mathbf{X}^{(g)})).$$

The map $f_1^{(g)} = c_1^{(g)} \circ s_1^{(g)}$ satisfies Definition 1. However, the intermediate continuous score introduces a discretization step whose bin boundaries are arbitrary (though data-driven), which can create an artificial categorical structure.

3.2. MCA-derived components

Multiple Correspondence Analysis (MCA) (Greenacre, 1984) is the natural dimensionality reduction method for categorical data. It operates directly on the *indicator matrix* $\mathbf{Z}_{\mathrm{ind}} \in \{0,1\}^{N \times K}$, where $K = \sum_j r_j$ is the total number of categories; each row of $\mathbf{Z}_{\mathrm{ind}}$ represents one observation as a concatenation of one-hot encoded vectors[1], one for each variable. It

1. One-hot encoded vectors are a way to represent categorical data as binary vectors, where exactly one element is 1 and all others are 0.

performs eigenvalue decomposition of the centered and scaled Burt matrix $\mathbf{B} = \mathbf{Z}_{\text{ind}}^{\top}\mathbf{Z}_{\text{ind}}$. The resulting components use the χ^2 distance, which treats categories according to their co-occurrence patterns rather than their label values.

MCA has two advantages over PCA for this application. First, no discretisation step is needed: the component scores are partitioned according to the χ^2 geometry of the category space via k-means clustering of category centroids, yielding a discrete component without an arbitrary binning rule. Second, the χ^2 metric is coherent with BIC: the eigenvalues of the Burt matrix are directly related to the χ^2 statistics of the within-group contingency tables, so the explained inertia of the first m MCA components quantifies the fraction of within-group categorical association captured, in the same units as the BIC score. The raw MCA eigenvalues are known to understate explained variance; we apply the Benzécri (1979) correction before selecting the number of components. The full extraction procedure is given in Algorithm 1.

4. Theoretical Analysis

4.1. CE family membership and validity of $\text{BIC}_{\text{joint}}$

Lemma 4 (CE family membership) *The augmented BN (G^+, θ^+) fitted by maximum likelihood on the fully observed dataset D^+ belongs to a curved exponential family. Consequently, the Jacobian of $\theta^+ \mapsto p_{\theta^+}(\mathbf{X}, \mathbf{Z})$ has rank $d(G^+)$ at the MLE, and $\text{BIC}_{\text{joint}}(G^+, D^+)$ is an asymptotically valid model-selection criterion.*

Proof D^+ is fully observed by construction (Definition 2): every component value $Z_l^{(g)} = f_l^{(g)}(\mathbf{x}^{(g,i)})$ is computed deterministically for each observation. A discrete BN over fully observed data is a CE family (Geiger et al., 2001): the sufficient statistics are cell counts N_{ijk}, the MLE is unique and interior, and the Fisher information is non-singular. BIC validity then follows from Haughton (1988). ■

The Rusakov-Geiger SE-family pathologies — rational BIC coefficient and $\log\log N$ term — arise when parameters are estimated by maximising a *marginal* likelihood with hidden nodes. In our method, structure and parameters are estimated on the fully observed D^+ via $\text{BIC}_{\text{joint}}$; the SE-family concern is absent from the learning step.

4.2. Why latent class models and SEM are not empirical baselines

The difficulties noted in the Introduction for latent variable approaches are not incidental but follow directly from marginalizing over an unobserved variable. For a K-class LCA, the *complete-data* likelihood factorizes into per-class multinomials and is itself exponential family; it is marginalizing over the unobserved class indicator that replaces this concave, log-linear structure with a non-concave $\log\sum_c(\cdot)$ mixture likelihood. This is the documented cause of EM's sensitivity to initialization for LCA (McLachlan and Peel, 2000), and it also invalidates the regularity conditions underlying standard BIC asymptotics for selecting the number of classes: under the k-class null, the added mixing proportion of a $(k{+}1)$-class model lies on the boundary of the parameter space (Lindsay, 1995). Our fully observed, deterministic-component construction (Section 3) avoids both failure modes by design, since no marginalization over an unobserved variable is required.

Algorithm 1: MCA-Based Discrete Component Extraction

Input: Data $\mathbf{W}$; variable groups $\mathcal{G} = \{G_1, \ldots, G_m\}$; variance threshold τ; $k_{\max}$; $n_{\min}$; balance ratio ρ; homogeneity threshold η

Output: Augmented data $\mathbf{W}^+ = [\mathbf{W} \mid \mathbf{M}]$

`// Benzécri-corrected dimension selection`

Function `GetNcp(`*MCA*, τ`):`
- $Q \leftarrow$ number of variables in *MCA*;
- $\tilde{\lambda}_l \leftarrow \left(\frac{Q}{Q-1}\right)^2 (\lambda_l - \frac{1}{Q})^2$ if $\lambda_l > \frac{1}{Q}$, else 0;
- $\tilde{V}_l \leftarrow \sum_{j=1}^{l} \tilde{\lambda}_j \,/\, \sum_j \tilde{\lambda}_j$;
- **return** $\min\{l : \tilde{V}_l \geq \tau\}$;

`// Voronoi discretization per dimension`

Function `VoronoiDiscretize(`$\mathbf{Z}$, $\mathbf{C}$, $\mathbf{k}$`):`
- **for** *each dimension* d **do**
  - Cluster category coordinates $\mathbf{C}_{:,d}$ into k_d groups via k-means;
  - Order cluster labels by centroid value;
  - Assign each observation to the label of its nearest category centroid;
- **return** discrete component matrix $\mathbf{M}$;

`// Main loop over variable groups`

$\mathbf{W}^+ \leftarrow \mathbf{W}$;

for *each group* $G_i \in \mathcal{G}$ **do**
- $\mathbf{W}_i \leftarrow$ columns of $\mathbf{W}$ in G_i;
- Run MCA on $\mathbf{W}_i$ with $d_{\max} = 20$;
- $d^* \leftarrow$ `GetNcp(`*MCA*, τ`)`;
- Rerun MCA with $ncp = d^*$; extract individual coords $\mathbf{Z}$ and category coords $\mathbf{C}$;
- `// Automated level selection`
- **for** *each dimension* $d = 1, \ldots, d^*$ **do**
  - Cluster $\mathbf{C}_{:,d}$ into $k_{\max}$ groups; assign observations;
  - $n_{\min}^{(d)} \leftarrow \min_k |\text{level } k|$; $\rho^{(d)} \leftarrow n_{\min}^{(d)} / \max_k |\text{level } k|$;
  - **if** $n_{\min}^{(d)} < n_{\min}$ ***or*** $\rho^{(d)} < \rho$ **then**
    - $h \leftarrow$ frequency of most common pattern in smallest level;
    - **if** $h \geq \eta$ **then**
      - $k_d \leftarrow k_{\max}$; `// meaningful subgroup`
    - **else**
      - $k_d \leftarrow k_{\max} - 1$; `// heterogeneous artifact`
  - **else**
    - $k_d \leftarrow k_{\max}$
- $\mathbf{M}_i \leftarrow$ `VoronoiDiscretize(`$\mathbf{Z}$, $\mathbf{C}$, $(k_1, \ldots, k_{d^*})$`)`;
- $\mathbf{W}^+ \leftarrow [\mathbf{W}^+ \mid \mathbf{M}_i]$;

Ordinal/categorical SEM based on polychoric correlations (Jöreskog, 1994) avoids the discrete latent class, but retains a different parametric commitment: each observed Likert item is treated as a threshold discretization of a linear-Gaussian latent response variable, and the latent factors are assumed linearly related – assumptions our discrete BN's structure-learning step does not impose. SEM's covariance- or polychoric-correlation-based fit function also operates on a different likelihood scale than the multinomial complete-data likelihood underlying a discrete BN, so a shared quantitative comparison (e.g., a common BIC table) is not available without a separate methodological bridge; SEM further fixes the measurement model in advance and estimates the inter-factor structure separately, in contrast to our jointly learned structure over original and component nodes. For these reasons we treat both LCA and SEM as related work rather than as empirical baselines.

4.3. Modified BIC for cross-class comparison

Models with zero, one, two, or more components per group induce different model classes over the original variables $\mathbf{X}$: a plain BN (no components), a structured mixture (components as parents of $\mathbf{X}$ nodes), or hybrids. To compare them on a common basis we evaluate all models on their implied marginal likelihood of $\mathbf{X}$.

Definition 5 (Modified BIC)

$$\mathrm{BIC}_{\mathrm{marg}}(G^+, D) \;=\; \ell_{\mathrm{marg}}(G^+, \hat{\theta}^+; D) - \frac{d(G^+)}{2} \log N \;, \tag{2}$$

where $\ell_{\mathrm{marg}} = \sum_{i=1}^{N} \log \sum_{\mathbf{z}} p_{\hat{\theta}^+}(\mathbf{x}^{(i)}, \mathbf{z})$ *and* $d(G^+)$ *counts* all *CPT parameters of the augmented network, including those of the component nodes. For a plain BN (no components),* $\mathrm{BIC}_{\mathrm{marg}} = \mathrm{BIC}_{\mathrm{joint}}$.

Proposition 6 (Conservative penalisation) *The true Geiger-Heckerman effective dimension for the marginal map* $\theta^+ \mapsto p_{\theta^+}(\mathbf{X})$ *satisfies* $\dim(G^+, \hat{\theta}^+) \leq d(G^+)$*, with the gap equal to at least the total parameter count of component nodes whose structural role is* leaf *(Remark 3).*

Proof For a leaf component $Z_l^{(g)}$ — one with no children among other nodes — the CPT parameters $\theta_{Z_l^{(g)}|\mathrm{pa}(Z_l^{(g)})}$ enter $p_{\theta^+}(\mathbf{X})$ only through the sum $\sum_z p_{\theta^+}(Z_l^{(g)} = z \mid \mathrm{pa}(Z_l^{(g)})) = 1$, which is identically one for all θ^+. These parameters therefore contribute zero to the Jacobian rank of $\theta^+ \mapsto p_{\theta^+}(\mathbf{X})$, giving a gap of $(k_l - 1)q_{Z_l^{(g)}}$ per leaf component. ■

The conservatism is a deliberate choice: computing the exact $\dim(G^+, \hat{\theta}^+)$ requires evaluating the Jacobian rank at the MLE, which depends on the data and the learned topology. Using $d(G^+)$ avoids this computation while biasing model selection toward *fewer* component nodes — a desirable property for interpretability.

Theorem 7 (Asymptotic equivalence) *Suppose the true distribution* P^* *is consistent with the deterministic component structure, i.e.,* $P^*(Z_l^{(g)} \neq f_l^{(g)}(\mathbf{X}^{(g)})) = 0$ *for all* g, l*. Then*

$$\frac{1}{N}\left[\ell_{\mathrm{joint}}(G^+, \hat{\theta}^+; D^+) - \ell_{\mathrm{marg}}(G^+, \hat{\theta}^+; D)\right] \xrightarrow{P} 0 \quad \text{as } N \to \infty,$$

so $\frac{1}{N}[\mathrm{BIC}_{\mathrm{joint}}(G^+, D^+) - \mathrm{BIC}_{\mathrm{marg}}(G^+, D)] \xrightarrow{P} 0$.

Proof For each observation $\mathbf{x}^{(i)} \in D$ let $\delta_i = \log p_{\hat{\theta}^+}(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) - \log \sum_{\mathbf{z}} p_{\hat{\theta}^+}(\mathbf{x}^{(i)}, \mathbf{z}) = \log(1 - \varepsilon_i)$ where $\varepsilon_i = P_{\hat{\theta}^+}(\mathbf{Z} \neq \mathbf{z}^{(i)} \mid \mathbf{X} = \mathbf{x}^{(i)})$. By consistency of the MLE (Wald, 1949) and $P^*(\varepsilon_i > 0) = 0$, we have $\varepsilon_i \xrightarrow{P} 0$, so $\delta_i \xrightarrow{P} 0$. The law of large numbers gives $N^{-1} \sum_i \delta_i \to 0$. ■

The gap $\ell_{\mathrm{marg}} - \ell_{\mathrm{joint}} = -\sum_i \log p_{\hat{\theta}^+}(\mathbf{z}^{(i)} \mid \mathbf{x}^{(i)})$ is large when the fitted model assigns non-negligible probability to $\mathbf{Z} \neq \mathbf{z}^{(i)}$ given $\mathbf{x}^{(i)}$, i.e. when the deterministic relationship $\mathbf{z} = f(\mathbf{x})$ is not tightly concentrated in the fitted model. This means that the component nodes act as *parents* of $\mathbf{X}$ variables rather than children. In this regime the augmented BN induces a structured mixture over $\mathbf{X}$: the marginal $p(\mathbf{X}) = \sum_{\mathbf{z}} p(\mathbf{X} \mid \mathbf{z})\, p(\mathbf{z})$ integrates over mixture components that genuinely shape $p(\mathbf{X})$, and the gap reflects the entropy of the mixture weights. By contrast, when all component nodes are leaves, the deterministic relationship forces $p_{\hat{\theta}^+}(\mathbf{z}^{(i)} \mid \mathbf{x}^{(i)}) \approx 1$, the gap vanishes, and $\mathrm{BIC}_{\mathrm{marg}} \approx \mathrm{BIC}_{\mathrm{joint}}$ in agreement with Theorem 7. A large gap is therefore evidence that the components provide genuine explanatory power for $\mathbf{X}$ rather than merely summarising within-group structure, and $\mathrm{BIC}_{\mathrm{marg}}$ is the appropriate criterion since it evaluates exactly $p(\mathbf{X})$.

The connection between BIC and χ^2 geometry is close: BIC selects edges whose χ^2 association exceeds the penalty threshold since the log-likelihood ratio for adding an edge between two discrete variables is asymptotically proportional to the χ^2 statistic:

$$2\big[\ell(G^+_{\mathrm{with}}) - \ell(G^+_{\mathrm{without}})\big] \;\approx\; \chi^2(X_i, X_j).$$

BIC therefore measures the association between discrete variables in χ^2 units asymptotically, and a component node is a useful routing hub to the extent that its discrete states concentrate the χ^2 association.

MCA components are coherent with the χ^2 metric in two ways. First, they are ordered by decreasing χ^2 inertia: the eigenvalues of the Burt matrix are proportional to the χ^2 statistics of the within-group contingency tables, so $Z_1^{(g)}$ captures the dominant χ^2 association in group g, and $Z_2^{(g)}$ captures the next largest residual, and so on. Second, the Voronoi discretization partitions observations according to χ^2 distance from category centroids, so the discrete states of each $Z_l^{(g)}$ correspond to natural χ^2 clusters rather than arbitrary bins. Together, these properties mean that MCA components are *pre-ordered by their marginal value to BIC*: when the structure learner adds $Z_2^{(g)}$, it finds genuine residual χ^2 association not already captured by $Z_1^{(g)}$ and can route it without displacing existing structure.

PCA components lack this alignment. The l-th PCA component captures a direction of maximum Euclidean variance orthogonal to the first $l-1$, but equal-frequency binning places bin boundaries to equalize observation counts rather than to respect χ^2 clusters. The discrete states of a PCA component therefore have no guaranteed concentration property in BIC terms. When the k-th component is added, the structure learner may find it advantageous under $\mathrm{BIC}_{\mathrm{joint}}$ to route some $\mathbf{X}$ variables through Z_k rather than Z_1 — not because Z_k better captures the marginal structure of $\mathbf{X}$, but because its arbitrary bin boundaries happen to reduce joint CPT complexity. Because $\mathrm{BIC}_{\mathrm{joint}}$ is optimized over a strictly larger model class when k increases, the joint score cannot decrease; but the induced

marginal $p(\mathbf{X})$ under G_k^+ may be worse than under G_{k-1}^+ if the new component introduces edges whose joint contribution is positive while their net effect on $p(\mathbf{X})$ is negative — a consequence of the misalignment between Euclidean variance and χ^2 association rather than any specific structural change. To summarize, if $G_k^+ = \arg\max_{G \in \mathcal{G}_k} \mathrm{BIC}_{\mathrm{joint}}(G, D_k^+)$ is the network learned by Hill-Climbing on the k-component PCA-augmented dataset, then $\ell_{\mathrm{marg}}(G_k^+, D)$ need not be monotonically increasing in k in general, as the misalignment between Euclidean variance and χ^2 association may cause additional components to worsen the induced marginal $p(\mathbf{X})$ even when $\mathrm{BIC}_{\mathrm{joint}}$ improves.

5. Experiments

5.1. Data and setup

The Czech Attitude Barometer (CAB) (Kudrnáč et al., 2024) is a large panel survey on political attitudes, trust, mental health, migration attitudes, climate change views, and other social and political topics in Czechia. We use wave 1 ($N = 1{,}160$ respondents, $p = 45$ variables after preprocessing), partitioned by domain experts into $G = 8$ thematic groups – Trust in Democracy (TD), General Trust (TG), Mental Wellbeing (MW), Mental Resilience (MR), General Migration (GM), Ukraine Migration (UM), Interest in Politics (IP), and Climate Change (CC), ranging from 3 to 10 variables each – plus two ungrouped variables (gender and left-right self-placement); the full variable listing and group sizes are given in Appendix A. All variables except gender are ordinal Likert scales. Structure learning uses Hill-Climbing with the BIC score and 10 random restarts, as implemented in bnlearn (Scutari, 2010).

5.2. Multiple Correspondence Analysis

Figure 1 (left panel) shows the Benzécri-corrected cumulative explained variance ratio against the number of MCA components per group, with the dashed line marking the $\tau = 0.90$ threshold used for the MC-all model. No group is unidimensional at this threshold: the first component alone explains between 46.7% (Mental Resilience) and 78.5% (General Trust) of within-group inertia, so every group needs at least two components – though part of this additional inertia, as discussed below, reflects a known nonlinear artifact of the first dimension rather than independent content. Six groups (TD, TG, MW, GM, UM, IP) reach the threshold at component 2, while Mental Resilience and Climate Change require a third; MC-all therefore adds two component nodes for six groups and three for the remaining two. Mental Resilience is the most internally complex group, with its first component explaining far less (46.7%) than the 60–78% range of the others, reflecting that its six items capture distinct coping dimensions rather than one underlying factor.

The substantive content of each first component also varies by group (Figure 1, right panel). For Trust in Democracy it is dominated by trust in government, parliament, the presidency, and the constitutional court (loadings 0.56–0.75), with trust in political parties loading distinctly lower (0.25) – separating depersonalized state institutions from contested party politics. In more homogeneous batteries such as general-migration and Ukraine-refugee attitudes, the first component instead loads evenly across items (0.65–0.76), indicating MCA recovers an already unidimensional construct rather than imposing structure.

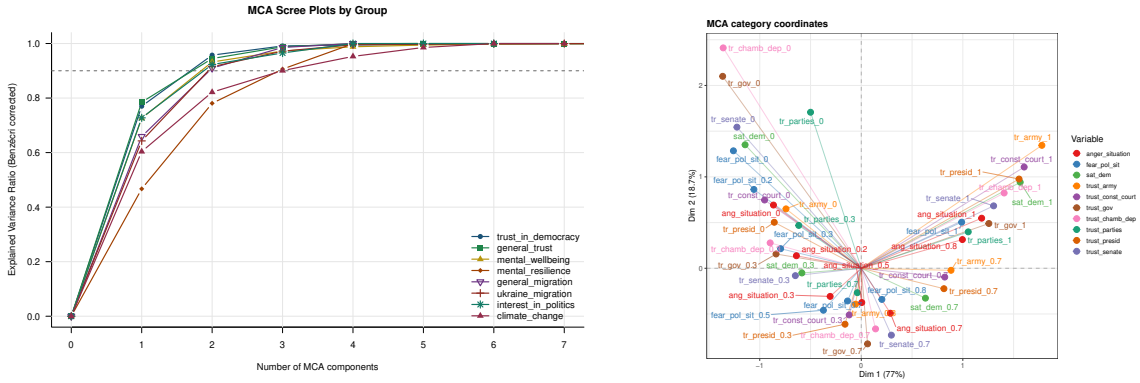


Figure 1: Left: Benzécri-corrected MCA explained variance ratio as a function of the number of components.
Right: MCA category coordinates for the Trust-in-Democracy (TD) group on its first two dimensions.

Across all eight groups, the second dimension shows the textbook parabolic arch (Greenacre, 1984) – both ordinal extremes of each item loading higher than intermediate categories – confirming later dimensions largely encode a nonlinear function of the first rather than independent content; we therefore restrict substantive interpretation to each group's first dimension throughout. Unlike EM-fitted latent class indicators, which are subject to label switching across refits (Stephens, 2000), each MCA component has this fixed, deterministic definition, giving a stable interpretive target – a further practical advantage over the latent-variable alternatives of Section 4.2.

5.3. Results

Table 1 reports, for each model, the parameter count $d(G^+)$ and in-sample marginal score $\mathrm{BIC}_{\mathrm{marg}}$ from a single fit on the full dataset, together with a genuinely held-out evaluation: we repeat 10-fold cross-validation, refitting the network structure, CPT parameters, and – for the PCA/MCA variants – the components themselves on each fold's 90% training portion, and evaluate the exact marginal log-likelihood $\log \sum_{\mathbf{z}} p_{\hat{\theta}^+}(\mathbf{x}, \mathbf{z})$ of the withheld 10%. Each held-out probability is computed exactly via evidence propagation (evidence set only on $\mathbf{X}$, summing out the component nodes rather than approximating them). We report the pooled held-out log-likelihood per observation, $\ell^{\mathrm{ho}}_{\mathrm{marg}}$, averaged over all N observations across the ten folds. Because structure, parameters, and components are refit independently on each fold for every model – No-PC, PCA, and MCA alike – no model retains information about its own held-out fold, keeping the comparison methodologically clean.

MC-1 wins on $\mathrm{BIC}_{\mathrm{marg}}$, and PC-1 wins on both ℓ_{marg} and $\mathrm{BIC}_{\mathrm{marg}}$ for PCA – there the penalty plays no role in the within-class ranking. For MCA, however, the raw-likelihood ranking (MC-all > MC-2 > MC-1) is reordered by the penalty term $d(G^+)/2 \log N$ to MC-1 > MC-2 > MC-all: it selects MC-1 in-sample and also promotes MC-2 over MC-all, since MC-2 matches MC-all's fit with substantially fewer parameters ($d(G^+) = 1{,}131$ vs. $1{,}210$).

Table 1: Model comparison on the Czech Attitude Barometer. $N = 1{,}160$; all values in log-likelihood units (natural logarithm). $\ell^{\text{ho}}_{\text{marg}}$ is the held-out marginal log-likelihood per observation, pooled over 10-fold cross-validation. Bold marks the best value in each column among the augmented models.

Method	Model	$d(G^+)$	ℓ_{marg}	BIC_{marg}	$\ell^{\text{ho}}_{\text{marg}}$
—	No-PC	810	−61,865	−64,723	−54.364
PCA	PC-1	731	−19,017	−21,596	−54.587
	PC-2	1,286	−19,100	−23,637	−54.765
	PC-all	2,082	−19,024	−26,370	−55.053
MCA	MC-1	824	−17,903	**−20,811**	−54.549
	MC-2	1,131	−17,142	−21,132	**−54.007**
	MC-all	1,210	**−17,110**	−21,379	−54.079

MCA augmentation improves held-out generalization; PCA augmentation does not. MC-2 and MC-all both outperform the plain BN ($\ell^{\text{ho}}_{\text{marg}} = -54.007$ and -54.079 vs. -54.364 for No-PC), gains of 0.357 and 0.285 nats per observation, with structure, parameters, and components alike never fit on the evaluation data. Every PCA-augmented model instead generalizes *worse* than the plain BN, degrading monotonically with the number of components (-54.587, -54.765, -55.053 for PC-1, PC-2, PC-all).

To confirm these pooled differences reflect consistent fold-level effects rather than a few favorable splits, we ran a paired Wilcoxon signed-rank test on the ten fold-level $\ell^{\text{ho}}_{\text{marg}}$ values for each pairwise comparison (fold assignment identical across all seven models), with Benjamini–Hochberg correction across the nine comparisons. MCA-2 and MCA-all each beat the plain BN in all ten folds ($p_{\text{adj}} = 0.003$ both), while MCA-1 and all three PCA variants underperform it in at least nine of ten folds ($p_{\text{adj}} \leq 0.007$) – so both the MCA benefit and the PCA/MCA-1 degradation are systematic rather than incidental. MCA-1 is also significantly worse than both MCA-2 and MCA-all (0/10 wins, $p_{\text{adj}} = 0.003$ each), confirming a single component per group is insufficient. The one exception is MCA-2 vs. MCA-all: MCA-2 wins 8 of 10 folds with a mean advantage of 0.072 nats, but the paired test cannot rule out chance ($p_{\text{adj}} = 0.105$).

The in-sample criterion favors the wrong number of MCA components: BIC_{marg} ranks MC-1 best ($-20{,}811$), yet MC-1 is the only MCA variant that fails to beat the plain BN out-of-sample (-54.549 vs. -54.364) and is in fact the worst-generalizing of the three, losing all ten folds to both MC-2 and MC-all. With only one summarizing node per group, MC-1 concentrates each group's explanatory content into a single channel, so imprecision in that channel's learned edges has an outsized effect on unseen data; MC-2 and MC-all spread the same information across more nodes and generalize more robustly. MC-2 gives the best point estimate for held-out generalization overall ($\ell^{\text{ho}}_{\text{marg}} = -54.007$) with substantially fewer parameters than MC-all ($d(G^+) = 1{,}131$ vs. $1{,}210$), though the two are not statistically distinguishable after correction ($p_{\text{adj}} = 0.105$). We recommend **MC-2** on the combined grounds of its point estimate and its more parsimonious structure, while noting that MC-all remains a defensible alternative. Figure 2 shows the MC-2 structure.

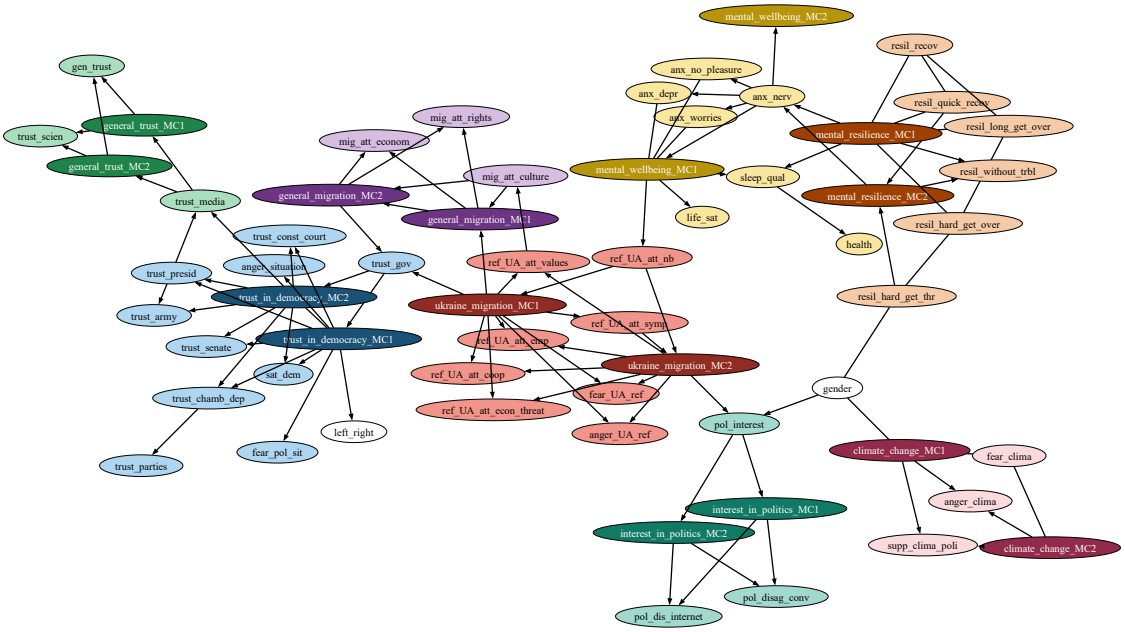


Figure 2: CPDAG of the MC-2 BN. Dark nodes are MCA component nodes and light nodes represent original observed variables colored by group.

This reversal is a genuine structure-learning effect, not preprocessing leakage: components are refit independently on each fold for MC-1, MC-2, and MC-all alike, so no leakage advantage differs across them and cannot explain why MC-1 alone underperforms. It must instead originate in how Hill-Climbing uses the available component nodes as their number varies – consistent with Remark 3, where a single component is more likely forced into a source or hub role carrying a disproportionate share of the group's explanatory burden. These results sharpen the paper's central claim: MCA is not just theoretically more principled than PCA for categorical data (Section 4), but in this application measurably improves out-of-sample fit while PCA augmentation measurably harms it.

6. Conclusions

We have shown that augmenting a Bayesian network with deterministically computed discrete components yields a fully observed model in a curved exponential family, making standard BIC valid for joint model scoring. The modified $\mathrm{BIC}_{\mathrm{marg}}$ provides fair cross-class comparison by evaluating all models on their marginal fit to the original variables. MCA is preferable to PCA for discrete data: it operates on the χ^2 geometry native to categorical variables, eliminates arbitrary continuous-score discretization, and yields stable joint likelihoods across model complexity. Empirically, the MCA models achieve the best $\mathrm{BIC}_{\mathrm{marg}}$ as well as the best held-out marginal log-likelihood across PCA variants and the base BN. The method can be extended to mixed datasets with both continuous and categorical variables, using PCA for the former and MCA for the latter.

Acknowledgments

This work was supported by the European Regional Development Fund project "Beyond Security: Role of Conflict in Resilience-Building" (reg. no.: CZ.02.01.01/00/22_008/0004595).

References

Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11:171–234, 2010. URL https://www.jmlr.org/papers/v11/aliferis10a.html.

Jean-Paul Benzécri. Sur le calcul des taux d'inertie dans l'analyse d'un tableau de données. *Cahiers de l'Analyse des Données*, 4(3):377–378, 1979. URL http://eudml.org/doc/87965.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977. URL https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Gal Elidan, Noam Lotner, Nir Friedman, and Daphne Koller. Discovering hidden variables: A structure-based approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 13, pages 479–485, 2001.

Nir Friedman and Daphne Koller. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003. URL https://doi.org/10.1023/A:1020249912095.

Dan Geiger, David Heckerman, and Christopher Meek. Asymptotic model selection for directed networks with hidden variables. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 158–168, San Francisco, CA, 1996. Morgan Kaufmann. URL https://arxiv.org/abs/1302.3580.

Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29(2):505–529, 2001. URL https://doi.org/10.1214/aos/1009210550.

Michael J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1984.

Dominique M. A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355, 1988. URL https://doi.org/10.1214/aos/1176350709.

Karl G. Jöreskog. A general method for analysis of covariance structures. *Biometrika*, 57 (2):239–251, 1970. URL https://10.1093/biomet/57.2.239.

Karl G. Jöreskog. On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 59:381—389, 1994. URL https://doi.org/10.1007/BF02296131.

Aleš Kudrnáč, Ivan Petrúšek, Tomáš Dvořák, Jakub Cigán, Arnošt Veselý, Magdalena Mouralová, Michaela Prokeš, Paulína Tabery, Naděžda Čadová, Jana Vitíková, Štěpánka Weikertová, Matouš Pilnáček, Martin Spurný, Jiří Vinopal, Marie Plačková, Monika Kyselá, Helena Hájková, Karolína Ježková, and Lucie Novotná. Czech Attitude Barometer, wave 1, 2024. URL https://doi.org/10.14473/CSDA/B73GVD.

Steffen L. Lauritzen. The EM algorithm for graphical association models with missing data. In *Computational Statistics & Data Analysis*, volume 19, pages 191–201. Elsevier, 1995. URL https://doi.org/10.1016/0167-9473(93)E0056-A.

Bruce G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, 1995.

Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley, New York, 2000.

Marina Meilă and Michael I. Jordan. Learning with mixtures of trees. In *Journal of Machine Learning Research*, volume 1, pages 1–48, 2001. URL https://jmlr.org/papers/v1/meila00a.html.

Dmitry Rusakov and Dan Geiger. Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, 6(1):1–35, 2005. URL http://jmlr.org/papers/v6/rusakov05a.html.

Marco Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. URL https://doi.org/10.18637/jss.v035.i03.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. In *Journal of Machine Learning Research*, volume 7, pages 191–246, 2006. URL https://jmlr.org/papers/v7/silva06a.html.

Charles Spearman. "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15(2):201–293, 1904. URL https://doi.org/10.2307/1412107.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*, pages 795–809, 2000. URL https://doi.org/10.1111/1467-9868.00265.

Louis L. Thurstone. *The Vectors of Mind*. University of Chicago Press, Chicago, 1935.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999. URL https://doi.org/10.1111/1467-9868.00196.

Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006. URL https://doi.org/10.1007/s10994-006-6889-7.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949. URL https://doi.org/10.1214/aoms/1177729952.

Nevin L. Zhang. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723, 2004. URL https://www.jmlr.org/papers/v5/zhang04a.html.

Appendix A. Variable Groups

This appendix lists all variable groups used in the analysis, along with a brief description of each variable's meaning.

A.1. Trust in Democracy (TD)

`sat_dem` — Satisfaction with the functioning of democracy
`trust_presid` — Trust: President of the Republic
`trust_gov` — Trust: Government
`trust_chamb_dep` — Trust: Chamber of Deputies
`trust_senate` — Trust: Senate
`trust_const_court` — Trust: Constitutional Court
`trust_parties` — Trust: Political parties
`trust_army` — Trust: Army
`fear_pol_sit` — Fear: political situation in the Czech Republic
`anger_situation` — Anger: overall situation in the Czech Republic

A.2. General Trust (TG)

`gen_trust` — Social (generalized) trust
`trust_media` — Trust: Public-service media
`trust_scien` — Trust: Scientists

A.3. Mental Well-Being (MW)

`anx_nerv` — Depression and anxiety: feeling nervous, anxious, or on edge
`anx_worries` — Depression and anxiety: not being able to stop or control worrying
`anx_depr` — Depression and anxiety: feeling down, depressed, or hopeless
`anx_no_pleasure` — Depression and anxiety: little interest or pleasure in doing things
`health` — Subjective health
`sleep_qual` — Sleep quality
`life_sat` — Life satisfaction

A.4. Mental Resilience (MR)

`resil_recov` — Resilience: I tend to bounce back quickly after hard times
`resil_hard_get_thr` — Resilience: I have a hard time making it through stressful events
`resil_quick_recov` — Resilience: It does not take me long to recover from a stressful event
`resil_hard_get_over` — Resilience: It is hard for me to snap back when something bad happens
`resil_without_trbl` — Resilience: I usually come through difficult times with little trouble
`resil_long_get_over` — Resilience: I tend to take a long time to get over setbacks in my life

A.5. General Migration (GM)

`mig_att_econom` — Attitudes towards migrants: economic benefit
`mig_att_culture` — Attitudes towards migrants: new ideas and culture
`mig_att_rights` — Attitudes towards migrants: equal rights

A.6. Ukraine Migration (UM)

`ref_UA_att_values` — Attitudes towards Ukrainian refugees: values in the Czech Republic
`ref_UA_att_econ_threat` — Attitudes towards Ukrainian refugees: threat to the economy
`ref_UA_att_emp` — Attitudes towards Ukrainian refugees: empathy/compassion
`ref_UA_att_symp` — Attitudes towards Ukrainian refugees: sympathy
`ref_UA_att_coop` — Attitudes towards Ukrainian refugees: cooperation
`ref_UA_att_nb` — Attitudes towards Ukrainian refugees: living in the neighborhood
`fear_UA_ref` — Fear: presence of Ukrainian refugees in the Czech Republic
`anger_UA_ref` — Anger: presence of Ukrainian refugees in the Czech Republic

A.7. Interest in Politics (IP)

`pol_interest` — Interest in politics
`pol_disag_conv` — Political disagreements in everyday conversation
`pol_dis_internet` — Discussing politics online

A.8. Climate Change (CC)

`fear_clima` — Fear: climate change
`anger_clima` — Anger: climate change
`supp_clima_poli` — Support for climate change policies

A.9. Other Variables (Not Assigned to a Group)

`gender` — Gender
`left_right` — Ideology: left–right self-placement