A Statistical Classification of Cryptocurrencies

Daniel Traian Pele, Niels Wesselhöfft, Wolfgang K. Härdle, Yannis Yatracos, Michalis Kolossiatis

International Research Training Group 1792
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt–Universität zu Berlin

Department of Statistics and Econometrics
Bucharest University of Economic Studies

Department of Mathematics and Statistics
University of Cyprus, Nicosia

Yau Mathematical Sciences Center, Tsinghua University, Beijing, China

Genus differentia approach

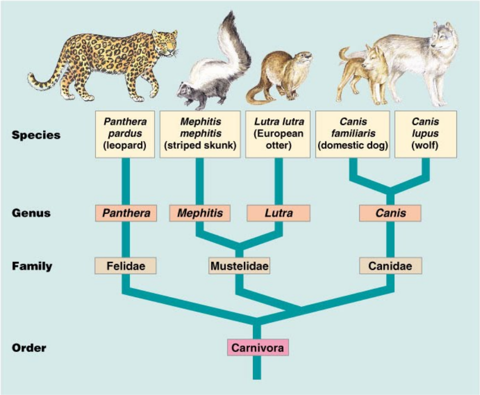


Figure: Genus differentia approach in biology

Genus differentia approach

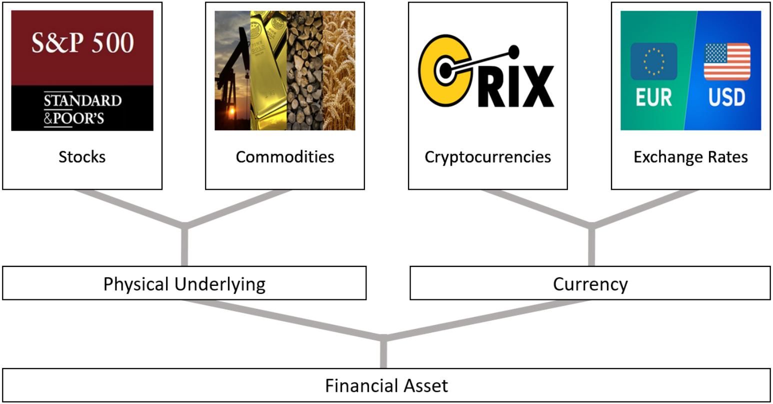


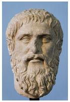
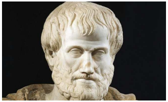
Figure: Genus differentia approach in finance

Aim of classification

- Genotypic differentiation
  - Biology - the change in DNA sequences.
  - Finance - the underlying process of price manifestation.
- Phenotypic differentiation
  - Biology - classification based on behavior and features of a species.
  - Finance - classification based on statistical features of the price series.

Motivation

- Question: What defines cryptocurrencies?

- Plato: man is an upright, featherless biped, with broad, fat nails.
- Aristotle: definition of a species consists of genus proximum and differentia specifica.
- Goal: Define cryptocurrencies in terms of their genus proximum and differentia specifica.
- Method: Find latent variables, to form groups of shared characteristics.
- Finding: Synchronic evolution, i.e. asymptotic speciation.
- Implication: Cryptocurrencies are a different species in the ecosystem of financial instruments.

Outline

1. Motivation
2. Data and descriptives
3. Factor model
4. Explanation
5. Expanding window
6. Conclusion

Literature review

- Dyhrberg (2016): BTC has similarities to both GOLD and the USD, being in between a currency and a commodity.
- Baur et al. (2018): BTC volatility and correlation characteristics are distinctively different compared to GOLD and USD.
- Härdle et al. (2018): BTC, XRP, LTC, ETH returns exhibit higher volatility, skewness and kurtosis compared to GOLD and S&P500 daily returns.
- Zhang et al. (2018): Cryptocurrencies presents heavier tails and higher Hurst exponent than the classical assets.
- Liu et al. (2019) developed a three-factor model using the CAPM approach and showed that the cross-sectional expected cryptocurrency returns can be captured by three factors: the market factor, the size factor and momentum factor.

Data

- Sample: $n = 679$ assets.
- New asset class
  - Cryptocurrencies: $n_1 = 150$
- Old asset classes
  - Stocks (S&P 500): $n_2 = 496$
  - Exchange rates: $n_3 = 13$ List
  - Commodities (Bloomberg Commodity Index): $n_4 = 20$ List
- Daily data from 01/02/2014 - 08/30/2019 (1426 trading days).

Statistical assessment

- Return X is a r.v. with cdf $F()$ from which $p = 23$ statistics are estimated.
- Moments of order $k \in \mathbb{R}^+$, $\mu_k = \mathrm{E}\left\{(X-\mu)^k\right\}$.
  - variance: $\sigma^2 = E\left\{(X-\mu)^2\right\}$;
  - skewness: $Skewness = E\left\{(X-\mu)^3\right\}/\sigma^3$;
  - kurtosis: $Kurtosis = E\left\{(X-\mu)^4\right\}/\sigma^4$.
- Tails: $\alpha \in \{0.005, 0.01, 0.025, 0.05, 0.95, 0.975, 0.99, 0.995\}$.
  - $Q_\alpha = \inf\{x \in \mathbb{R} : \alpha \leq F(x)\}$;
  - $CTE_\alpha = \begin{cases} \mathrm{E}\{X \mid X < Q_\alpha\}, & \alpha < 0.5 \\ \mathrm{E}\{X \mid X > Q_\alpha\}, & \alpha > 0.5 \end{cases}$
- Scaling and memory parameters
  - Alpha-stability ▸ Alpha-stability
  - ARCH parameter (GARCH (1,1))
  - GARCH parameter (GARCH (1,1))

Assets profile

Variable	Commodities	Cryptocurrencies	Exchange rates	Stocks
$\sigma^2 \cdot 10^3$	0.365	14.563	0.028	0.270
Skewness	0.245	0.723	−1.233	−0.520
Kurtosis	22.461	28.037	38.201	13.392
$Stable_\alpha$	1.721	1.410	1.714	1.711
$Stable_\gamma$	0.010	0.047	0.003	0.009
$Q_{5\%}$	−0.027	−0.159	−0.008	−0.025
$Q_{2.5\%}$	−0.034	−0.210	−0.010	−0.033
$Q_{1\%}$	−0.044	−0.296	−0.013	−0.044
$Q_{0.5\%}$	−0.054	−0.378	−0.015	−0.054
$CTE_{5\%}$	−0.038	−0.250	−0.011	−0.038
$CTE_{2.5\%}$	−0.047	−0.319	−0.014	−0.047
$CTE_{1\%}$	−0.060	−0.428	−0.017	−0.062
$CTE_{0.5\%}$	−0.073	−0.525	−0.021	−0.076
$Q_{95\%}$	0.026	0.169	0.008	0.024
$Q_{97.5\%}$	0.034	0.243	0.010	0.030
$Q_{99\%}$	0.046	0.364	0.013	0.040
$Q_{99.5\%}$	0.057	0.480	0.015	0.049
$CTE_{95\%}$	0.039	0.297	0.011	0.034
$CTE_{97.5\%}$	0.049	0.393	0.013	0.042
$CTE_{99\%}$	0.064	0.544	0.016	0.055
$CTE_{99.5\%}$	0.080	0.671	0.018	0.066
GARCH parameter	0.706	0.796	0.728	0.637
ARCH parameter	0.118	0.159	0.078	0.130

Table: Assets profile

Factor analysis

- Estimate the correlation matrix for all variables.
- Factor extraction based on the correlation of the coefficients.
- Factor rotation.

Correlation matrix

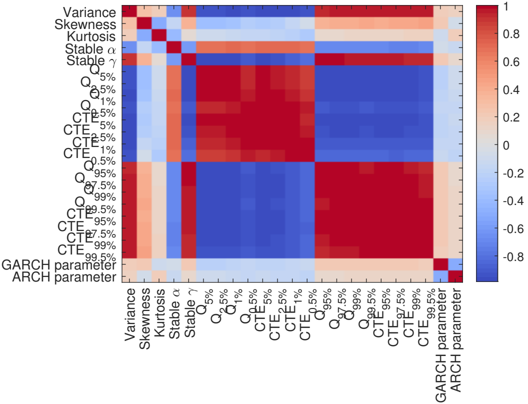


Figure: Correlation matrix of the statistical estimates. SFA_cryptos

Factor model

- Linear Factor model

$$X = QF + \mu + \varepsilon,\ \varepsilon \sim G() \tag{1}$$

- X is the initial matrix of p variables
- Q is a matrix of the non-random loadings
- F are the common k factors ($k < p$)
- μ is the vector of the means of initial p variables
- ε is a matrix of the random specific factors
- Random vectors F and U are unobservable and uncorrelated

Factors loadings and scree plot

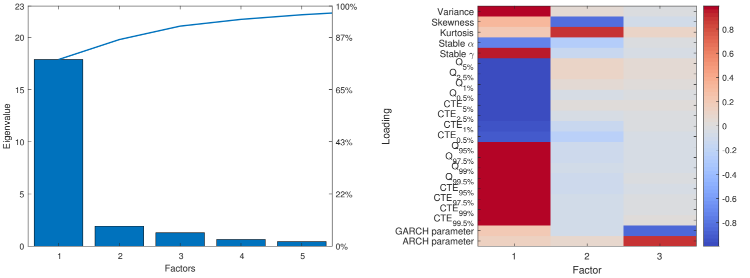


Figure: Scree plot and factors loadings. SFA_cryptos

Factor rotation

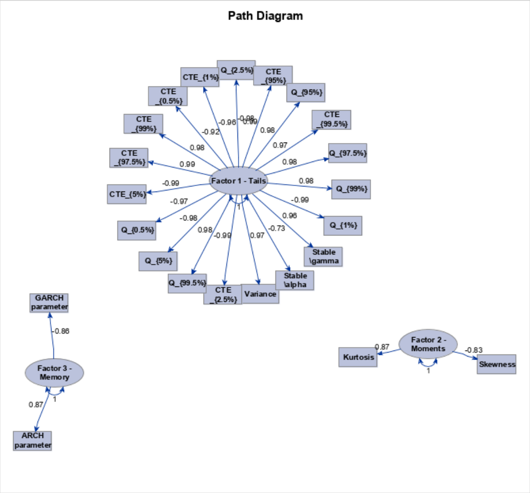


Figure: Path diagram. FA_cryptos

Mapping of the factors

1. Tail factor - 77% of the total variance
   - Alpha-stable parameters S_α, S_γ
   - Lower and upper quantiles
   - Conditional tail expectations
   - Variance
2. Moment factor - 8% of the total variance
   - Skewness
   - Kurtosis
3. Memory factor - 6% of the total variance
   - ARCH parameter
   - GARCH parameter

Tail factor vs Moment factor

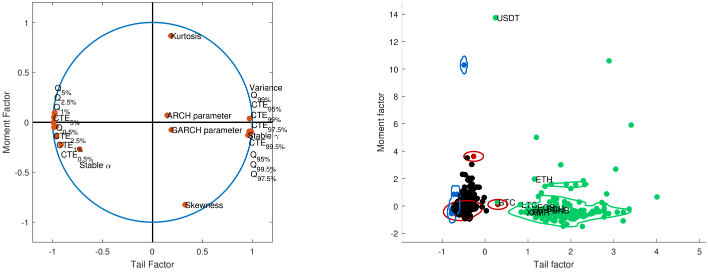


Figure: Loadings (left) and scores (right) based on tail and moment factor. SFA_cryptos

Tail factor vs Memory factor

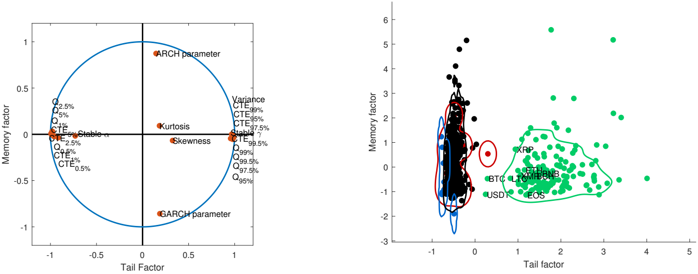
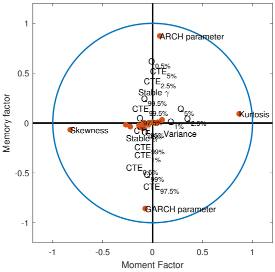


Figure: Loadings (left) and scores (right) based on tail and memory factor. SFA_cryptos

Moment factor vs Memory factor



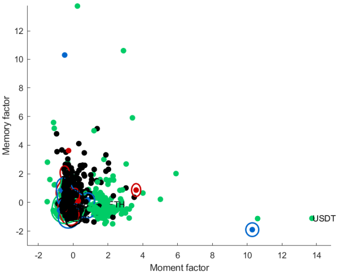


Figure: Loadings (left) and scores (right) based on moment and memory factor. SFA_cryptos

Factor explanation

- Classify between Cryptocurrencies and other asset classes
- Binary logistic regression for each factor $F_k, \;\; k \in \{1, 2, 3\}$

$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 F_k)}{1 + \exp(\beta_0 + \beta_1 F_k)}, \tag{2}$$

$$Y = \begin{cases} 1, & \text{if Cryptocurrency} \\ 0, & \text{if otherwise} \end{cases} \tag{3}$$

Factor explanation

Exogenous factor	Factor 1	Factor 2	Factor 3
Estimated β_1	15.679***	-0.030	-0.084
	(3.278)	(0.077)	(0.093)
$\widetilde{R}^2$	0.992	0.0003	0.002

Note: Standard errors in (); ** denotes significance at 95% confidence level.

$$\widetilde{R}^2 = \frac{1 - \left\{ \frac{L(\mathbf{0})}{L(\widehat{\beta})} \right\}^{\frac{2}{n}}}{1 - \{L(\mathbf{0})\}^{\frac{2}{n}}} \tag{4}$$

- $L(0)$ is the likelihood of the intercept-only model
- $L(\widehat{\beta})$ is the likelihood of the full model

Linear Discriminant Analysis

- Finding a projection that maximizes the separability between classes.
- Assumes Gaussianity with equal covariances.

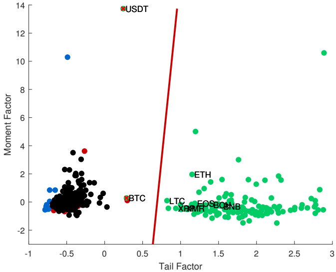


Figure: LDA ▸ LDA

Quadratic Discriminant Analysis

- Finding a projection that maximizes the separability between classes.
- Assumes Gaussianity with different covariances.

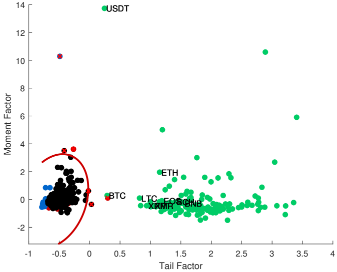


Figure: Quadratic Discriminant Analysis

Support Vector Machines

- Finding a projection that maximizes margin in a hyperplane of the original data.
- No parametric assumptions on the underlying probability distribution function.

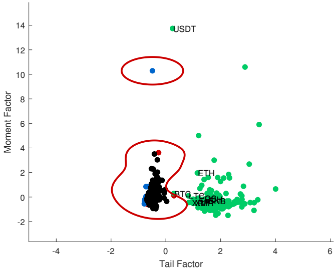


Figure: SVM ▸ SVM

K-means clustering

- Projection of the clusters on the 3D space extracted trough Factor Analysis.
- Each cryptocurrencies cluster was labeled with its leader in terms of market capitalization.

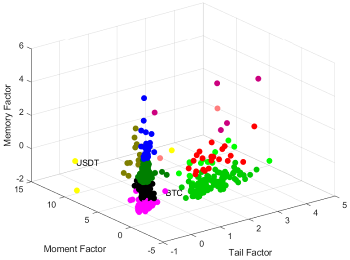


Figure: 3D. Cluster_cryptos

Maximum Variance Components Split

- These method have goals to separate, respectively, the components of a structure like the types of assets herein, and clusters defined as the components of a mixture distribution.
- They are based on an unusual variance decomposition in between-group variations.

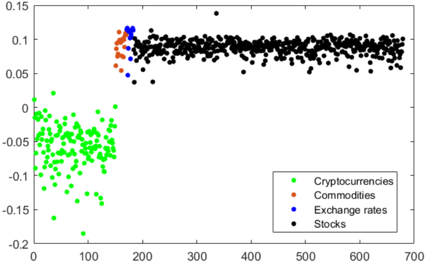
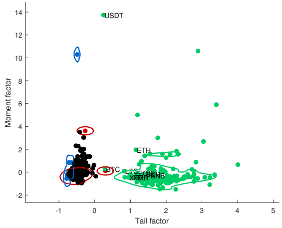


Figure: MVCS. VCS_cryptos

MVCS

Video

- Expanding rolling window estimation
  - Starting window 2014-01-02 until 2016-10-231 (1/2 of the data)
  - Increases daily up to full window 2014-01-02 until 2019-08-30
  - Kernel density contour level 0.015
- Clusters converge over time



DFA_cryptos

Synchronic evolution

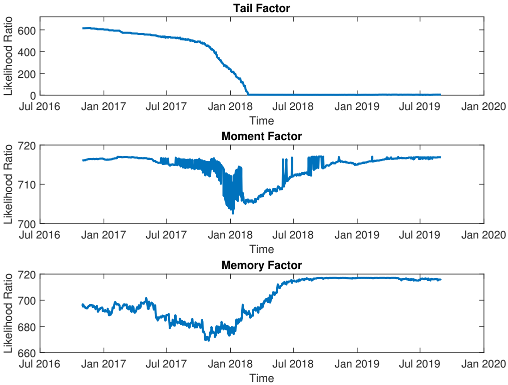


Figure: Likelihood Ratios for the binary logistic model, estimated for the period 10/31/2016- 08/30/2019. CONV_cryptos

Conclusion

- Financial perspective
  - Main statistical difference between Cryptocurrencies and other asset classes: tail behavior.
  - Moments and memory are of subliminal importance.
  - Nonlinear classification with SVM provides proficient results for risk analysts and regulators.
  - Cryptocurrencies are completely separated by the other types of assets, as proved by Maximum Variance Components Split method.
- Biological perspective
  - Speciation takes time to form distinct species, which potentially evolve further away from each other.
  - Cryptocurrencies establish themselves as unique asset classes.

A Statistical Classification of Cryptocurrencies

Daniel Traian Pele, Niels Wesselhöfft, Wolfgang K. Härdle, Yannis Yatracos, Michalis Kolossiatis

International Research Training Group 1792
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt–Universität zu Berlin

Department of Statistics and Econometrics
Bucharest University of Economic Studies

Department of Mathematics and Statistics
University of Cyprus, Nicosia

Yau Mathematical Sciences Center, Tsinghua University, Beijing, China

Exchange rates

▸ Data

1. EUR/USD Euro
2. JPY/USD Japanese Yen
3. GBP/USD Great Britain Pound
4. CAD/USD Canada Dollar
5. AUD/USD Australia Dollar
6. NZD/USD New Zealand Dollar
7. CHF/USD Swiss Franc
8. DKK/USD Danish Krone
9. NOK/USD Norwegian Krone
10. SEK/USD Swedish Krone
11. CNY/USD Chinese Yuan Renminbi
12. HKD/USD Hong Kong Dollar
13. INR/USD Indian Rupee

Commodities

Data

1. WTI Crude oil USCRWTIC Index
2. Natural Gas NGUSHHUB Index
3. Brent oil EUCRBRDT Index
4. Unleaded Gasoline RBOB87PM Index
5. ULS Diesel DIEINULP Index
6. Live cattle SPGSLC Index
7. Lean hogs HOGSNATL Index
8. Wheat WEATTKHR Index
9. Corn CRNUSPOT Index
10. Soybeans SOYBCH1Y Index
11. Aluminum LMAHDY Comdty
12. Copper LMCADY Comdty
13. Zinc ZSDY Comdty
14. Nickel CKEL Comdty
15. Tin JMC1DLTS Index
16. Gold XAU Curncy
17. Silver XAG Curncy
18. Platinum XPT Curncy
19. Cotton COTNMAVG Index
20. Cocoa MLCXCCSP Index

Lévy-Stable distributions

- Fourier transform of characteristic function $\varphi_X(u)$

$$S(X \mid \alpha, \beta, \gamma, \delta) = \frac{1}{2\pi} \int \varphi_X(u) \exp(-iuX) du$$

- Characteristic function representation, $0 < \alpha < 2, \alpha \neq 1$

$$\log \varphi_X(u) = iu\delta - \gamma |u|^{\alpha} \{1 + i\beta (u/|u|) \tan(\alpha\pi/2)\} \qquad (5)$$

- Stability or invariance under addition

$$n \log \varphi_X(u) = iu(n\delta) - (n\gamma)|u|^{\alpha} \{1 + i\beta (u/|u|) \tan(\alpha\pi/2)\}$$

- Limiting distribution of n i.i.d. stable r.v., $0 < \alpha \leq 2$
GCLT (Gnedenko and Kolmogorov, 1954)

$$n^{-\frac{1}{\alpha}} \sum_{i=1}^{n} (X_i - \delta) \xrightarrow{\mathcal{L}} S(\alpha, \beta, \gamma, 0) \qquad (6)$$

Linear Discriminant Analysis

- Let $X_i \sim N(\mu_i, \Sigma_i)$ belonging to class ω_i, $\Sigma_i = \Sigma_j$
- Project samples X onto a line $Y = w^\top X$
- Select the projection that maximized the separability
- Maximize normalized, squared distance in the means of the classes

$$w^* = \arg\max_w \frac{|w^\top(\mu_i - \mu_j)|^2}{s_i^2 + s_j^2}, \tag{7}$$

$$s_i^2 = \sum_{x_i \in \omega_i} (w^\top x_i - w^\top \mu_i)^2 = w^\top S_i w \tag{8}$$

- Linear Discriminant of Fisher (1936)

$$w^* = S_W^{-1}(\mu_i - \mu_j), \; S_W = S_i + S_j \tag{9}$$

▸ LDA

Support Vector Machines

- Given training data set D with n samples and 2 dimensions

$$D = (X_1, Y_1), \ldots (X_n, Y_n), \quad X_i \in \mathbb{R}^2, \quad Y_i \in [0, 1]$$

- Finding a hyperplane that maximizes the margin

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

$$\text{s.t.} \quad Y_i \left(w^\top X_i + b \right) \geq 1, \quad i = 1, \ldots, n$$

Figure: SVM

Variance Component Split

- Consider the groups $X_{(1)}, \ldots, X_{(i)}$ and $X_{(i+1)}, \ldots, X_{(n)}$ with averages, respectively, $\overline{X}_{[1,i]}$ and $\overline{X}_{[i+1,n]}$, $i = 1, \ldots, n-1$, then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n-1}\frac{i(n-i)}{n^2}(\overline{X}_{[i+1,n]} - \overline{X}_{[1,i]})(X_{(i+1)} - X_{(i)}). \quad (10)$$

- The relative contribution of the groups $X_{(1)}, \ldots, X_{(i)}$ and $X_{(i+1)}, \ldots, X_{(n)}$ in the sample variability:

$$W_i = W_i(X_1, \ldots, X_n) = \frac{i(n-i)}{n}\frac{(\overline{X}_{[i+1,n]} - \overline{X}_{[1,i]})(X_{(i+1)} - X_{(i)})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad (11)$$

- Index $\mathcal{I}_n = \max\{W_i, i = 1, \ldots, n-1\}$ determines two potential clusters or parts of a structure and is based on averages and inter-point distances.

Maximum Variance Component Split

- The Maximum Variance Component Split (MVCS) method compares known components of a structure, *e.g.* cryptocurrencies herein, with data splits for a set of unit projection directions $\mathcal{D}_M$ usually determined by M positive equidistant angles of $[0, \pi]$; *e.g.* when $r = 2$ and $M = 3$ the angles used are $\pi/3, 2\pi/3, \pi$.
- When one of the data split along projection direction **a** coincides with a component of the structure we have complete separation of this component along **a**.
- A set of projection directions $\mathcal{D}_M$ can be

$$(\Pi_{l=1}^{r} cos\theta_l,\ \ sin\theta_1 \Pi_{l=2}^{r} cos\theta_l, ...,\ \ sin\theta_{r-1} cos\theta_r,\ \ sin\theta_r), \tag{12}$$

where θ_l takes values in $\{\frac{m\pi}{M}, m = 1, ..., M\}, l = 1, ..., r$.

▸ MVCS