Simultaneous Inference for Latent Variables in Factor Analytic Models

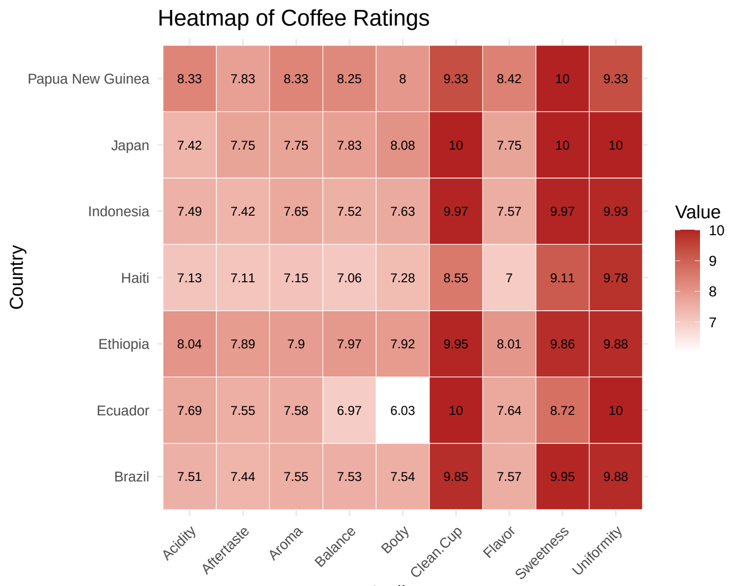
Zhining Wang, Emi Tanaka, and Francis Hui

Research School of Finance, Actuarial Studies and Statistics,
College of Business and Economics, The Australian National University

Motivating Data: Coffee Bean Exploration

- The coffee bean dataset records consumer ratings (0–10) on nine sensory attributes (e.g., sweetness, aroma, acidity) and aggregates them by country. The dataset consists of $n = 32$ countries in total.
  - Data are publicly available from Kaggle `https://www.kaggle.com/datasets/adampq/coffee-quality-with-locations-of-origin?resource=download`, which itself is sourced from the Coffee Quality Instiute (CQI) database `https://database.coffeeinstitute.org/`

Motivating Data: Coffee Bean Exploration



Motivating Data: Coffee Bean Exploration

- We hypothesise a few latent factors can be used to represent these observed ratings e.g., latent axis representing aspects of coffee taste or richness. That is, we wish to perform dimension reduction.
- Our interest is to perform statistical inference in this latent space e.g., do two countries differ statistically in terms of coffee taste/richness? Is the ”best” country statistically different from the rest?

The Factor Analytic Model

Suppose we have n independent observations $\mathbf{y}_i; i = 1, 2, \ldots, n$, each of dimension p. Assume the j-th response y_{ij} can be written as (West et al., 2003)

$$y_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{f}_i^\top \boldsymbol{\lambda}_j + \epsilon_{ij},$$

where

- $\mathbf{x}_i$ is a q-dimensional covariate and $\boldsymbol{\beta}_j$ the associated coefficients;
- $\mathbf{f}_i$ is an m-dimensional latent factor assumed to be standard normal in distribution, $f_{ik} \sim \mathcal{N}(0, 1)$;
- $\boldsymbol{\lambda}_j$ is an m-dimensional loading vector. Furthermore, we define the **loading matrix** $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1 \; \boldsymbol{\lambda}_2 \; \cdots \; \boldsymbol{\lambda}_p]^\top$;
- ϵ_{ij} is an error term where we assume $\epsilon_{ij} \sim \mathcal{N}(0, \psi_{jj})$.

- We write the fully vectorised form of the factor analytic model as,

$$\boldsymbol{y} = (\mathbf{I}_p \otimes \mathbf{X})\, \boldsymbol{\beta} + (\boldsymbol{\Lambda} \otimes \mathbf{I}_n)\, \boldsymbol{f} + \boldsymbol{\epsilon},$$

  where $\boldsymbol{y} = (y_{11}, \cdots, y_{n1}, y_{12}, \cdots, y_{np})^\top$, $\mathbf{X}^\top = \left[\mathbf{x}_1^\top\ \mathbf{x}_2^\top\ \ \ldots\ \mathbf{x}_n^\top\right]$ is an $q \times n$ matrix, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_p^\top\right)^\top$, $\boldsymbol{f} = (f_{11}, \cdots, f_{n1}, f_{12}, \cdots, f_{np})^\top$ and $\boldsymbol{\epsilon} = (\epsilon_{11}, \cdots, \epsilon_{n1}, \epsilon_{12}, \cdots, \epsilon_{np})^\top$.
- Marginally, we have

$$\boldsymbol{y} \sim \mathcal{N}_{np}\left((\mathbf{I}_p \otimes \mathbf{X})\, \boldsymbol{\beta}, \mathbf{V} \otimes \mathbf{I}_n\right),$$

  where $\mathbf{V} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$ is the marginal variance of each observation $\mathbf{y}_i$, and we call $\boldsymbol{\theta} := \left[\mathrm{vec}(\boldsymbol{\Lambda})^\top, \mathrm{diag}(\boldsymbol{\Psi})^\top\right]^\top$ the **variance components**.

Model Identifiability

- Without any constraints on the loading matrix $\mathbf{\Lambda}$, the factor model is not identifiable, since for any orthogonal matrix $\mathbf{M}$, the transformed loading matrix $\mathbf{\Lambda M}$ yields the same covariance structure and the same likelihood (Mardia et al., 1979).
- We put a **corner constraint** on the loading matrix by setting the upper-triangular elements of $\mathbf{\Lambda}$ to zero, and restrict the diagonal elements of $\mathbf{\Lambda}$ to be positive.

Parameter Estimation and Factor Prediction

Given variance component $\boldsymbol{\theta}$,

- the MLE of $\boldsymbol{\beta}$ is
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{I}_p \otimes (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right) \boldsymbol{y},$$
- the best linear unbiased prediction (BLUP) (Searle et al., 2009) of $\boldsymbol{f}$ is
$$\tilde{\boldsymbol{f}} = \mathbb{E}\left[\boldsymbol{f} \mid \boldsymbol{y}\right] = \left(\boldsymbol{\Lambda}^\top \mathbf{V}^{-1} \otimes \mathbf{I}_n\right)\left(\boldsymbol{y} - (\mathbf{I}_p \otimes \mathbf{X})\hat{\boldsymbol{\beta}}\right).$$

In practice, we can plug in estimated values of $\hat{\boldsymbol{\theta}}$ (e.g. obtained via REML (Corbeil and Searle, 1976)) to obtain $\hat{\boldsymbol{f}}$.

General Mixed Parameters

We are interested in conduct simultaneous inference for a set of L general mixed parameters. For known designed vectors $\boldsymbol{c}_l$ and $\boldsymbol{\kappa}_l$, the **general mixed parameter** (Reluga et al., 2023) for level $l = 1, \ldots, L$ is given by

$$\mu_l = \boldsymbol{c}_l^\top \boldsymbol{\beta} + \boldsymbol{\kappa}_l^\top \boldsymbol{f}$$

The BLUP for the general mixed parameter is then given by plugging $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{f}}$,

$$\tilde{\mu}_l = \boldsymbol{c}_l^\top \hat{\boldsymbol{\beta}} + \kappa_l^\top \tilde{\boldsymbol{f}},$$

and the empirical BLUP or EBLUP follows us

$$\hat{\mu}_l = \boldsymbol{c}_l^\top \hat{\boldsymbol{\beta}} + \kappa_l^\top \hat{\boldsymbol{f}}.$$

Prediction Mean Square Error

- For level $l = 1, \ldots, L$, the **prediction mean square error** (PMSE) $\sigma^2(\hat{\mu}_l)$ can be decomposed as

$$\begin{aligned}\mathbb{E}[(\hat{\mu}_l - \mu_l)^2] &= \mathbb{E}[(\hat{\mu}_l - \tilde{\mu}_l)^2] + \mathbb{E}[(\tilde{\mu}_l - \mu_l)^2] + 2\mathbb{E}[(\hat{\mu}_l - \tilde{\mu}_l)(\tilde{\mu}_l - \mu_l)] \\ &= \mathbb{E}[(\hat{\mu}_l - \tilde{\mu}_l)^2] + \sigma^2(\tilde{\mu}_l).\end{aligned}$$

Prediction Mean Square Error

- For level $l = 1, \ldots, L$, the **prediction mean square error** (PMSE) $\sigma^2(\hat{\mu}_l)$ can be decomposed as

$$\begin{aligned}\mathbb{E}[(\hat{\mu}_l - \mu_l)^2] &= \mathbb{E}[(\hat{\mu}_l - \tilde{\mu}_l)^2] + \mathbb{E}[(\tilde{\mu}_l - \mu_l)^2] + 2\mathbb{E}[(\hat{\mu}_l - \tilde{\mu}_l)(\tilde{\mu}_l - \mu_l)] \\ &= \mathbb{E}[(\hat{\mu}_l - \tilde{\mu}_l)^2] + \sigma^2(\tilde{\mu}_l).\end{aligned}$$

- As $n \to \infty$ the second term dominates the PMSE of $\hat{\mu}_l$, where

$$\sigma^2(\tilde{\mu}_l) = \boldsymbol{\kappa}_l^\top \boldsymbol{\kappa}_l - \boldsymbol{\kappa}_l^\top ((\boldsymbol{\Lambda}^\top \mathbf{V}^{-1} \boldsymbol{\Lambda}) \otimes \mathbf{I}_n) \boldsymbol{\kappa}_l + (\boldsymbol{c}_l - (\mathbf{I}_p \otimes \mathbf{X}^\top) \boldsymbol{o}_l)^\top \mathbf{Q} (\boldsymbol{c}_l - (\mathbf{I}_p \otimes \mathbf{X}^\top) \boldsymbol{o}_l)$$

  where $\boldsymbol{o}_l = (\mathbf{V}^{-1} \boldsymbol{\Lambda} \otimes \mathbf{I}_n) \boldsymbol{\kappa}_l$ and $\mathbf{Q} = \mathbf{V} \otimes (\mathbf{X}^\top \mathbf{X})^{-1}$.
- We approximate $\hat{\sigma}^2(\hat{\mu}_l) \approx \hat{\sigma}^2(\tilde{\mu}_l)$, based on by plugging the estimated variance components.

Basic Idea: CPI vs. SPI

- A cluster-level prediction interval (CPI) is an region $\mathcal{C}_{1-\alpha}$ such that it has probability $100(1-\alpha)\%$ of covering the general mixed parameter **for a single level**.

$$\mathbb{P}(\mu_l \in \mathcal{C}_{1-\alpha}) = 1 - \alpha, \forall l \in [L]$$

Basic Idea: CPI vs. SPI

- A cluster-level prediction interval (CPI) is an region $\mathcal{C}_{1-\alpha}$ such that it has probability $100(1-\alpha)\%$ of covering the general mixed parameter **for a single level**.

$$\mathbb{P}(\mu_l \in \mathcal{C}_{1-\alpha}) = 1-\alpha,\ \forall l \in [L]$$

- A simultaneous prediction interval (SPI) is a region $\mathcal{I}_{1-\alpha}$ such that it has probability $100(1-\alpha)\%$ of covering the general mixed parameters **across all levels**,

$$\mathbb{P}(\mu_l \in \mathcal{I}_{1-\alpha}, \forall l \in [L]) = 1-\alpha.$$

Constructing SPIs

- From the equation $\mathbb{P}(\mu_l \in \mathcal{I}_{1-\alpha}, \forall l \in [L]) = 1-\alpha$ suggests a maximum t-type statistic:

$$\begin{aligned}\alpha &= \mathbb{P}(\mu_l \notin \mathcal{I}_{1-\alpha}, \exists l \in [L]) \\ &= \mathbb{P}\left(\left|\frac{\hat{\mu}_l - \mu_l}{\hat{\sigma}(\hat{\mu}_l)}\right| \geqslant c_{1-\alpha}, \exists l \in [L])\right) \\ &= \mathbb{P}\left(\max_{l=1,2,\ldots,L}\left|\frac{\hat{\mu}_l - \mu_l}{\hat{\sigma}(\hat{\mu}_l)}\right| \geqslant c_{1-\alpha}\right).\end{aligned}$$

- Theoretically, the critical value $c_{1-\alpha}$ is then the $(1-\alpha)^{\text{th}}$-quantile of the maximum t-type statistic i.e.,

$$c_{1-\alpha} = \inf_t \{t \in \mathbb{R} : \mathbb{P}(\tau \leqslant t) \geqslant 1-\alpha\},$$

where $\tau = \max_{l=1,2,\ldots,L} |\tau_l|$ and $\tau_l = \frac{\hat{\mu}_l - \mu_l}{\hat{\sigma}(\hat{\mu}_l)}$.

- If $c_{1-\alpha}$ is known, the SPI can be constructed as

$$\mathcal{I}_{1-\alpha} = \bigtimes_{l=1}^{L} \left[\hat{\mu}_l \mp c_{1-\alpha}\hat{\sigma}(\hat{\mu}_l)\right].$$

Approximate the Critical Value $c_{1-\alpha}$

We propose three ways to approximate the $c_{1-\alpha}$: the **Bootstrap method**, the **Monte Carlo method**, and the **Bonferroni method**. In this talk, we will focus on the bootstrap method.

Approximate the Critical Value $c_{1-\alpha}$

We propose three ways to approximate the $c_{1-\alpha}$: the **Bootstrap method**, the **Monte Carlo method**, and the **Bonferroni method**. In this talk, we will focus on the bootstrap method.

1. **Fit the model** and obtain the EBLUP $\hat{\mu}_l$ and the estimated PMSE $\hat{\sigma}(\hat{\mu}_l)$ for levels $l = 1, 2, \ldots, L$.
2. **Create B bootstrap datasets** using the estimated parameters.

Approximate the Critical Value $c_{1-\alpha}$

We propose three ways to approximate the $c_{1-\alpha}$: the **Bootstrap method**, the **Monte Carlo method**, and the **Bonferroni method**. In this talk, we will focus on the bootstrap method.

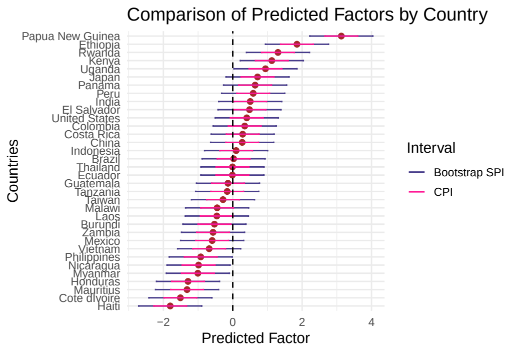
1. **Fit the model** and obtain the EBLUP $\hat{\mu}_l$ and the estimated PMSE $\hat{\sigma}(\hat{\mu}_l)$ for levels $l = 1, 2, \ldots, L$.
2. **Create B bootstrap datasets** using the estimated parameters.
3. For each bootstrap dataset, fit the model again to **obtain the bootstrap version of the maximum t-type statistics**.
4. Use the $1 - \alpha$ quantile of the bootstrap empirical distribution of the maximum t-type statistics $c_{1-\alpha}^{BS}$.

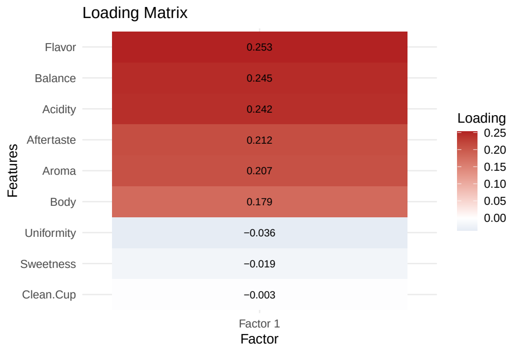
Approximate the Critical Value $c_{1-\alpha}$

We propose three ways to approximate the $c_{1-\alpha}$: the **Bootstrap method**, the **Monte Carlo method**, and the **Bonferroni method**. In this talk, we will focus on the bootstrap method.

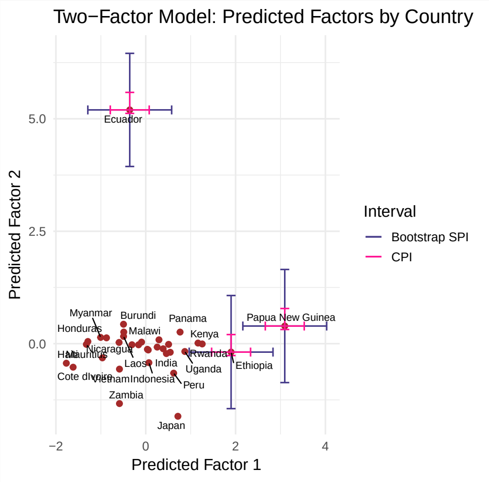
1. **Fit the model** and obtain the EBLUP $\hat{\mu}_l$ and the estimated PMSE $\hat{\sigma}(\hat{\mu}_l)$ for levels $l = 1, 2, \ldots, L$.
2. **Create** B **bootstrap datasets** using the estimated parameters.
3. For each bootstrap dataset, fit the model again to **obtain the bootstrap version of the maximum t-type statistics**.
4. Use the $1-\alpha$ quantile of the bootstrap empirical distribution of the maximum t-type statistics $c_{1-\alpha}^{BS}$.
5. **Construct SPI** using $c_{1-\alpha}^{BS}$ and the results from Step 1, $\mathcal{I}_{1-\alpha}^{BS} = \times_{l=1}^{L} \left[\hat{\mu}_l \mp c_{1-\alpha}^{BS} \hat{\sigma}(\hat{\mu}_l)\right]$.

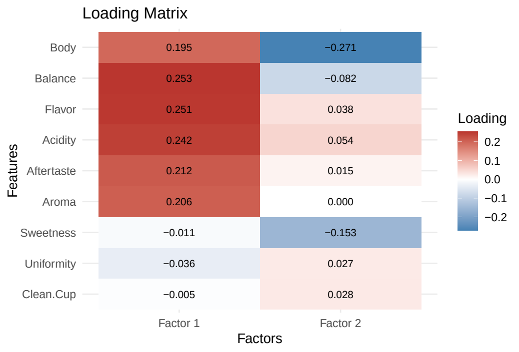
Application to Coffee Bean Data: 1-factor Model





Application to Coffee Bean Data: 2-factor Model





Discussion

- Simultaneous inference can be useful in the context of factor analytic models, if we want to perform (say) formal statistical inference on the latent variables.
- We proposed constructing SPI using a bootstrap approach. In simulations (not shown), we show **this generically performs better than** simpler Monte Carlo- and Bonferroni-based SPIs.
- In future work, we will run more simulations under more complicated models and general mixed parameters, as well as investigate simultaneous inference the case of non-Gaussian response factor analytic models.

Thank You!

Let's drink a cup of coffee and see whether my research aligns with your tastes!

Questions?

References

Corbeil, R. R. and Searle, S. R. (1976). Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, London ; New York.

Reluga, K., Lombardía, M.-J., and Sperlich, S. (2023). Simultaneous inference for linear mixed model parameters with an application to small area estimation. *International Statistical Review*, 91(2):193–217.

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance Components*. John Wiley & Sons, Hoboken, NJ, 2 edition.

West, M., Carvalho, C. M., others, Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., and Smith, A. F. (2003). Bayesian factor regression models in the “large p, small n” paradigm. In *Bayesian Statistics 7*, pages 733–742. Oxford University Press.

Appendix

Using the mixed model equation, we find

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{f}} - \boldsymbol{f} \end{bmatrix} \sim \mathcal{N}_{qp+nm}\left(\mathbf{0}_{qp+nm}, \left(\hat{\mathbf{A}}^\top \hat{\mathbf{K}}^{-1} \hat{\mathbf{A}} + \hat{\mathbf{B}}\right)^{-1}\right),$$

where $\hat{\mathbf{A}} := \begin{bmatrix} \mathbf{I}_p \otimes \mathbf{X} & \hat{\boldsymbol{\Lambda}} \otimes \mathbf{I}_n \end{bmatrix}$, $\hat{\mathbf{B}} := \begin{bmatrix} \mathbf{0}_{qp \times qp} & \mathbf{0}_{qp \times nm} \\ \mathbf{0}_{nm \times qp} & \mathbf{I}_{nm} \end{bmatrix}$, and $\hat{\mathbf{K}} := \hat{\boldsymbol{\Psi}} \otimes \mathbf{I}_n$.

Therefore for a Monte Carlo Procedure, for $s = 1, 2, \ldots, S$:

1. Sample $\begin{bmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{f}} - \boldsymbol{f} \end{bmatrix}^{(s)}$ from its asymptotic distribution.
2. Calculate the centred general mixed effect $(\mu_l - \mu)^{(s)} = \boldsymbol{c}_l^\top (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^{(s)} + \boldsymbol{\kappa}_l^\top (\tilde{\boldsymbol{f}} - \boldsymbol{f})^{(s)}$.
3. Calculate $\tau_{MC}^{(s)} := \max_l \frac{|(\mu_l - \mu)^{(s)}|}{\hat{\sigma}(\hat{\mu}_l)}$.

Order $\tau_{MC}^{(s)}$ for all $s \in [S]$ and find the upper $1 - \alpha$ quantile as the empirical critical value $c_{1-\alpha}^{\mathsf{MC}}$.

Bonferroni Procedure

In a Bonferroni procedure, we directly approximate the critical value as $c^{\mathsf{BO}}_{1-\alpha} = \Phi^{-1}\left(1 - \frac{\alpha}{2L}\right)$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

Simulation Settings

- It is of interest to see how SPIs constructed by different methods behave under different n, p, q, m, along with the true variance explained (VE) by the loadings

$$\mathrm{VE} = \frac{\mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)}{\mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})}.$$

- We ran 200 simulation rounds for different combinations of $(n, p, q, m, \mathrm{VE})$.
- The **target** of the intervals was all the latent variables in a one-factor model i.e., $L = n$.
- We used $B = 500$ bootstrap datasets and $S = 2000$ Monte Carlo samples.

Simulation Results

Table 1: Summary of SPI methods across different simulation settings ($p = 20$, $m = 1$, $q = 3$). The nominal level of intervals is 0.95.

Setting	Method (Intervals)	ECP	Mean Len.	Avg Var-Width ($\times 10^{-2}$)
$n = 500$, VE$= 0.7$	Bootstrap SPI	0.949	1.30	0.13
	Monte Carlo SPI	0.913	1.26	0.11
	Bonferroni SPI	0.924	1.27	0.10
	CPI	0	0.57	0.03
$n = 200$, VE$= 0.7$	Bootstrap SPI	0.965	1.43	0.27
	Monte Carlo SPI	0.915	1.35	0.22
	Bonferroni SPI	0.930	1.37	0.19
	CPI	0	0.57	0.09
$n = 500$, VE$= 0.5$	Bootstrap SPI	0.940	1.82	0.31
	Monte Carlo SPI	0.935	1.78	0.25
	Bonferroni SPI	0.935	1.79	0.25
	CPI	0	0.85	0.07

Takeaways from Simulations

- **Bootstrap SPI performs well**, achieving empirical coverage closest to the nominal 0.95 level.
- **As sample size increases**, Monte Carlo and Bonferroni SPIs are still slightly undercovered.
- **Lower VE increases uncertainty**, resulting in longer interval lengths and larger average variance of interval widths.
- **CPI consistently fails to provide valid coverage**, highlighting the necessity of simultaneous inference.