Data Synthesis with Selective Differential Privacy

Fang Liu

Applied and Computational Mathematics and Statistics
University of Notre Dame

acknowledgement: Gina Mannino

JSM 2024, Portland, OR

Taking points

- Differential Privacy (DP)
- Selective DP
- Data synthesis with DP and Selective DP
- Examples

Differential Privacy (DP) [Dwork et al., 2006a,b]

- A mathematical framework to provide privacy guarantees to individuals in a dataset when releasing info from the dataset: A randomized algorithm $\mathcal{M}$ is of (ϵ, δ)-DP if for all neighboring data sets (x, x') differing by one record and $\forall \mathcal{S} \subseteq \text{image}(\mathcal{M})$,

$$\Pr(\mathcal{M}(x) \in S) \leq e^{\epsilon} \Pr(\mathcal{M}(x') \in \mathcal{S}) + \delta$$

- $\epsilon > 0$ and $\delta \in [0, 1)$ are the privacy loss or budget parameters.
- DP mechanisms, when ϵ and δ are small, guarantee that their outputs do not reveal personal info about any individual in the input data.
- When $\delta = 0$, (ϵ, δ)-DP reduces to pure ϵ-DP; when $\delta \neq 0$, it can be intuitively interpreted as the probability that ϵ pure DP is violated.
- There are various extensions of (ϵ, δ)-DP (e.g. zero-concentrated DP [Bun and Steinke, 2016], Rényi DP [Mironov, 2017], Gaussian DP [Dong et al., 2022]).

DP

In the classical DP framework, whether (ϵ, δ)-DP or its extensions, the privacy guarantee is uniform

- for all users/individuals in the input data
- for all attributes/variables in the input data

This "one size fits all" concept does not consider, in the real world, that

- different individuals may have different privacy expectations for their personal data.
- different variables may have different privacy/utility trade-off needs.

Selective DP

- Selective DP is an extension of DP that allow privacy loss to various across users or attributes in a dataset.
- Development and application selective DP concepts and procedures can potentially improve the utility of outputs with DP guarantees, compared to the uniform DP framework.

User-Selective (US) DP

- Each individual may have their own privacy requirement.
- Roughly, three groups (Westin privacy segmentation): fundamentalists/conservative, pragmatists/moderate, and unconcerned/liberal [Harris et al., 1998, Jensen et al., 2005, Jorgensen et al., 2015]
- Different privacy loss parameters can be specified for different groups, depending the needs in each group.
- Savings in privacy budget in the non-F/C groups may help increase utility.
- Existing work on US DP:
  - personalized DP [Jorgensen et al., 2015, Niu et al., 2021, Acharya et al., 2024]
  - one-sided DP [Doudalis et al., 2017]
  - per-instance DP [Wang, 2018]
  - heterogeneous DP (two groups) [Chaudhuri and Courtade, 2023]

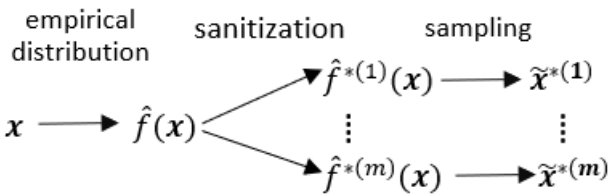
Attribute-Selective (AS) DP

- Some attributes may contain less sensitive information or are public (e.g. QID)
- Some are more sensitive but oftentimes also of the primary interest and their accuracy needs to be preserved as much as possible (e.g. health status).
- Most existing work on AS-DP tweaks the definition of "neighboring" sets.
  - differ by one attribute in at most one user [Kifer and Machanavajjhala, 2011, Kenthapadi et al., 2013, Ahmed et al., 2016, Asi et al., 2019]
  - privacy weights $w_j \in [0, 1]$ for different attribute j [Alaggan et al., 2015]
  - partial DP [Ghazi et al., 2022]
  - language models: sensitive vs non-sensitive tokens, neighbors are defined for sensitive tokens only [Shi et al., 2022]

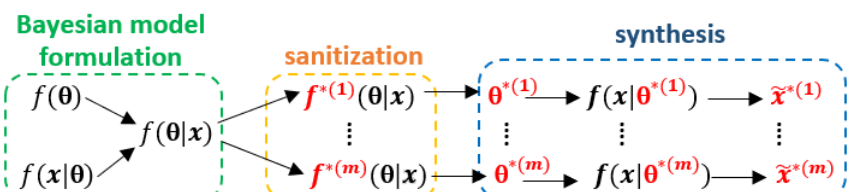
Data Synthesis with DP

Integration of DP guarantees during the data synthesis process.

- *"non-parametric"* (e.g, histograms, KDE, deep generative models):



- *parametric*: e.g.



Data Synthesis with Selective DP

- US DP: generative models based on data from different privacy groups are sanitized separately
  - if the users' data in different groups follow the same pop'n dist'n, we may use weighted averages of sanitized statistics across the groups to formulate one DP generative model;
  - o.w., group-specific synthesis should be used and the generated data subsets can then be merged and release.
- AS DP
  - independent sanitization if attributes are mutually independent, with proper privacy loss accounting over multiple attributes
  - o.w., simultaneous or sequential sanitization; it is critical to ensure relationships among attributes are not distorted during sanitization and synthesis.
- doubly-selective DP: data synthesis with US- and AS- DP

Toy Examples

US-DP: 3 user groups with different privacy expectations

- $X \sim \mathcal{N}(0, 1)$, $n = [100, \ldots, 1200]$
- 3 privacy groups in SDP: 55% conservative ($\epsilon_1 = 0.1$), 35% moderate ($\epsilon_2 = 2$), 10% liberal ($\epsilon_3 = 8$) [Jorgensen et al., 2015]
- Unif-DP: $\epsilon = 0.1$

AS-DP: 2 attributes with different privacy requirements

- $X_1 \sim \mathcal{N}(0, 1); X_2 \sim \mathcal{N}(1, 1); \rho = (0, 0.2, 0.5, 0.8); n = [100, \ldots, 2000]$
- X_1 (sensitive): $\epsilon_1 = 0.5$; X_2 (somewhat insensitive): $\epsilon_2 = 2$
- Unif-DP: $\epsilon = 0.5$

1000 repeats and $m = 3$ synthetic datasets (multiple synthesis [Liu, 2022]) per repeat in both simulations settings.

Toy Examples: results

- evaluation metrics; bias, RMSE, CP (coverage probability of 95% CI) for the means and variances of the variables and correlation in the AS-DP example.
- Figures are in the back-up slides
- US-DP keeps Conservative happy with the same level of privacy ($\epsilon_1 = 0.1$) as in uniform DP while offering better utility from relaxing the privacy guarantees in the other 2 groups.
- AS-DP keeps the sensitive attribute protected with the same level privacy ($\epsilon_1 = 0.5$) as in uniform DP while offering better utility for outputs that involve X_2 from relaxing its privacy guarantees.

	mean	variance	correlation
X_1 ($\epsilon_1 = 0.5$)	similar	similar	improved
X_2 ($\epsilon_2 = 2$)	improved	improved	

Conclusions and Discussion

Compared to "uniform" DP, selective DP

- is more reflective of the real-life privacy perceptions, expectations, and needs
- provides a framework to potentially improve the utility of sanitized outputs

Research on selective DP is still scarce and can be difficult

- **conceptual challenges**: new selective DP notions need to satisfy the basic properties of DP, such as privacy loss composition, immunity to post-processing.
- **scalability/generality**: selective DP mechanisms should be provided, when new concepts are defined, and be general enough to be applicable to a variety of queries/analyses

References

K. Acharya, F. Boenisch, R. Naidu, and J. Ziani. Personalized differential privacy for ridge regression, 2024. URL http://arxiv.org/abs/2401.17127.

F. Ahmed, A. X. Liu, and R. Jin. Social graph publishing with privacy guarantees. In *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, pages 447–456, 2016. doi: 10.1109/ICDCS.2016.74. URL https://ieeexplore.ieee.org/abstract/document/7536543. ISSN: 1063-6927.

M. Alaggan, S. Gambs, and A.-M. Kermarrec. Heterogeneous differential privacy, 2015. URL http://arxiv.org/abs/1504.06998.

H. Asi, J. Duchi, and O. Javidbakht. Element level differential privacy: The right granularity of privacy, 2019. URL http://arxiv.org/abs/1912.04042.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography: 14th International Conference, TCC 2016-B, Beijing, China, October 31-November 3, 2016, Proceedings, Part I*, pages 635–658. Springer, 2016.

S. Chaudhuri and T. A. Courtade. Mean estimation under heterogeneous privacy: Some privacy can be free, 2023. URL http://arxiv.org/abs/2305.09668.

J. Dong, A. Roth, and W. J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

S. Doudalis, I. Kotsogiannis, S. Haney, A. Machanavajjhala, and S. Mehrotra. One-sided differential privacy, 2017. URL http://arxiv.org/abs/1712.05888.

References (cont.)

C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer, 2006a.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006b.

B. Ghazi, R. Kumar, P. Manurangsi, and T. Steinke. Algorithms with more granular differential privacy guarantees, 2022. URL `http://arxiv.org/abs/2209.04053`.

. Harris, . et, and . al. E-commerce & privacy: what net users want., 1998. URL `http://www.pandab.org/ecommercesurvey.html`.

C. Jensen, C. Potts, and C. Jensen. Privacy practices of internet users: Self-reports versus observed behavior. *International Journal of Human-Computer Studies*, 63(1-2):203–227, 2005.

Z. Jorgensen, T. Yu, and G. Cormode. Conservative or liberal? personalized differential privacy. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1023–1034, 2015. doi: 10.1109/ICDE.2015.7113353. URL `https://ieeexplore.ieee.org/document/7113353`. ISSN: 2375-026X.

K. Kenthapadi, A. Korolova, I. Mironov, and N. Mishra. Privacy via the johnson-lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1), Aug. 2013. ISSN 2575-8527. doi: 10.29012/jpc.v5i1.625. URL `http://dx.doi.org/10.29012/jpc.v5i1.625`.

References (cont.)

D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011. ISBN 978-1-4503-0661-4. doi: 10.1145/1989323.1989345. URL https://dl.acm.org/doi/10.1145/1989323.1989345.

F. Liu. Model-based differentially private data synthesis and statistical inference in multiply synthetic differentially private data. *Transactions on Data Privacy*, 15(2), 2022.

I. Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

B. Niu, Y. Chen, B. Wang, Z. Wang, F. Li, and J. Cao. AdaPDP: Adaptive Personalized Differential Privacy. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, May 2021. doi: 10.1109/INFOCOM42981.2021.9488825. URL https://ieeexplore.ieee.org/document/9488825. ISSN: 2641-9874.

W. Shi, A. Cui, E. Li, R. Jia, and Z. Yu. Selective differential privacy for language modeling, 2022. URL http://arxiv.org/abs/2108.12944.

Y.-X. Wang. Per-instance differential privacy, 2018. URL http://arxiv.org/abs/1707.07708.

This page is intentionally left blank.

Inference based on Synthetic Data

multiple synthesis (MS) to account for sanitization and synthesis randomness.

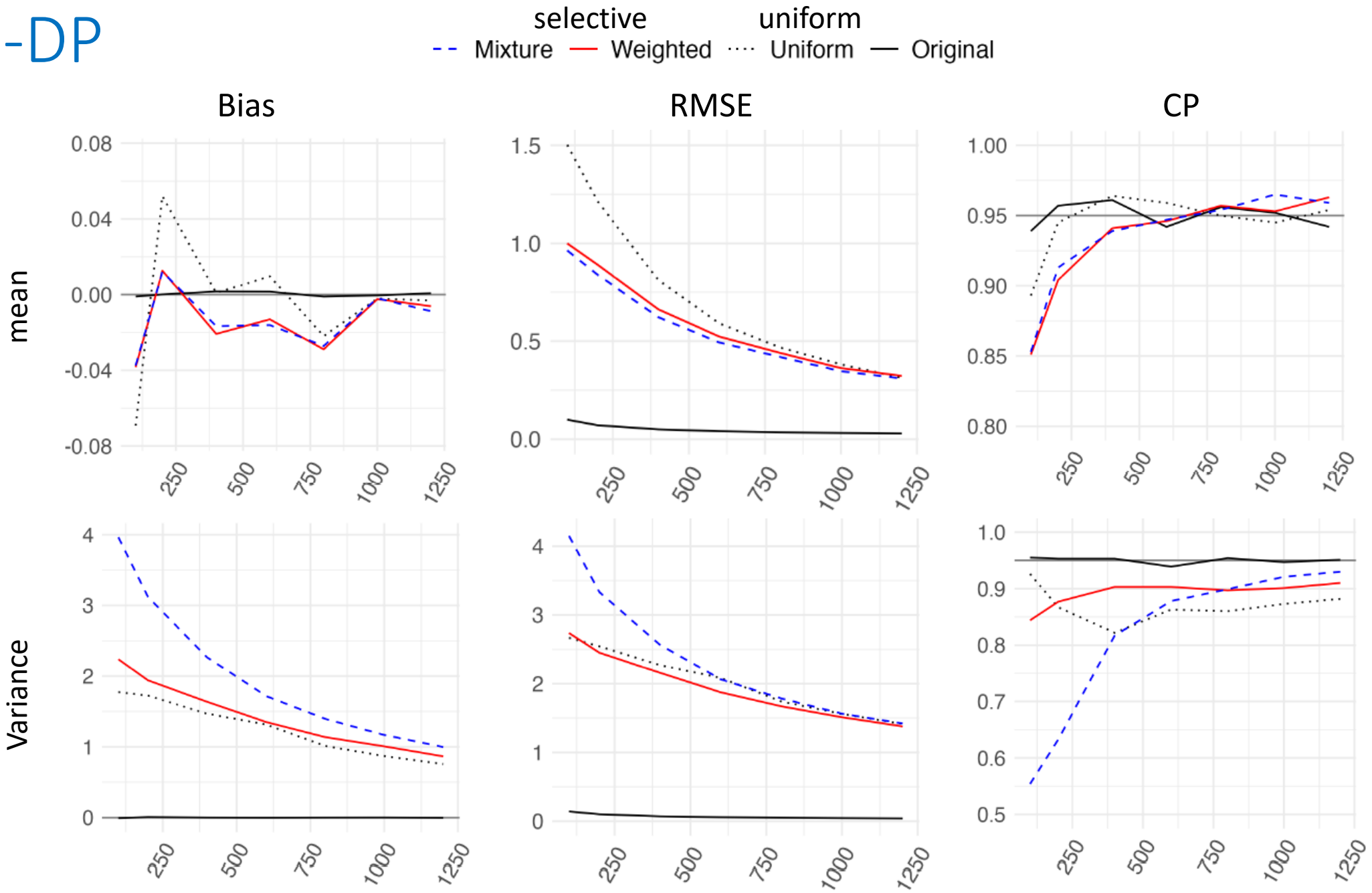
- $\bar{\beta}^* = m^{-1} \sum_{j=1}^{m} \hat{\beta}^{*(j)}$
- its variance is estimated by

$$u = \varpi + m^{-1} b, \text{ where}$$
$$\varpi = m^{-1} \sum_{j=1}^{m} \hat{v}^{*(j)} \text{ and } b = (m-1)^{-1} \sum_{j=1}^{m} (\hat{\beta}^{*(j)} - \bar{\beta}^*)^2;$$

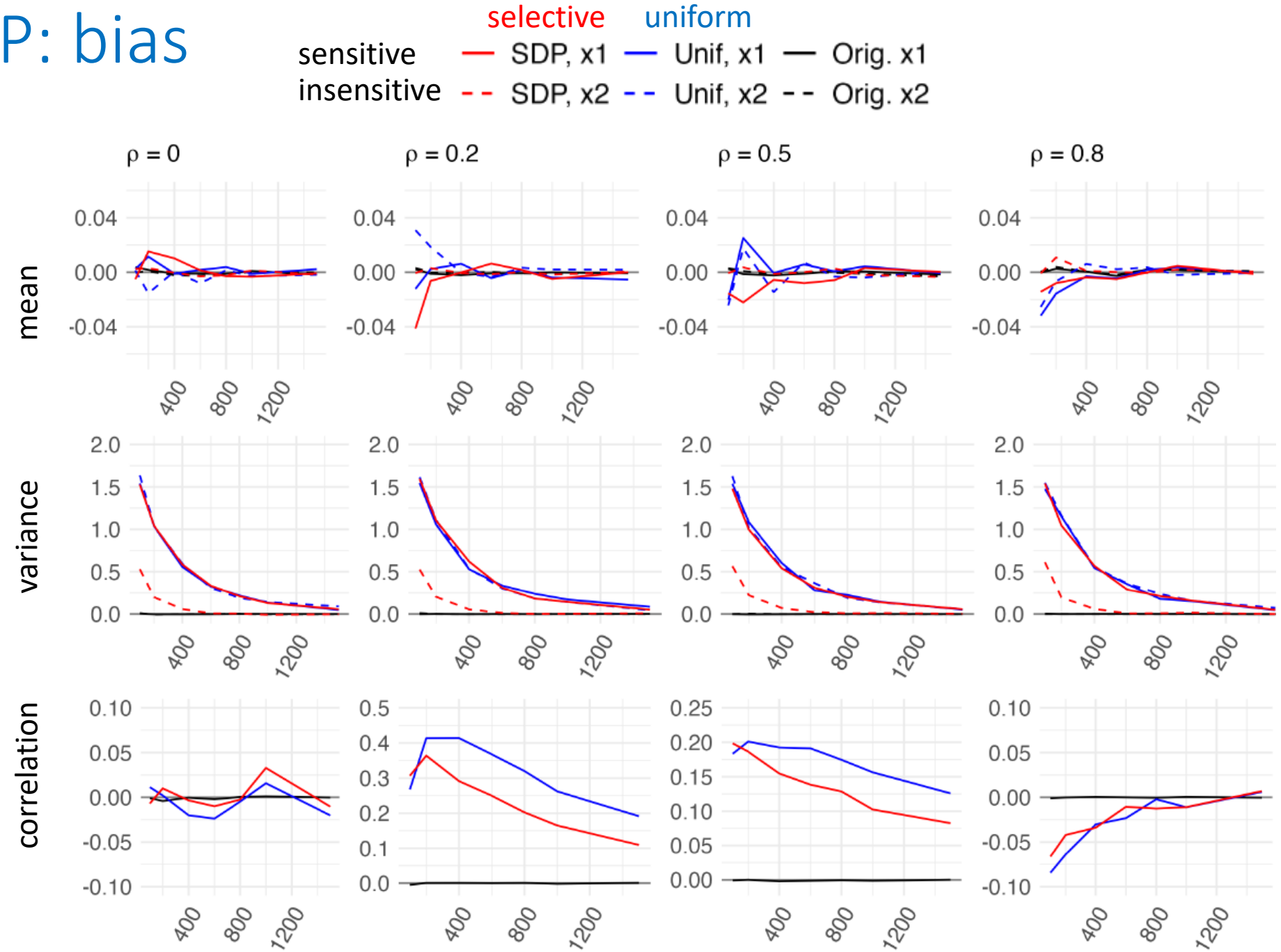
  ϖ is the average within-set variance and b is the between-set variance comprised of the variability due to sanitization and that due to synthesis.

- Inference of β is based on $t_\nu(\bar{\beta}^*, m^{-1}b + \varpi)$ with the degree of freedom $\nu = (m-1)(1 + m\varpi/b)^2$.
- $m \in [3, 5]$ is recommended [Liu, 2022]

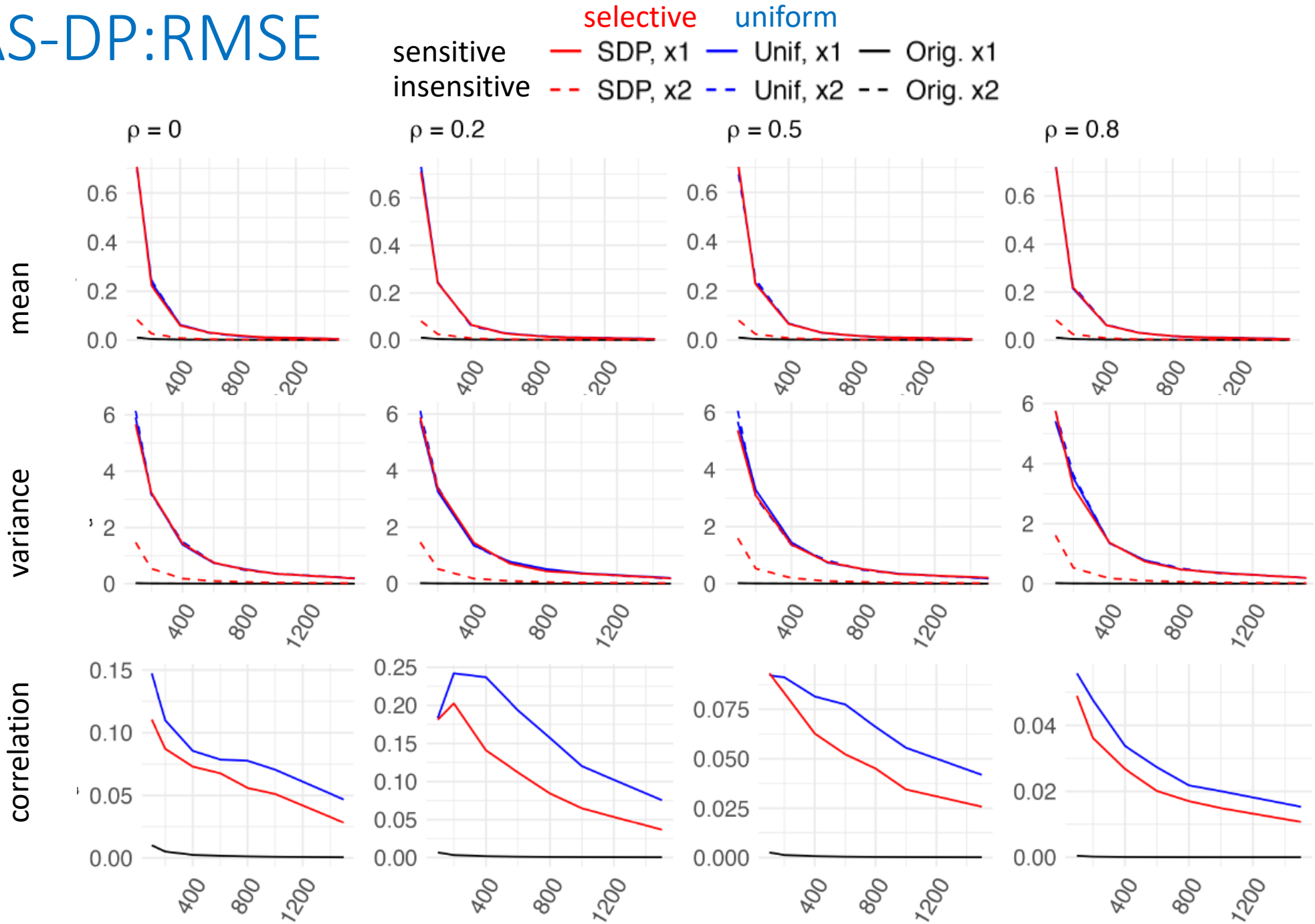
US-DP


selective
uniform
Mixture
Weighted
Uniform
Original
Bias
RMSE
CP
mean
Variance
0.08
0.04
0.00
-0.04
-0.08
1.5
1.0
0.5
0.0
1.00
0.95
0.90
0.85
0.80
4
3
2
1
0
1.0
0.9
0.8
0.7
0.6
0.5
250
500
750
1000
1250

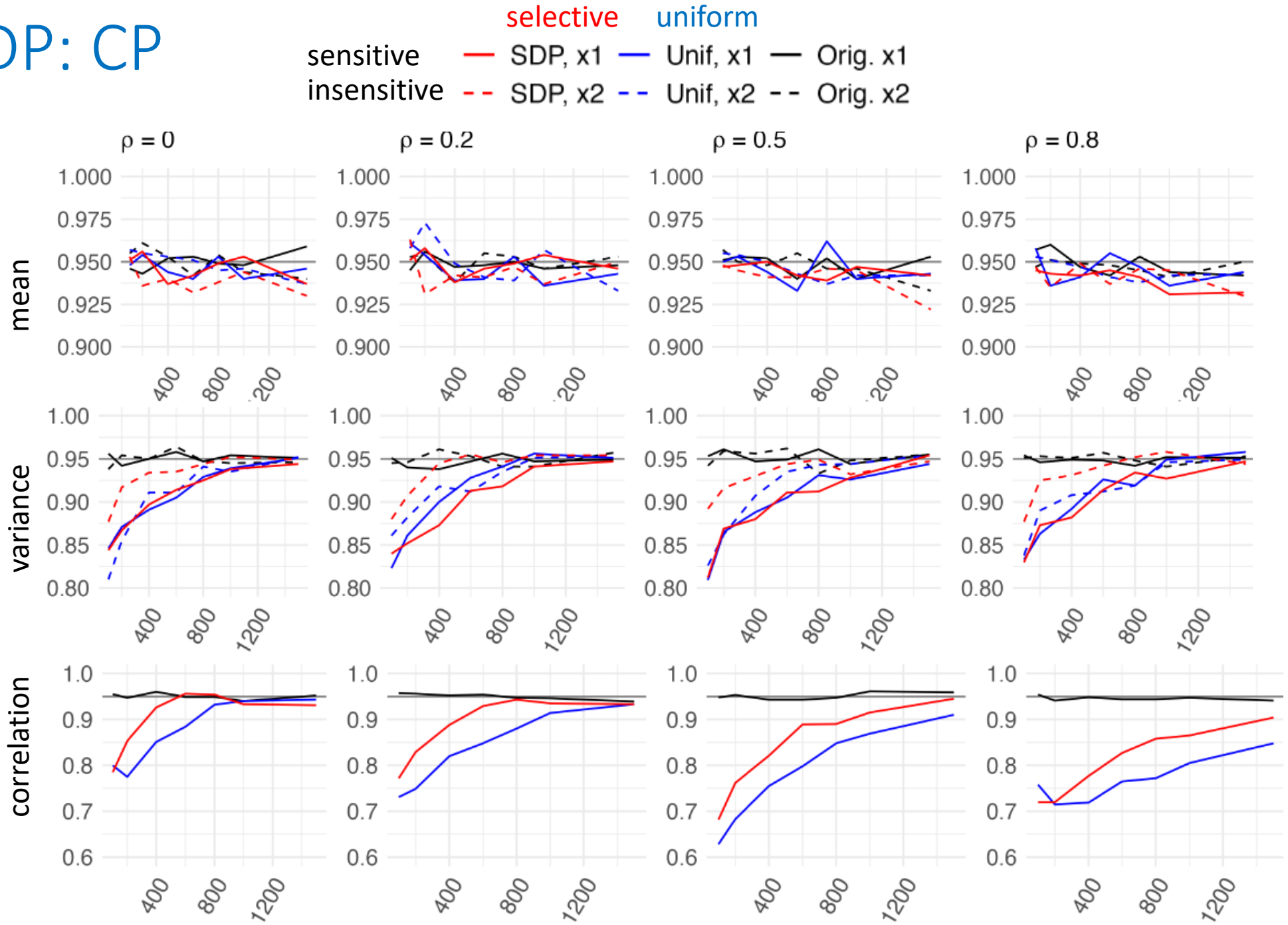
AS-DP: bias


selective
uniform
sensitive
insensitive
SDP, x1
Unif, x1
Orig. x1
SDP, x2
Unif, x2
Orig. x2
ρ = 0
ρ = 0.2
ρ = 0.5
ρ = 0.8
mean
variance
correlation
0.04
0.00
-0.04
2.0
1.5
1.0
0.5
0.0
0.10
0.05
-0.05
-0.10
0.4
0.3
0.2
0.1
0.25
0.20
0.15
400
800
1200

AS-DP:RMSE


selective
uniform
sensitive
insensitive
SDP, x1
Unif, x1
Orig. x1
SDP, x2
Unif, x2
Orig. x2
ρ = 0
ρ = 0.2
ρ = 0.5
ρ = 0.8
mean
variance
correlation

AS-DP: CP


selective
uniform
sensitive
insensitive
SDP, x1
Unif, x1
Orig. x1
SDP, x2
Unif, x2
Orig. x2
ρ = 0
ρ = 0.2
ρ = 0.5
ρ = 0.8
mean
variance
correlation