Biometrika (2020), **xx**, x, *pp*. 1–4
Printed in Great Britain

Discussion of 'Network cross-validation by edge sampling'

BY J. LEI
Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, U.S.A
jinglei@andrew.cmu.edu

AND K. LIN
Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, U.S.A
kevinl1@andrew.cmu.edu

1. INTRODUCTION

We congratulate the authors for a nice contribution to model selection and assessment for complex network data. Li et al. (2020) enrich our tools for network data analysis, bringing together different fields such as low rank matrix estimation, cross-validation, and network modeling. This discussion focuses on the V-fold variant of the edge cross-validation (ECV) method and its combination with other recently developed cross-validation methods and network models.

2. V-FOLD ECV AND COMPARISON WITH NCV

The V-fold cross-validation is perhaps the most popular variant of cross-validation. It partitions the training set into V disjoint folds, and performs sample splitting validation by leaving out each fold as the validation set and using the remaining $V-1$ folds for estimation. The final validated risk is obtained by aggregating the sample splitting validation risks over the folds. This idea can be extended to the ECV framework straightforwardly as mentioned in Section 2.2 of Li et al. (2020).

A related cross-validation method for network and other low rank matrix estimation problems is the "network cross-validation" (NCV) method developed by Chen & Lei (2018), where the sample splitting is realized by holding out a diagonal submatrix of the observed adjacency matrix. The simulation study in Li et al. (2020) reports some empirical comparison between ECV and NCV under various settings, and suggests that ECV may be preferable in more challenging settings. Given the nature of these algorithms, it is hard to imagine a difference in rates of convergence of estimation error, and it seems more likely the estimation errors differ at a constant factor level. A possible explanation would be the more efficient use of validation sample points in ECV. In V-fold ECV, each node pair appears exactly once as a validation sample point, so that the validation sample size is roughly $n^2/2$. On the other hand, in V-fold NCV, only the node pairs in the diagonal sub-blocks are used for validation. Hence, the validation sample size for V-fold NCV is roughly $n^2/(2V)$.

3. CROSS-VALIDATION WITH CONFIDENCE

Li et al. (2020) provide some theoretical guarantees, for example, in Theorems 2 and 3, for the ECV model selection. Such results say that with high probability, the selected model order will not be lower than the true value. In other words, ECV does not underfit. There is no guarantee about overfitting. This agrees with our understanding about cross-validation – although it

Table 1. *Model selection performance of ECV and ECVC with $n = 100$ and equal-sized communities with sparsity $\rho = n^{-r}$. Reported are counts out of 100 independent repetitions.*

		Correct		Overfit		Underfit	
K	r	ECV	ECVC	ECV	ECVC	ECV	ECVC
3	0	98	100	2	0	0	0
3	0.05	99	100	1	0	0	0
3	0.1	98	99	2	0	0	1
3	0.15	100	100	0	0	0	0
3	0.2	100	98	0	2	0	0
4	0	98	100	2	0	0	0
4	0.05	98	100	2	0	0	0
4	0.1	94	99	6	1	0	0
4	0.15	69	68	21	9	10	23
4	0.2	18	9	20	5	62	86
5	0	97	99	3	1	0	0
5	0.05	62	74	38	24	0	2
5	0.1	20	22	57	27	23	51
5	0.15	16	3	12	3	72	94
5	0.2	1	0	0	0	99	100

effectively avoids underfitting, it is not guaranteed to prevent overfitting unless the majority of sample points are used for validation (Shao, 1993; Zhang, 1993; Yang, 2007). The intuition is that a slightly overfitting model will have nearly the same predictive accuracy as the true model, and can have lower cross-validated risk due to the sampling randomness.

Recently Lei (2019) developed a method called cross-validation with confidence (CVC) that takes into account the randomness in the cross-validated risk. Instead of simply comparing the cross-validated risks, it uses cross-validated test errors as input data and approaches the model selection problem by testing the statistical hypothesis that a given candidate model is the best model, and outputs a confidence set containing the best candidate model with guaranteed probability under certain regularity conditions. If a single model is to be selected, one can choose the most parsimonious model in the confidence set. We apply this rule to our simulations below.

Here we combine the 5-fold ECV method with CVC at nominal type I error level 0.05, which we call ECVC, and empirically examine its performance in a simple stochastic block model setting. We use $n = 100$ nodes, with K equal sized communities for different values of $K \in \{3, 4, 5\}$. The community-wise edge probability matrix $B = B_0/n^r$ for $r \in \{0, 0.05, 0.1, 0.15, 0.2\}$ and B_0 has diagonal entries 0.8 and off-diagonal entries 0.2. The candidate set is $\{1, 2, 3, 4, 5, 6\}$. We use weighted spectral clustering as the base estimator, where each eigenvector is multiplied by the square root of its corresponding absolute eigenvalue before applying k-means clustering. The code is available at `https://github.com/linnylin92/edgeCV`.

We repeat the experiment 100 times for each combination of K and r. The results are reported in Table 1, which includes the number of times each algorithm correctly selects the value of K, selects a value larger than the truth (overfitting), or selects a value smaller than the truth (underfitting). The main observation is that when underfitting is not a big concern for ECV, ECVC is able to improve the model selection accuracy. When ECV underfits, ECVC cannot improve the performance as it underfits even more.

Table 2. *Model selection performance of ECV and ECVC in tensor setting with $n = 100$, $p = 10$, and equal-sized communities with sparsity $\rho = n^{-r}$. Reported are counts out of 100 independent repetitions.*

		Correct		Overfit		Underfit	
K	r	ECV	ECVC	ECV	ECVC	ECV	ECVC
3	0.3	100	100	0	0	0	0
3	0.35	98	99	2	1	0	0
3	0.4	98	100	2	0	0	0
3	0.45	96	99	4	1	0	0
3	0.5	94	100	6	0	0	0
4	0.3	98	100	2	0	0	0
4	0.35	95	99	5	1	0	0
4	0.4	90	98	10	2	0	0
4	0.45	47	68	51	30	2	2
4	0.5	41	23	20	7	39	70
5	0.3	96	100	4	0	0	0
5	0.35	71	78	29	22	0	0
5	0.4	42	50	39	16	19	34
5	0.45	11	2	8	2	81	96
5	0.5	0	0	0	0	100	100

4. MULTI-LAYER NETWORKS

Multi-layer networks have been a very active research topic recently (Paul & Chen, 2017; Bhattacharyya & Chatterjee, 2018; Pensky et al., 2019; Lei et al., 2019; Arroyo et al., 2019). ECV can be implemented in the multi-layer stochastic block model or other variants. For binary data, we can bypass the matrix completion step by treating the incomplete training data as a realization from a model with downscaled edge probabilities.

Table 2 reports simulation results for ECV and ECVC in the same stochastic block model setting as in the previous section, except that there are $p = 10$ independent layers of the stochastic block model. Due to the increased sample size, we consider sparser layers with $r \in \{0.3, 0.35, 0.4, 0.45, 0.5\}$. The base estimator is the multi-layer maximum likelihood estimator fitted using a greedy algorithm given in Lei et al. (2019). Again, ECVC can improve the accuracy when most model selection mistakes for ECV come from overfitting.

REFERENCES

ARROYO, J., ATHREYA, A., CAPE, J., CHEN, G., PRIEBE, C. E. & VOGELSTEIN, J. T. (2019). Inference for multiple heterogeneous networks with a common invariant subspace. *arXiv preprint arXiv:1906.10026* .

BHATTACHARYYA, S. & CHATTERJEE, S. (2018). Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594* .

CHEN, K. & LEI, J. (2018). Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association* **113**, 241–251.

LEI, J. (2019). Cross-validation with confidence. *Journal of the American Statistical Association* **to appear**.

LEI, J., CHEN, K. & LYNHC, B. (2019). Consistent community detection in multi-layer network data. *Biometrika* **to appear**.

LI, T., LEVINA, E. & ZHU, J. (2020). Network cross-validation by edge sampling. *Biometrika* .

PAUL, S. & CHEN, Y. (2017). Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *arXiv preprint arXiv:1704.07353* .

PENSKY, M. et al. (2019). Dynamic network models and graphon estimation. *The Annals of Statistics* **47**, 2378–2403.

SHAO, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association* **88**, 486–494.

YANG, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics* , 2450–2473.
ZHANG, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics* , 299–313.