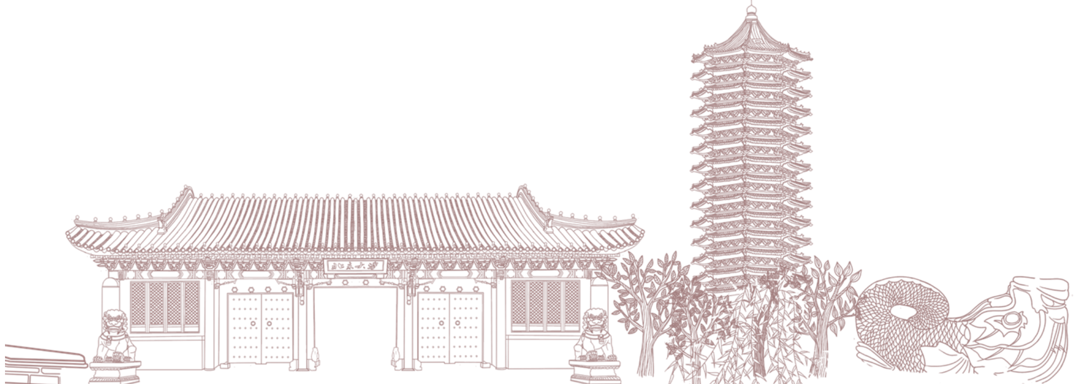
Estimating graphical models for count data with applications to single-cell gene network

Feiyi Xiao, Junjie Tang, Huaying Fang, Ruibin Xi

Speaker : Feiyi Xiao

School of Mathematical Science, Peking University

10/10/2022



Background

Gaussian Graphical model

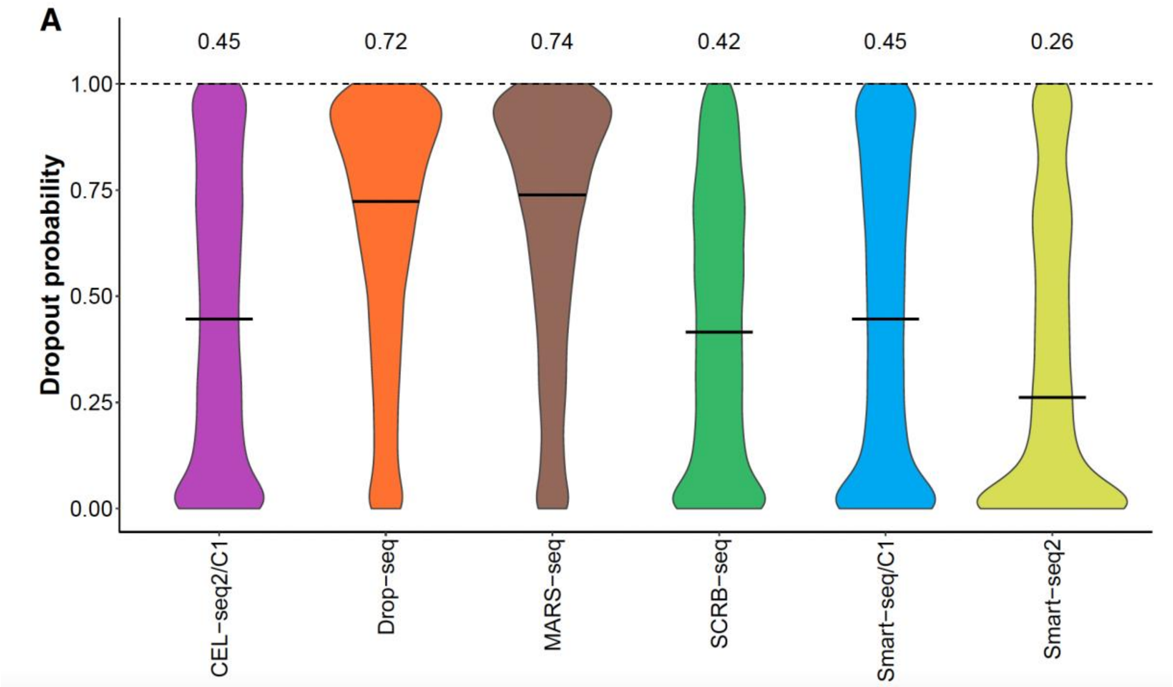
- In Gaussian graphical model $\boldsymbol{x} \sim \mathrm{N}_p(0, \Sigma)$:
  - Precision matrix: $\Theta \ = \Sigma^{-1}$.
  - Nonzero elements of Θ correspond to edges in Gaussian graphical model. If $\boldsymbol{x} \sim N_p(0, \Sigma), \Theta_{\mathrm{ij}} = 0$ iff $x_i \perp \ x_j \ |\{x_k, \ k \neq \ i, j\}$ (Wittaker, 1990).
  - We can impose sparsity on Θ to study the Gaussian graphical model.
- glasso: Yuan and Lin (2006) and Friedman et al. (2007) proposed to estimate Θ by minimizing:

$$-\log \det{(\Theta)} + \ \mathrm{tr}\left(\Theta\hat{\Sigma}\right) + \lambda \, |\Theta|_{1,off}$$

Background

Challenges in Analyzing scRNA–seq Data

- High dimensional and large number of cells.
- Essential count data, many methods developed for continuous data would not work well.
- High dropout (ratio of zeros) and increased variation.

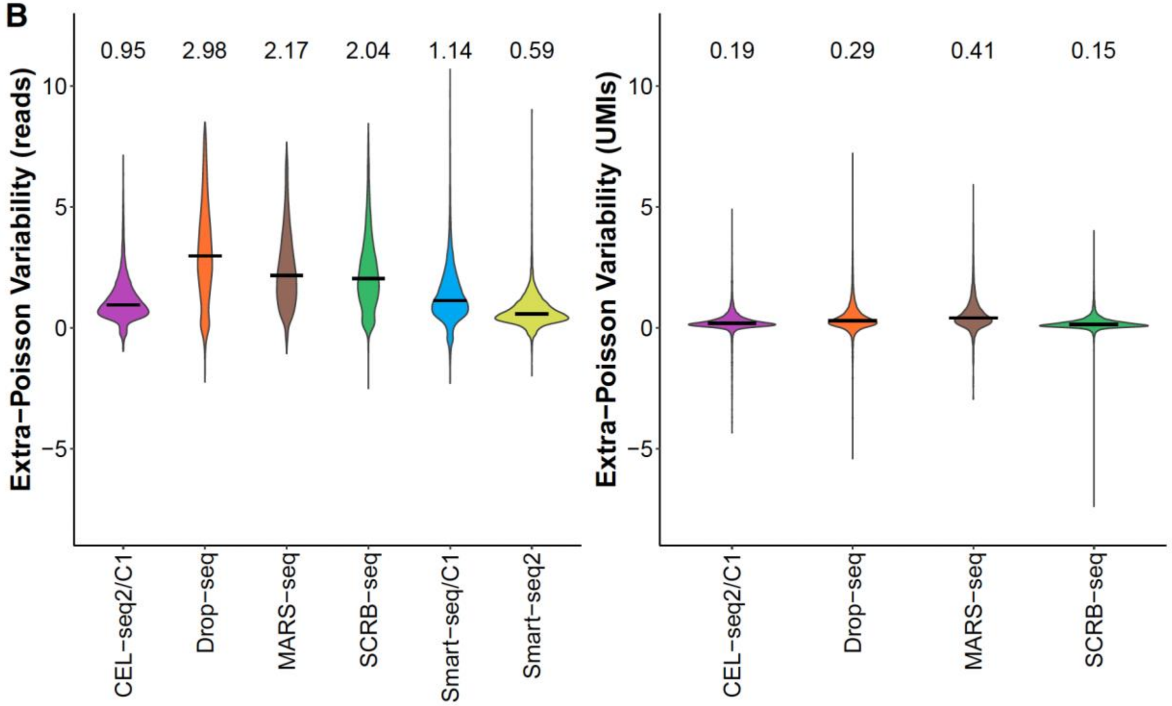


Ziegenhain et al. 2017 Molecular Cell

Background

Challenges in Analyzing scRNA–seq Data

- High dimensional and large number of cells.
- Essential count data, many methods developed for continuous data would not work well.
- High dropout (ratio of zeros) and increased variation.



Ziegenhain et al. 2017 Molecular Cell

Method

Poisson Log–normal (PLN) model

- scRNA-seq data with n cells and p genes.
- Observed expression : $\boldsymbol{Y_i} = \left(Y_{i1}, \dots, Y_{ip}\right)^T$.
- Underlying true expressions: $\boldsymbol{X_i} = \left(X_{i1}, \dots, X_{ip}\right)^T$.
- S_i: library size.
- Network: precision matrix Θ^*.
- The PLN model for scRNA-seq data:

$$\boldsymbol{Y_i} \,|\boldsymbol{X_i} \sim \prod_{j=1}^{p} \text{Poisson}\left(S_i X_{ij}\right)$$

$$\log(\boldsymbol{X_i}) \sim \text{N}(\boldsymbol{\mu}^*, (\Theta^*)^{-1})$$

Method

PLNet Procedure

- Estimate the covariance matrix $\Sigma^* = (\Theta^*)^{-1}$ using maximum marginal likelihood estimator (MMLE).
  - Newton-Raphson algorithm.
  - Initial values: moment estimator $\tilde{\boldsymbol{\mu}}^m$ and $\tilde{\Sigma}^m$.
  - Positive semi-definite projection.
- Plug-in the MMLE $\hat{\Sigma}$ to the lasso penalized D-trace loss (Zhang and Zou (2014)) to estimate Θ^*:

$$\hat{\Theta} = \mathrm{argmin}_{\Theta \succcurlyeq 0} \ \frac{1}{2}\mathrm{tr}(\hat{\Sigma}\Theta^2) \ - \mathrm{tr}(\Theta) + \lambda_n ||\Theta||_{1,\mathrm{off}}.$$

- Tuning parameter λ_n selection: approximated Bayesian information criterion (BIC):

$$\left\| \frac{1}{2}\left(\hat{\Theta}\hat{\Sigma} + \hat{\Sigma}\hat{\Theta}\right) - \ I_p \right\|_F \ + \frac{\log(n)}{n} \left\| \hat{\Theta} \right\|_0$$

Main Theoretical Results

Consistency Theory

Theorem 1 (Rate of convergence and sign consistency)

Under some mild conditions, there exist positive constants A, B, C, such that for some $\eta > 2$, if $n > C_p C$, choosing $\lambda_n = 12\gamma^{-1}\left(k_\Sigma k_\Gamma^2 + k_\Gamma\right) C_p^{1/2} n^{-\frac{1}{2}}$, then with probability $1 - p^{2-\eta}$,

$$\left\|\widehat{\Theta} - \Theta\right\|_\infty \leq \left(12\gamma^{-1}\left(k_\Sigma k_\Gamma^3 + k_\Gamma^2\right) + 5dk_\Gamma^2\right) C_p^{1/2} n^{-\frac{1}{2}},$$

and $\widehat{\Theta}$ recovers all zeros and nonzeros in Θ, where C_p is defined as $B^{-1}(\eta \log p + \log A)$.

- Largely speaking, for any $\eta > 2$, the sign consistency holds for $n \sim CB^{-1}\eta \log p$, the rate of convergence for $\widehat{\Theta}$ is $O\left([\eta(\log p)/n]^{1/2}\right)$ under l_∞-norm.

Simulation

Simulation Settings

- 48 different scenarios:
  - 2 sample size setups (n = 500, 2000).
  - 3 dimension setups (p = 100, 300, 500).
  - 2 dropout levels (low: about 40 percent of the counts are zeros, high: about 60 percent of the counts are zeros).
  - 4 graph structures (Banded Graph, Random Graph, Scale-free Graph, Blocked Graph).
- Competitors:
  - PLNet-MOM (using moment estimator instead of MMLE in PLNet)
  - VPLN
  - glasso

Simulation

AUPR Results

Sample size	$n = 2000$		$n = 2000$		$n = 2000$	
Dimension	$p = 100$		$p = 300$		$p = 500$	
Dropout	Low	High	Low	High	Low	High
Banded graph						
PLNet	**0.99 (0.01)**	**0.96 (0.01)**	**0.99 (0.01)**	**0.94 (0.02)**	**0.98 (0.01)**	**0.89 (0.08)**
PLNet-MOM	0.97 (0.01)	0.92 (0.01)	0.91 (0.01)	0.83 (0.02)	0.83 (0.02)	0.75 (0.01)
VPLN	0.95 (0.01)	0.89 (0.03)	0.94 (0.01)	0.79 (0.15)	0.94 (0.01)	0.81 (0.01)
glasso	0.62 (0.03)	0.04 (0.01)	0.82 (0.01)	0.07 (0.01)	0.85 (0.01)	0.15 (0.02)
Random graph						
PLNet	**0.98 (0.01)**	**0.88 (0.04)**	**0.98 (0.03)**	**0.85 (0.05)**	**0.99 (0.01)**	**0.83 (0.04)**
PLNet-MOM	0.94 (0.02)	0.82 (0.06)	0.94 (0.01)	0.77 (0.05)	0.93 (0.01)	0.74 (0.05)
VPLN	0.78 (0.08)	0.69 (0.07)	0.88 (0.03)	0.67 (0.1)	0.86 (0.11)	0.67 (0.11)
glasso	0.55 (0.06)	0.18 (0.03)	0.8 (0.03)	0.24 (0.04)	0.84 (0.02)	0.26 (0.04)
Scale-free Graph						
PLNet	**0.89 (0.17)**	**0.85 (0.11)**	**0.97 (0.02)**	**0.85 (0.03)**	**0.96 (0.03)**	**0.83 (0.02)**
PLNet-MOM	0.85 (0.11)	0.81 (0.08)	0.86 (0.01)	0.75 (0.02)	0.83 (0.01)	0.71 (0.01)
VPLN	0.74 (0.16)	0.67 (0.15)	0.79 (0.04)	0.68 (0.11)	0.8 (0.05)	0.66 (0.13)
glasso	0.59 (0.14)	0.45 (0.06)	0.78 (0.02)	0.5 (0.03)	0.81 (0.02)	0.53 (0.02)
Blocked graph						
PLNet	**0.94 (0.02)**	**0.83 (0.07)**	**0.97 (0.01)**	**0.81 (0.08)**	**0.97 (0.01)**	**0.77 (0.05)**
PLNet-MOM	0.88 (0.04)	0.75 (0.08)	0.91 (0.02)	0.72 (0.08)	0.89 (0.02)	0.68 (0.05)
VPLN	0.73 (0.03)	0.66 (0.07)	0.78 (0.04)	0.62 (0.07)	0.8 (0.06)	0.59 (0.11)
glasso	0.47 (0.05)	0.2 (0.03)	0.7 (0.04)	0.21 (0.04)	0.75 (0.03)	0.21 (0.04)

Simulation

BIC, Banded graph

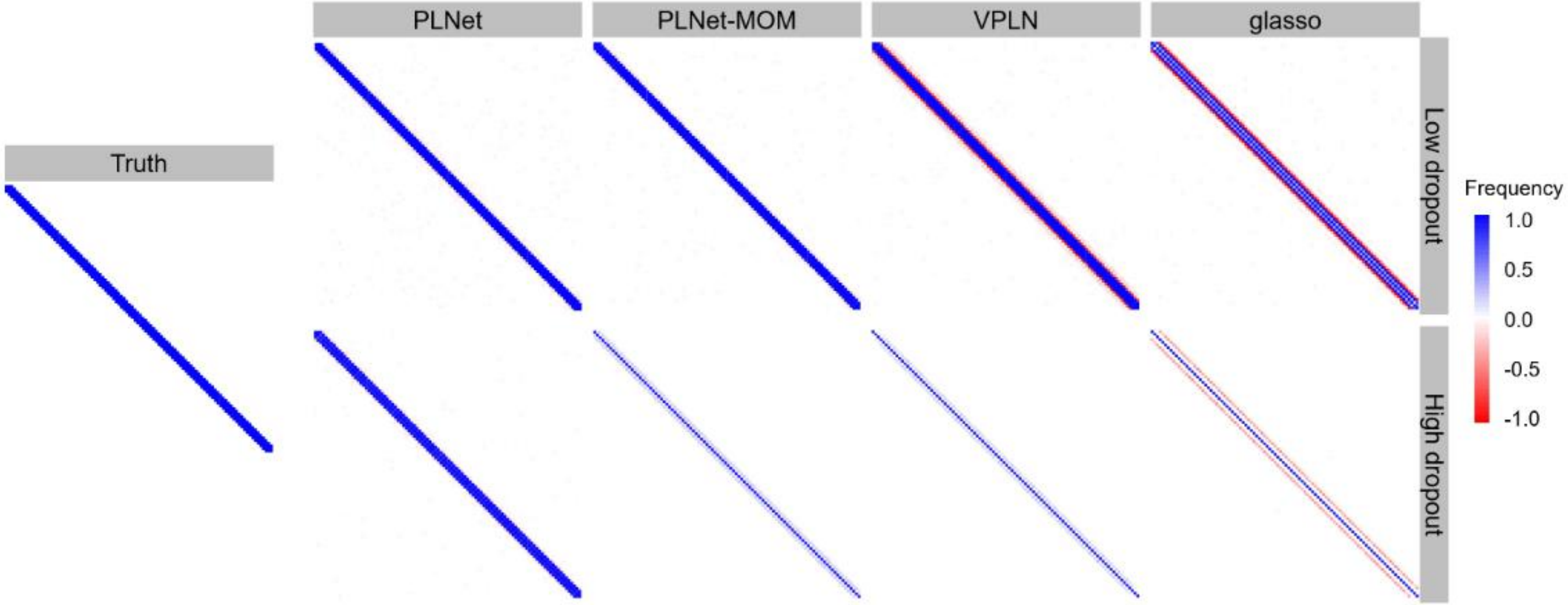


Fig. 1. The mean networks predicted by PLNet, VPLN, glasso and PLNet-MOM for the banded graph with 100 nodes and $n = 2000$. False edges are colored in red and true edges are in blue. The left panel is the true network matrix for reference.

Application to a scRNA–seq dataset

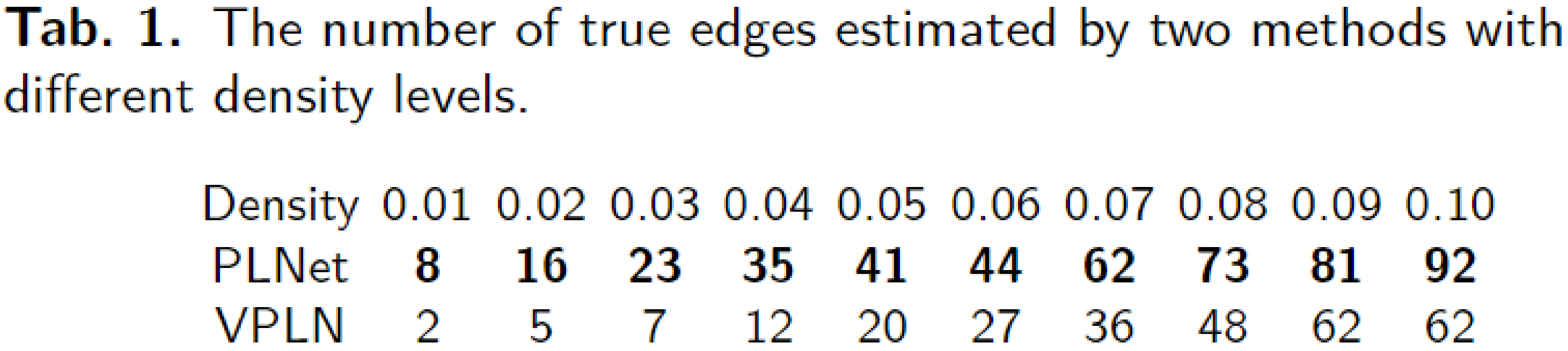
Peripheral Blood Mononuclear Cells (PBMC) Dataset

- A large scale scRNA-seq dataset with ctrl group and stim group stimulated by interferon β (IFN- β) from Kang et al.(2018).
- The CD14+ monocytes (2147 cells) in stim group to infer gene networks.
- Gene set: Top 200 highly variable genes + additional 26 TFs from the top 500 highly variable genes.
- The silver standard is based on an available regulatory network database obtained from ChIP-seq experiments (the hTFtarget database).

Application to a scRNA–seq dataset

Performance

- PLNet has a higher true discovery rate than VPLN.



Tab. 1. The number of true edges estimated by two methods with different density levels.

Density	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
PLNet	**8**	**16**	**23**	**35**	**41**	**44**	**62**	**73**	**81**	**92**
VPLN	2	5	7	12	20	27	36	48	62	62

Thanks for watching!