Poisson lognormal model

On-going works in the 'SingleStatOmics' perspective

MIA Paris

J. Chiquet, J. Kwon, M. Mariadasou, S. Robin
+ L. Sansonnet + futur Post-doc

Some virtual place, October 2020

Outline

1. Our PLN framework
2. "Canal historique" estimation strategy: variational
3. Ongoing work

Multivariate Poisson-log normal (PLN) distribution

A latent Gaussian model

Originally proposed by Atchisson [1]

$$\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

$$\mathbf{Y}_i \,|\, \mathbf{Z}_i \sim \mathcal{P}(\exp\{\mathbf{O}_i + \mathbf{X}_i^\intercal \mathbf{B} + \mathbf{Z}_i\})$$

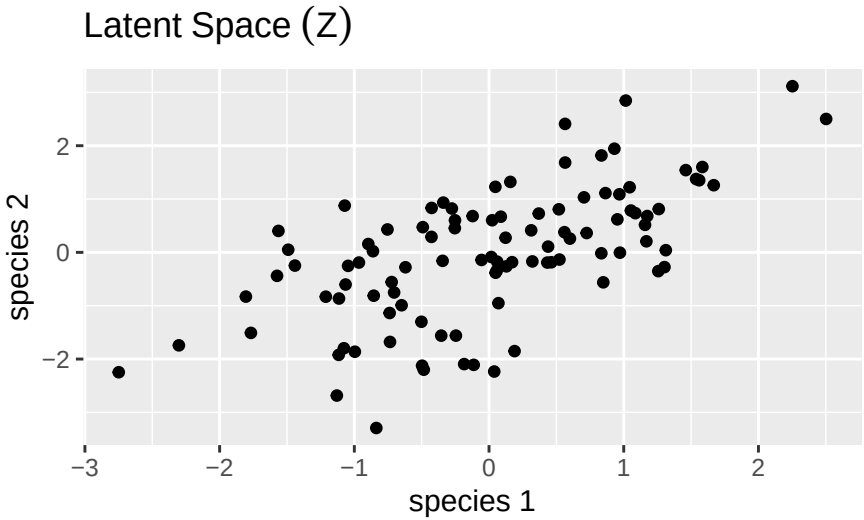
Interpretation

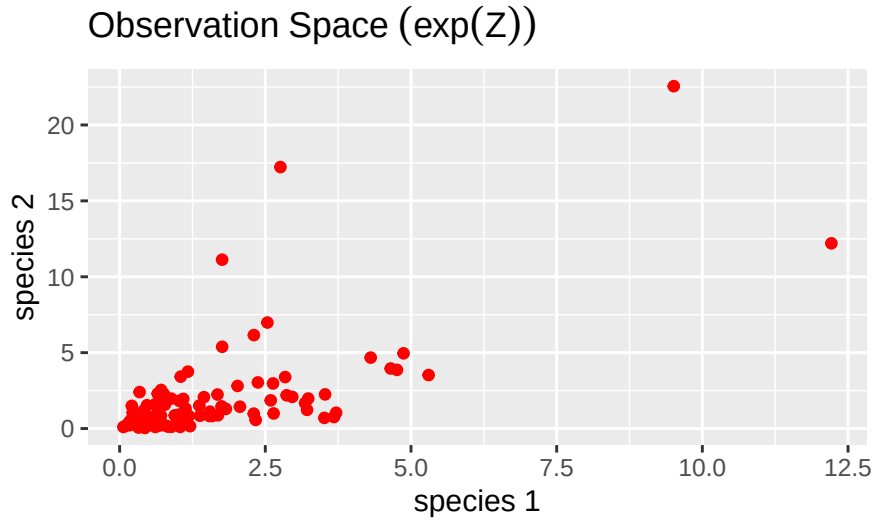
- Dependency structure encoded in the latent space (i.e. in $\mathbf{\Sigma}$)
- Additional effects are fixed
- Conditional Poisson distribution = noise model

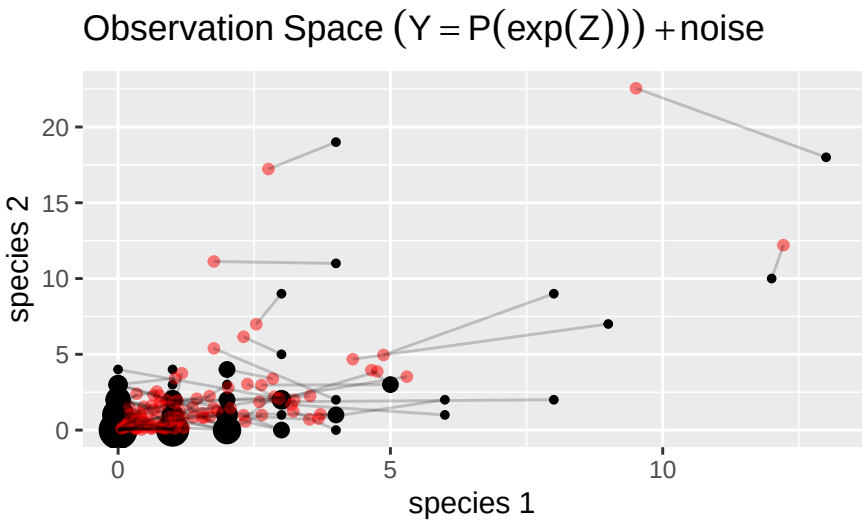
Properties

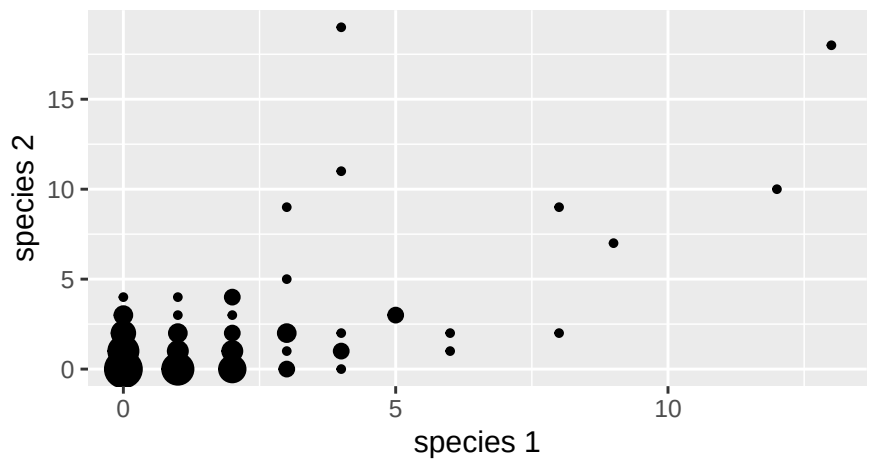
+ over-dispersion
+ covariance with arbitrary signs
- maximum likelihood via EM algorithm is limited to a couple of variables

Geometrical view

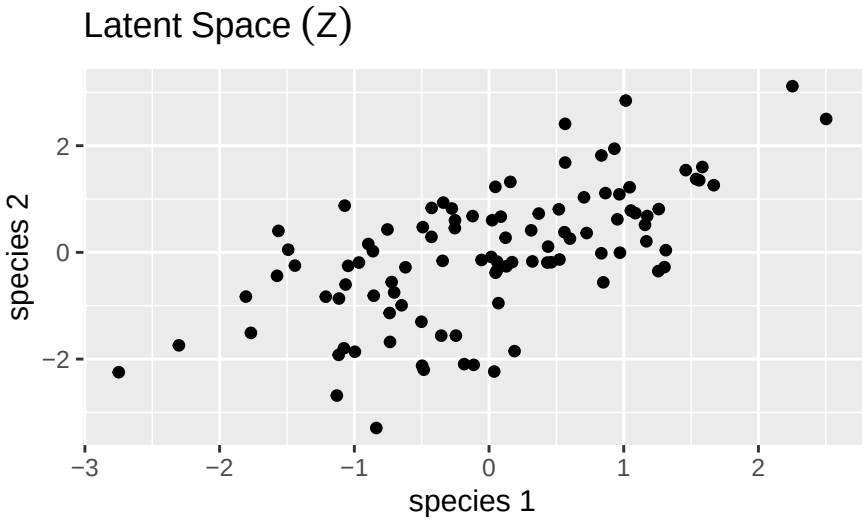


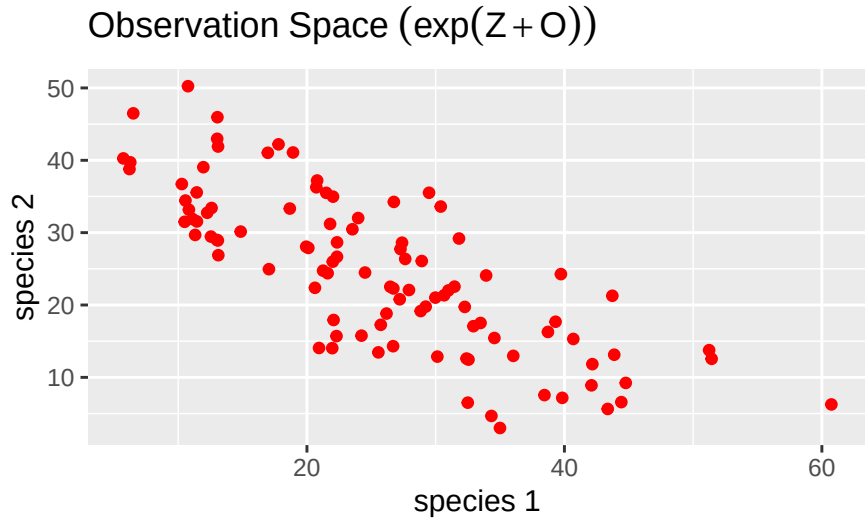


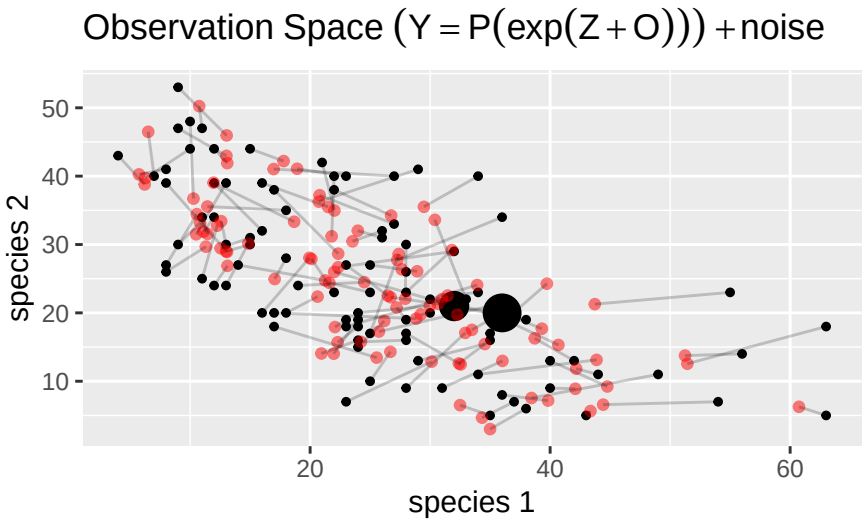


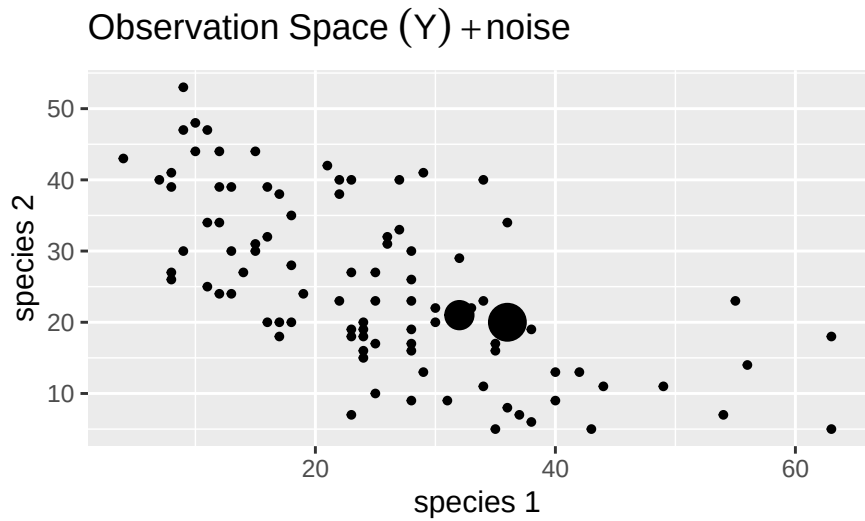


Geometrical view (with offset)









PLN: natural extensions towards multivariate analysis

- PCA: rank constraint on $\boldsymbol{\Sigma}$.

$$\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\top), \quad \mathbf{C} \in \mathcal{M}_{pk} \text{ with orthogonal columns.}$$

- Network: sparsity constraint on inverse covariance.

$$\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}), \quad \|\boldsymbol{\Omega}\|_1 < c.$$

- LDA: maximize separation between groups with means $\mathbf{M} = [\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_K^\top]^\top$

$$\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i = \mathbf{g}_i^\top \mathbf{M}, \boldsymbol{\Sigma}), \quad \mathbf{g}_i \text{ a group indicator vector.}$$

- Clustering: mixture model in the latent space

$$\mathbf{Z}_i \sim \prod_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \text{with, e.g., } \boldsymbol{\Sigma}_k \text{ diagonal matrices}$$

Outline

1. Our PLN framework
2. "Canal historique" estimation strategy: variational
3. Ongoing work

EM approach

Aim of the inference:

- estimate $\boldsymbol{\theta} = (\mathbf{B}, \boldsymbol{\Sigma})$
- predict the $\mathbf{Z}_i$

Maximum likelihood for incomplete data mode: EM

Le $\mathcal{H}(p) = -\mathbb{E}_p \log(p)$ be the entropy, then

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}) = \mathbb{E}_{\mathbf{Z} \mid \mathbf{Y}}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}[p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})].$$

Requires to evaluate (some moments of)

$$p(\mathbf{Z} \mid \mathbf{Y}) = \prod_i p(\mathbf{Z}_i \mid \mathbf{Y}_i)$$

but no close form for $p(\mathbf{Z}_i \mid \mathbf{Y}_i)$.

- [3] resorts to numerical or Monte-Carlo integration.
- Variational approach [5]: use a proxy of $p(\mathbf{Z} \mid \mathbf{Y})$.

Variational approach

Approximation of $p(\mathbf{Z} \mid \mathbf{Y})$: choose a class of distribution $\mathcal{Q}$

$$\mathcal{Q} = \left\{ \tilde{p} : \quad \tilde{p}(\mathbf{Z} \mid \mathbf{Y}) = \prod_i \mathcal{N}(\mathbf{Z_i}; \tilde{\mathbf{m}}_\mathbf{i}, \tilde{\mathbf{s}}_\mathbf{i}^\mathbf{2}) \right\}$$

and maximize the lower bound ($\tilde{\mathbb{E}}$ = expectation under $\tilde{p}$)

$$\begin{aligned} J(\boldsymbol{\theta}, \tilde{p}) &= \tilde{\mathbb{E}}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}[\tilde{p}(\mathbf{Z})] \\ &= \log p_{\boldsymbol{\theta}}(\mathbf{Y}) - KL[\tilde{p}(\mathbf{Z} \mid \mathbf{Y}) \,||\, p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})] \end{aligned}$$

Variational EM

1. **VE step:** find the optimal $\tilde{p}$, i.e., $(\mathbf{M}, \mathbf{S}) = \{(\mathbf{m}_i, \mathbf{s}_i)\}_{i=1}^n$:

$$\tilde{p}^h = \arg\max_{(\mathbf{M},\mathbf{S})} J(\boldsymbol{\theta}^h, \tilde{p}) = \arg\min_{\tilde{p} \in \mathcal{Q}} KL[\tilde{p}(\mathbf{Z}) \,||\, p_{\boldsymbol{\theta}^h}(Z \mid Y)]$$

2. **M step:** update $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}}^h = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \tilde{p}^h) = \arg\max_{\boldsymbol{\theta}} \tilde{\mathbb{E}}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})]$$

Optimization & Implementation

Property

The lower bound $J(\boldsymbol{\theta}, \tilde{p})$ is bi-concave, i.e.

- wrt $\tilde{p} = (\mathbf{M}, \mathbf{S})$ for given $\boldsymbol{\theta}$
- wrt $\boldsymbol{\theta} = (\boldsymbol{\Sigma}, \mathbf{B})$ for given $\tilde{p}$ (close form for $\widehat{\Sigma} = n^{-1}(\mathbf{M}^\top \mathbf{M} + \text{diag}(\mathbf{s}_+))$

but not jointly concave in general.

Optimization

Gradient ascent for the complete parameter $(\mathbf{M}, \mathbf{S}, \boldsymbol{\theta})$

- algorithm: conservative convex separable approximations [4]
- implementation: `NLopt` nonlinear-optimization package [2]
- initialization: LM after log-transformation applied independently on each variables + concatenation of the regression coefficients + Pearson residuals

$\rightsquigarrow$ Existing variants for PLN-network, PLN-PCA, PLN-LDA and PLN-mixture.

Contributions

J.C., Mahendra Mariadassou, Stéphane Robin,

The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances

https://doi.org/10.1101/2020.10.07.329383 bioRxiv, submitted to Frontiers Ecology And Evolution

J.C., Mahendra Mariadassou, Stéphane Robin,

Variational inference for probabilistic Poisson PCA

http://dx.doi.org/10.1214/18-AOAS1177 Ann Appl Statist 12: 2674–2698, 2018

J.C., Mahendra Mariadassou, Stéphane Robin,

Variational inference for sparse network reconstruction from count data

In Proceedings of the 19th International Conference on Machine Learning (ICML'19)

PLNmodels package, development version on github

```
install.packages("PLNmodels")
```

https://jchiquet.github.io/PLNmodels/

Outline

1. Our PLN framework
2. "Canal historique" estimation strategy: variational
3. Ongoing work
   - Other/better models
   - Large-scale problems
   - Convergence: optimization, statistics

Outline

1. Our PLN framework
2. "Canal historique" estimation strategy: variational
3. Ongoing work
   - Other/better models
   - Large-scale problems
   - Convergence: optimization, statistics

Zero-inflated versions

with Mahendra, François Gindraud?

$(Y_i) \in \mathbb{N}^p$ vector counts for observation i distributed as:

$$\mathbf{Z}_i \sim \mathcal{N}_p(\mathbf{B}\mathbf{x}_i, \mathbf{\Sigma})$$

$$N_{ij} \sim \mathcal{B}(\text{logit}(\mathbf{x}_i^\top \mathbf{B}_j^0))$$

$$Y_{ij}|N_{ij}, Z_{ij} \sim W_{ij}\delta_0 + (1 - W_{ij})\mathcal{P}\left(\exp\left\{o_{ij} + Z_{ij}\right\}\right)$$

- N_{ij}, Z_{ij} independent
- $Y_{ij}|N_{ij}, Z_{ij}$ independent
- $\mathbf{B}$: $d \times p$ matrix of regression coefficients for the observed counts,
- $\mathbf{B}^0$: $d \times p$ matrix of regression coefficients for the zero-inflation.

$\rightsquigarrow$ Extension of PLNnetwork to ZI-PLNnetwork would also be greatly appreciated!

Variational Approximation

Assume the following factorization

$$p(\mathbf{Z}_i, \mathbf{N}_i | \mathbf{Y}_i) = \tilde{p}_\psi(\mathbf{Z}_i, \mathbf{N}_i) = \tilde{p}_\psi(\mathbf{Z}_i, \mathbf{N}_i) = \tilde{p}_{\psi_1}(\mathbf{Z}_i)\tilde{p}_{\psi_2}(\mathbf{N}_i)$$

with

- $\tilde{p}_{\psi_1}(\mathbf{Z}_i)$ is a Gaussian distribution with diagonal covariance matrix

$$\tilde{p}_{\psi_1}(Z_i) = \mathcal{N}_p(Z_i; \mathbf{m}_i, \mathbf{s}_i).$$

- $\tilde{p}_{\psi_2}(\mathbf{W}_i)$ is a product-distribution of p Bernoulli distributions:

$$\tilde{p}_{\psi_2}(N_i) = \otimes_{j=1}^{p} \mathcal{B}(N_{ij}, \pi_{ij})$$

Three levels of approximations

1. Conditional independence of $\mathbf{N}_i$ and $\mathbf{Z}_i$ (given $\mathbf{Y}_i$)
2. Shape of the posterior distribution of $\mathbf{Z}_i$ (which is likely not Gaussian in practice)
3. Independence of the components of $\mathbf{N}_i$ (each N_{ij} is indeed a Bernoulli variable but they are likely not independent in practice)

Outline

1. Our PLN framework
2. "Canal historique" estimation strategy: variational
3. Ongoing work
   - Other/better models
   - Large-scale problems
   - Convergence: optimization, statistics

Large-scale optimizers

Variational lower bound

Keep on optimizing the variational criterion, by relying on

- tools for automatic differentiation
- stochastic-gradient descent variants
- GPU

$\rightsquigarrow$ first prove en concept with `pytorch` for PLN + PLNPCA

Can we do better than that?

Pave the way for optimizing log-likelihood directly

- Statistical guarantees eaasiest to derive
- Work started for PLN-PCA

PLN-PCA model and target log-likelihood

- $n \geqslant 1$ sample size
- $p \geqslant 1$ dimension of observation space
- $q \geqslant 1$ dimension of latent space
- $d \geqslant 1$ dimension of covariates

$$
\begin{aligned}
W_i &\sim \mathcal{N}(0, I_q), \text{ iid}, \quad i = 1, \ldots, n \\
Z_i &= \mathbf{B}\mathbf{x}_i + \mathbf{C}\, W_i, \quad i \in 1, \ldots, n, \\
Y_{ij} | Z_{ij} &\sim \mathcal{P}(\exp\left(o_{ij} + Z_{ij}\right)).
\end{aligned}
$$

Log-likelihood of an observation Y_i $(i \in [n])$ writes:

$$
\log p_\theta(\, Y_i) = \mathbb{E}_{W \sim \mathcal{N}(0, I_q)} \left[\exp \left(\sum_{j=1}^{p} \quad \ldots \quad \right) \right]
$$

Function to minimize:

$$
F(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(\, Y_i).
$$

(Stochastic) Gradient descent

Gradient descent:

$$\theta_{t+1} = \theta_t - \gamma_t \nabla F(\theta_t)$$

We have no access to $\nabla F(\theta_t)$. But it can be replaced by an unbiased estimator $\hat{g}_t$. This gives stochastic gradient descent:

$$\theta_{t+1} = \theta_t - \gamma_t \hat{g}_t, \qquad \text{where} \quad \mathbb{E}\left[\hat{g}_t \mid \theta_t\right] = \nabla F(\theta_t).$$

Since

$$\begin{aligned}
\nabla F(\theta) &= -\frac{1}{n}\sum_{i=1}^{n} \frac{\nabla_\theta p_\theta(Y_i)}{p_\theta(Y_i)} \\
&= -\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{E}_{W\sim\mathcal{N}(0,I_q)}\left[\nabla_\theta \exp\left(\sum_{j=1}^{p}(\quad\cdots\quad)\right)\right]}{\mathbb{E}_{W\sim\mathcal{N}(0,I_q)}\left[\exp\left(\sum_{j=1}^{p}(\quad\cdots\quad)\right)\right]}
\end{aligned}$$

Estimator $\hat{g}_t$ can be constructed by

- Sampling values of W
- Possibly sampling a subset $I \subset \{n\}$ of a given cardinality (for lowering computational cost and scaling for large n).

Adaptive algorithms: sophistications of SGD

- AdaGrad (2011) uses adaptive coordinate-wise step-sizes:

$$\theta_{t+1} = \theta_t - \frac{\gamma}{\sqrt{\varepsilon + G_t}} \odot \hat{g}_t \qquad \text{where} \quad G_t = \sum_{s=1}^{t} \hat{g}_s^{\odot 2}.$$

- RMSProp (2012) adds momentum to the step-sizes:

$$\theta_{t+1} = \theta_t - \frac{\gamma}{\sqrt{\varepsilon + G_t}} \odot \hat{g}_t \qquad \text{where} \quad G_t = \alpha G_{t-1} + (1-\alpha)\hat{g}_t^{\odot 2}.$$

- Adam (2015) also adds momentum to the gradients:

$$\theta_{t+1} = \theta_t - \frac{\gamma}{\sqrt{\varepsilon + G_t}} \odot \hat{m}_t \qquad \text{where} \quad m_t = \beta m_{t-1} + (1-\beta)\hat{g}_t.$$

Outline

1. Our PLN framework
2. "Canal historique" estimation strategy: variational
3. Ongoing work
   - Other/better models
   - Large-scale problems
   - Convergence: optimization, statistics

Post-doc: theoretical guarantees

With practical implications!

Properties of VEM inference / V-estimator

VEM stationary point $\neq$ log-likelihood stationary point

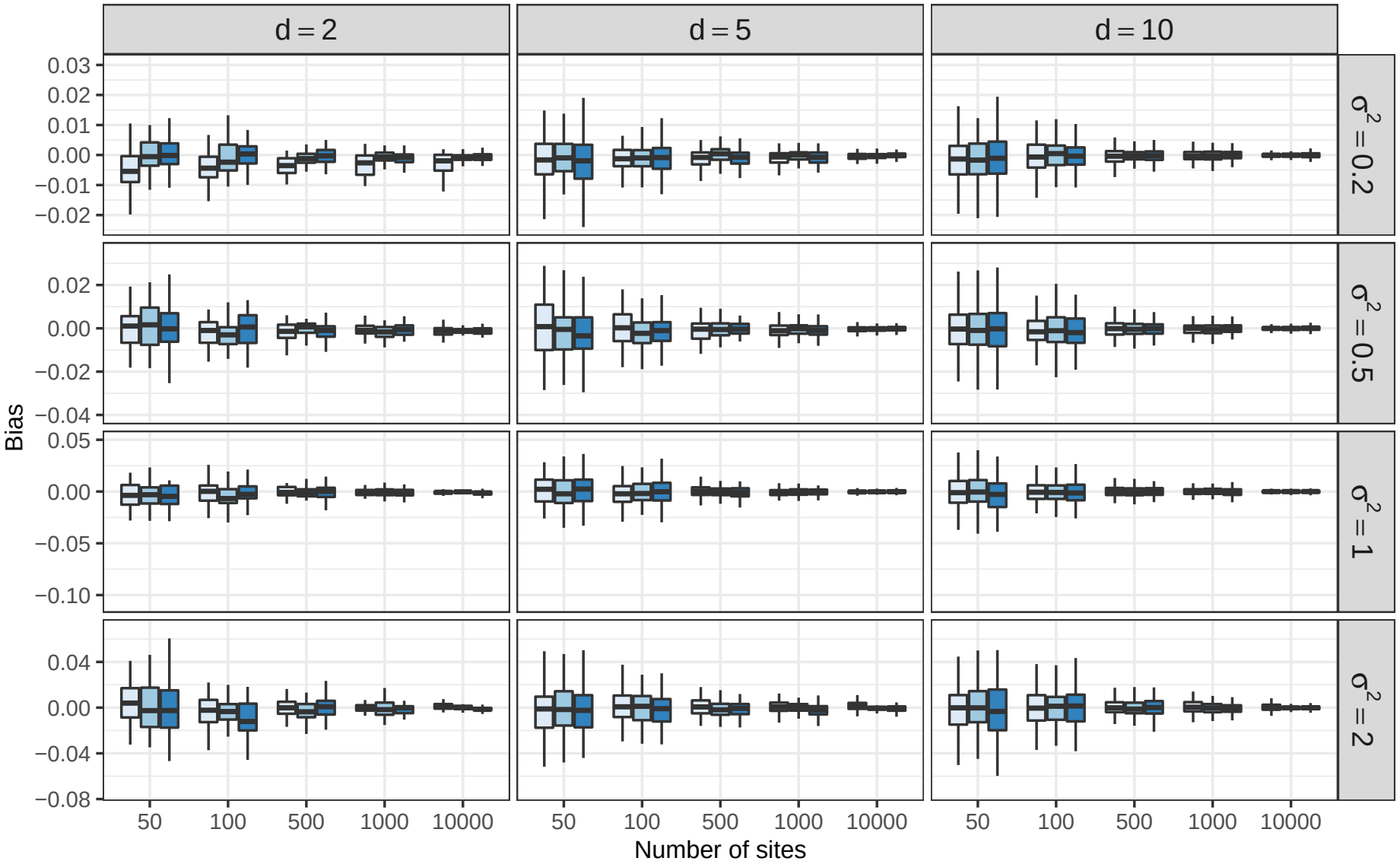
- Consistency results: M-estimator
- Asymptotic Variance of V-estimator: Louis formula

$\rightsquigarrow$ Numerical study: asymptotically unbiased, but variance underestimated
$\rightsquigarrow$ No trusted confidence intervals can be derived

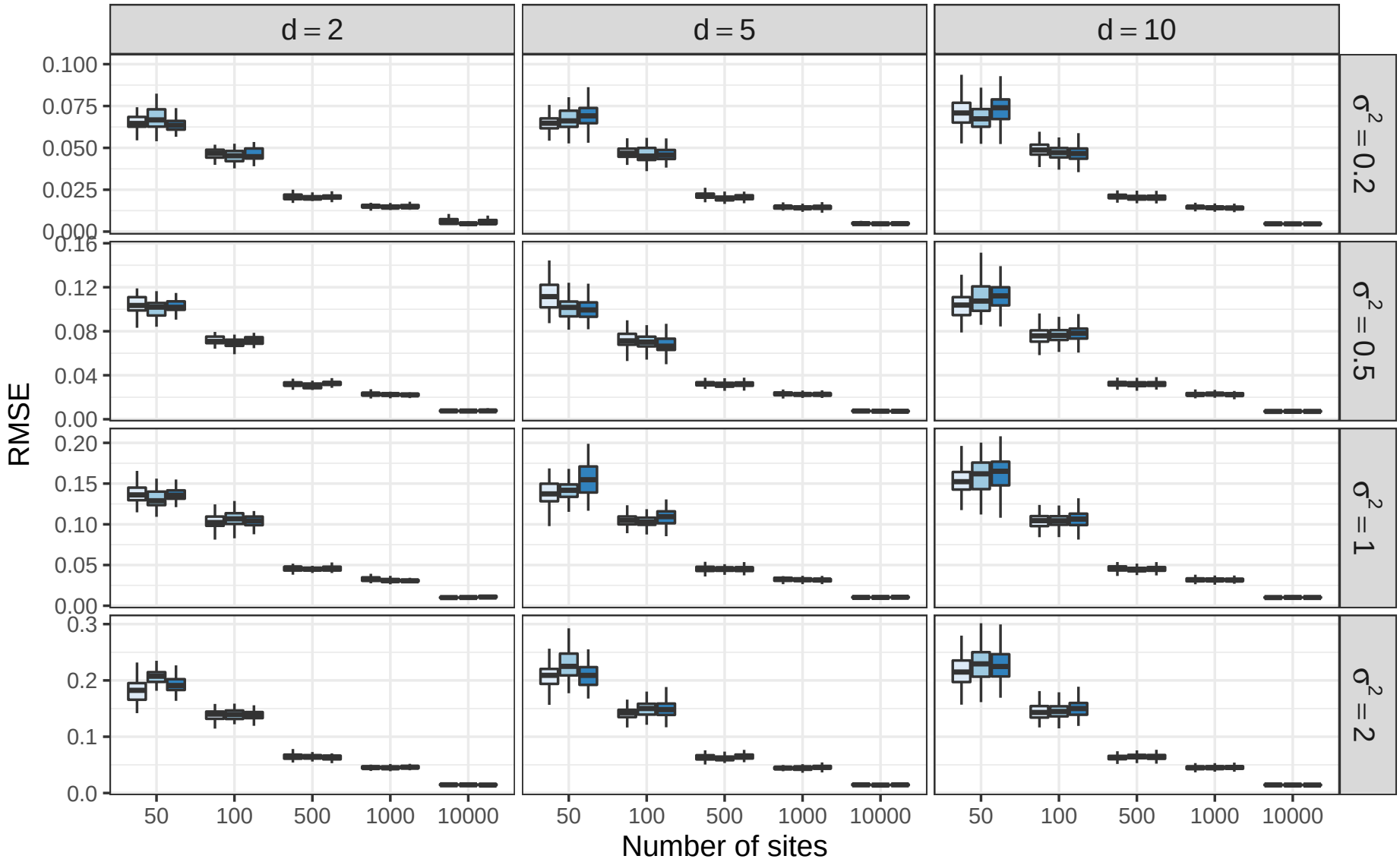
Simulations

- number of samples $n \in \{50, 100, 500, 1000, 10000\}$
- number of species/genes $p \in \{20, 200\}$
- number of covariates $d \in \{2, 5, 10\}$
- sampling effort $TSS \in \{\text{low}, \text{medium}, \text{high}\}$ ($\approx 10^4$, 10^5 and 10^6)
- noise level $\sigma^2 \in \{0.2, 0.5, 1, 2\}$
- Σ as $\sigma_{jk} = \sigma^2 \rho^{|j-k|}$, with $\rho = 0.2$
- $\mathbf{B}$ with entries sampled from $\mathcal{N}(0, 1/d)$

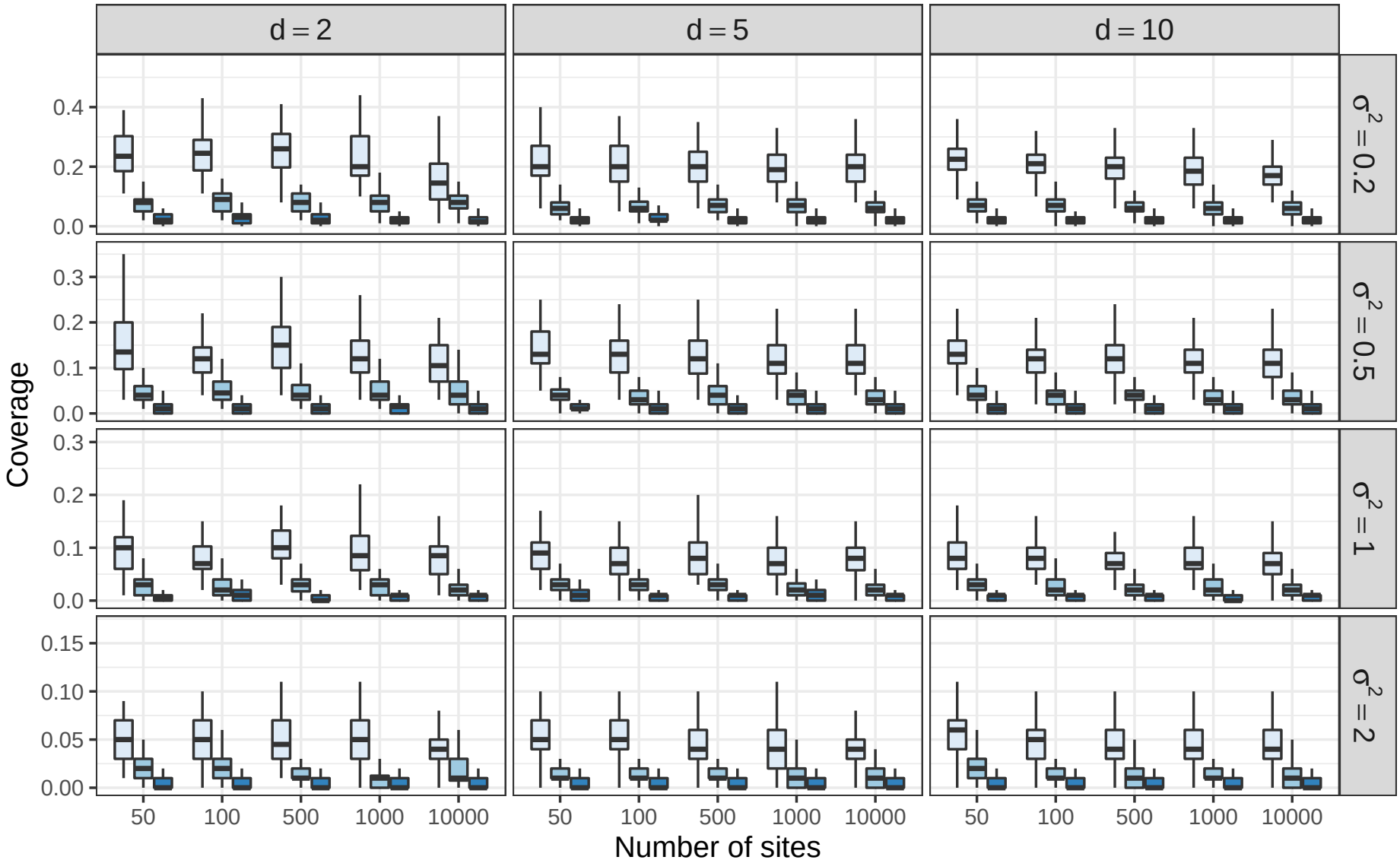
Bias


d = 2
d = 5
d = 10
σ² = 0.2
σ² = 0.5
σ² = 1
σ² = 2
Bias
Number of sites
50
100
500
1000
10000

RMSE


d = 2
d = 5
d = 10
σ² = 0.2
σ² = 0.5
σ² = 1
σ² = 2
RMSE
0.100
0.075
0.050
0.025
0.000
0.16
0.12
0.08
0.04
0.00
0.20
0.15
0.10
0.05
0.00
0.3
0.2
0.1
0.0
50
100
500
1000
10000
Number of sites

Coverage


d = 2
d = 5
d = 10
σ² = 0.2
σ² = 0.5
σ² = 1
σ² = 2
Coverage
0.4
0.2
0.0
0.3
0.2
0.1
0.0
0.3
0.2
0.1
0.0
0.15
0.10
0.05
0.00
50
100
500
1000
10000
Number of sites

Confidence interval: a new hope

Use regular inference

Using VEM output as a starting point for regular inference:

- Get maximum-(composite-)likelihood estimates starting from $\hat{\boldsymbol{\theta}}_{VEM}$
- Get log-likelihood estimates with SGD starting from $\hat{\boldsymbol{\theta}}_{VEM}$

$\rightsquigarrow$ Hopefully: few iterations are needed

SGD convergence

$\rightsquigarrow$ Theoretical guarantees ?

Use alternative algorithm for which proof is easiest

$\rightsquigarrow$ Connexion between convergence from optim/stat viewpoint

References


J. Aitchison and C.H. Ho.
The multivariate poisson-log normal distribution.
Biometrika, 76(4):643–653, 1989.


Steven G Johnson.
The NLopt nonlinear-optimization package, 2011.


D. Karlis.
EM algorithm for mixed Poisson and other discrete distributions.
Astin bulletin, 35(01):3–24, 2005.


Krister Svanberg.
A class of globally convergent optimization methods based on conservative convex separable approximations.
SIAM journal on optimization, 12(2):555–573, 2002.


M. J. Wainwright and M. I. Jordan.
Graphical models, exponential families, and variational inference.
Found. Trends Mach. Learn., 1(1–2):1–305, 2008.