Structured Poisson-lognormal models for the genetics of the 'social microbiota' in crop mixtures

FMJH thematic post-doctoral position 2027, Mathematics for Life Sciences

MIA Paris-Saclay, in collaboration with GQE-Le Moulon (Université Paris-Saclay)

Keywords. Latent variable models for multivariate counts; high-dimensional multivariate regression; structured regularisation; variational inference; mixed models; statistical genetics; microbial ecology.

Summary. The project develops statistical methodology for regressing thousands of sparse, over-dispersed and correlated counts on hundreds of thousands of predictors, with structure on both sides of the regression. The starting point is the zero-inflated Poisson-lognormal (PLN) model developed in the host team, and the open questions lie in high-dimensional statistics, variational inference and optimisation. The motivation comes from plant genetics: finding the genes through which a crop shapes its microbial communities, and those of its neighbour when two species are grown together. A large field experiment provides the data.

1 Motivation and data

Plant genes partly control the microbial communities living on leaves and roots: host genotype typically explains 5 to 30% of the variation in their composition, that is, in the presence and abundance of bacterial and fungal taxa (Bergelson et al. 2021). In species mixtures, where two or more crops are grown together in the same field (e.g. wheat and pea), a plant's genes may also act on its neighbour's microbes. Such effects of an individual on the phenotype of its neighbours are social interactions in a broad sense (Bilas et al. 2021). Their quantitative genetics, whose foundations go back to Griffing (1967), has recently gained momentum in crop mixtures and intercrops (Sampoux et al. 2020; Haug et al. 2021; Bančič et al. 2021; Salomon et al. 2026). Although social interactions may play an important role in microbiota recruitment, this has not been investigated so far, largely for lack of field experiments of sufficient scale and of suitable statistical tools.

Our partners at GQE-Le Moulon ran a field experiment designed for this question. A panel of 200 genotyped wheat varieties was grown either as sole crops or in mixture with each of two pea varieties, in two randomised replicates plus control plots (about 1,300 plots in total). Microbial samples were collected from leaves and soil (3,900 and 412 respectively, analysis by metabarcoding pending), and yield-related traits were measured on both species. In statistical terms, this gives count tables with thousands of taxa and 80 to 90% of zeros, strong dependence between taxa, and 2×10^6 genetic markers (loci), about 10^5 after filtering, for 200 genotypes.

2 Statistical starting point

Let $\mathbf{Y}_i \in \mathbb{N}^p$ be the vector of counts of p taxa in sample $i = 1, \dots, n$, $\mathbf{x}_i \in \mathbb{R}^d$ a vector of covariates and $\mathbf{o}_i$ a known offset accounting for sampling effort. The zero-inflated PLN model (Aitchison and Ho 1989; Chiquet et al. 2021; Batardière et al. 2025) assumes

$$\mathbf{Z}_i \sim \mathcal{N}(\mathbf{o}_i + \mathbf{B}^\top \mathbf{x}_i,\ \Sigma), \qquad W_{ij} \sim \mathcal{B}(\pi_{ij}), \qquad Y_{ij} \mid Z_{ij}, W_{ij} \sim (1 - W_{ij})\, \mathcal{P}\big(e^{Z_{ij}}\big),$$

Here $\mathbf{Z}_i$ is a latent Gaussian vector of log-abundances, and W_{ij} is a latent Bernoulli variable accounting for structural zeros (absence of taxon j from sample i), with $\mathrm{logit}(\pi_{ij})$ possibly depending on covariates. The regression matrix $\mathbf{B} \in \mathbb{R}^{d \times p}$ carries the effects of covariates on the latent log-abundances, and Σ the residual dependence between taxa. Maximum likelihood is the natural route to inference in this model, but the likelihood is intractable. Estimation therefore relies on a variational EM algorithm, which maximises an evidence lower bound $J(\mathbf{B}, \Sigma, \psi)$ over model and variational parameters (Chiquet et al. 2018). It is implemented in the `PLNmodels` (R/C++) and `pyPLNmodels` (Python/PyTorch, GPU) packages maintained by the team (Batardière et al. 2024). The framework currently handles a moderate number d of covariates and independent samples.

3 Research programme

The programme has three stages of increasing ambition. Depending on the skills of the recruited candidate, the emphasis may be put on theory, methodology or computation.

3.1 Genetic effects in the latent layer: high-dimensional structured regression

With $\mathbf{x}_i$ containing the marker profile of the variety grown in plot i, estimating $\mathbf{B}$ is a multivariate regression with $d \gg n$ and p in the thousands, to be solved through a non-Gaussian emission layer. The candidate will develop penalised variational estimators of the form

$$(\widehat{\mathbf{B}}, \widehat{\Sigma}) \in \arg\max_{\mathbf{B},\Sigma,\psi} \; J(\mathbf{B}, \Sigma, \psi) - \lambda \, \mathrm{pen}(\mathbf{B}),$$

for three families of penalties, each corresponding to a specific genetic framework.

- *Polygenic effects and heritability.* A ridge penalty amounts to a random genetic effect: the latent layer becomes a multivariate mixed model $\mathbf{Z}_i = \mathbf{o}_i + \mathbf{u}_{g(i)} + \varepsilon_i$, where $g(i)$ is the variety grown in sample i, $\mathbf{u}_g \in \mathbb{R}^p$ is the genetic effect of variety g on the p taxa and $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \Sigma_E)$. The matrix $\mathbf{U}$ stacking the $\mathbf{u}_g$ over the 200 varieties is matrix-normal, $\mathbf{U} \sim \mathcal{MN}(\mathbf{0}, \mathbf{K}, \Sigma_G)$, where $\mathbf{K}$ is the genomic relationship matrix between varieties and Σ_G, Σ_E are the $p \times p$ genetic and residual covariances between taxa. This yields latent-scale heritabilities and genetic correlations for every taxon, but raises two difficulties: Σ_G and Σ_E cannot both be estimated without structure (low rank, sparse precision, blocks) when only 200 genotypes are available, and $\mathbf{K}$ makes samples dependent, which breaks the sample-wise factorisation on which current variational algorithms rely. The spectral tricks that make Gaussian multivariate mixed models tractable (Zhou and Stephens 2014) do not carry over through the Poisson layer; new variational families and algorithms are needed.
- *Locus detection.* Sparse and group-sparse penalties on the rows of $\mathbf{B}$ select loci acting on one or several taxa, accounting for the dependence between responses as in Chiquet et al. (2017) for the Gaussian case.
- *Clusters of co-controlled taxa.* We conjecture that many taxa share similar genetic effect profiles. Clustering the columns of $\mathbf{B}$ during estimation (fusion-type penalties, mixture or low-rank structure on effect profiles) should both increase detection power by pooling weak signals and reveal groups of taxa under common host control.

3.2 Neighbour-dependent effects and two interacting species

Samples collected from a plot grown with a wheat-pea mixture capture the leaf microbiota of both species and, for a subset of plots, that of the soil (412 samples, against 3,900 for leaves). The latent vector is decomposed accordingly into blocks, $\mathbf{Z}_i = (\mathbf{Z}_i^{\text{wheat}}, \mathbf{Z}_i^{\text{pea}}, \mathbf{Z}_i^{\text{soil}})$, the soil block being missing for most plots. Likewise, the matrix of regression coefficients $\mathbf{B}$ can be split into direct effects (of a genotype on its own microbiota, in sole crops and in mixtures alike), social or indirect effects (of a genotype on the microbiota of its companion species, or of its neighbours of the same variety in a sole crop), and genotype-by-genotype interactions between associated partners. This follows the decomposition proposed by Salomon et al. (2026) for agronomic traits in sole crops and intercrops, carried over here to microbial counts. The post-doctoral researcher will formulate this block-structured model for general two-species factorial designs, determine which effects are identifiable under a given experimental design (here the two pea varieties act as testers), and handle the missing blocks.

3.3 From microbiota to crop performance

The last stage adds the agronomic traits of both species, $\mathbf{t}_i \in \mathbb{R}^r$, as a further layer,

$$\mathbf{t}_i = \mathbf{A}^\top \mathbf{x}_i + \Gamma^\top \mathbf{Z}_i + \mathbf{e}_i,$$

where $\mathbf{A} \in \mathbb{R}^{d \times r}$ holds the effects of markers and covariates on the traits that do not involve the microbiota, $\Gamma \in \mathbb{R}^{p \times r}$ the effects of the latent abundances, and $\mathbf{e}_i$ is a Gaussian error term. Latent abundances thus act as high-dimensional mediators between genotype and performance: the total effect of markers on traits is $\mathbf{A} + \mathbf{B}\Gamma$, of which $\mathbf{B}\Gamma$ transits through the microbiota. For a trait of the pea neighbour, the social effects of wheat that bypass the microbiota, such as competition for light through plant height, sit in the corresponding column of $\mathbf{A}$, and the microbiota-mediated ones in that of $\mathbf{B}\Gamma$. The objective is to quantify the share of genetic variance in mixture performance carried by the mediated term, and the gain in predictive accuracy it provides. A two-step approach plugging in variational means of $\mathbf{Z}_i$ is a natural baseline, but it ignores the uncertainty on a latent regressor; a joint hierarchical model is the goal. The post-doctoral researcher will implement this joint model of traits and microbiota, and investigate to what extent it can identify causal effects of the microbiota on traits.

3.4 Computation and software

Problems with $d \times p$ of order 10^8 to 10^9 coefficients call for dedicated optimisation. Computational efficiency could be improved through stochastic proximal methods for the non-convex variational objective, screening rules, low-rank parametrisations and GPU implementation. Methods will be released in the team's open-source packages and assessed first on simulations built from the marker data of the experiment, then on its microbiota data as they become available. Protocols that co-amplify a host gene to measure microbial load (Lundberg et al. 2021) may be used, in which case the offset $\mathbf{o}_i$ becomes a modelling object in its own right.

4 Candidate profile

PhD in statistics, applied mathematics or machine learning, with a solid background in at least one of: latent variable models and variational inference, high-dimensional statistics, mixed models, optimisation. Good programming skills (Python/PyTorch or R). Interest in applications to the life sciences; no prior knowledge of genetics or microbial ecology is required.

5 Environment and practical information

The post-doctoral researcher will join the MIA Paris-Saclay research unit (Palaiseau), whose members develop the PLN framework and its software and have long-standing activity in statistical genetics and network inference. The work will be carried out in close interaction with the quantitative geneticists of GQE-Le Moulon (Gif-sur-Yvette, a few kilometres away), who designed and ran the field experiment and coordinate the biological and molecular analyses with a group of soil and microbial ecologists.

The position is a two-year FMJH fellowship starting 1 October 2027 (the PhD must be defended before that date). Applications (a CV and a cover letter) are submitted through the FMJH call, open from 16 October to 6 December 2026: https://fondation-hadamard.fr/fr/programmes/les-programmes-transverses/allocation-postdoc/post-docs-thematiques/. Candidates are encouraged to contact us beforehand.

Contact. Julien Chiquet (MIA Paris-Saclay), julien.chiquet@inrae.fr, Jérôme Enjalbert, Tristan Mary-Huard and Timothée Flutre (GQE-Le Moulon).

References

Aitchison, John, and C. H. Ho. 1989. "The Multivariate Poisson-Log Normal Distribution." *Biometrika* 76 (4): 643–53.

Bančič, Jon et al. 2021. "Modeling Illustrates That Genomic Selection Provides New Opportunities for Intercrop Breeding." *Frontiers in Plant Science* 12: 605172. https://doi.org/10.3389/fpls.2021.605172.

Batardière, Bastien, Julien Chiquet, François Gindraud, and Mahendra Mariadassou. 2025. "Zero-Inflation in the Multivariate Poisson Lognormal Family." *Statistics and Computing* 35. https://doi.org/10.1007/s11222-025-10729-0.

Batardière, Bastien, Joon Kwon, and Julien Chiquet. 2024. "pyPLNmodels: A Python Package to Analyze Multivariate High-Dimensional Count Data." *Journal of Open Source Software* 9 (104): 6969. https://doi.org/10.21105/joss.06969.

Bergelson, Joy, Benjamin Brachi, Fabrice Roux, and Fabienne Vailleau. 2021. "Assessing the Potential to Harness the Microbiome Through Plant Genetics." *Current Opinion in Biotechnology* 70: 167–73. https://doi.org/10.1016/j.copbio.2021.05.007.

Bilas, R. D., A. Bretman, and T. Bennett. 2021. "Friends, Neighbours and Enemies: An Overview of the Communal and Social Biology of Plants." *Plant, Cell & Environment* 44 (4): 997–1013.

Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin. 2018. "Variational Inference for Probabilistic Poisson PCA." *The Annals of Applied Statistics* 12 (4): 2674–98. https://doi.org/10.1214/18-AOAS1177.

Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin. 2021. "The Poisson-Lognormal Model as a

Versatile Framework for the Joint Analysis of Species Abundances." *Frontiers in Ecology and Evolution* 9: 588292. https://doi.org/10.3389/fevo.2021.588292.

Chiquet, Julien, Tristan Mary-Huard, and Stéphane Robin. 2017. "Structured Regularization for Conditional Gaussian Graphical Models." *Statistics and Computing* 27 (3): 789–804. https://doi.org/10.1007/s11222-016-9654-1.

Griffing, B. 1967. "Selection in Reference to Biological Groups. I. Individual and Group Selection Applied to Populations of Unordered Groups." *Australian Journal of Biological Sciences* 20: 127–39.

Haug, Benedikt, Monika M. Messmer, Jérôme Enjalbert, et al. 2021. "Advances in Breeding for Mixed Cropping – Incomplete Factorials and the Producer/Associate Concept." *Frontiers in Plant Science* 11: 620400. https://doi.org/10.3389/fpls.2020.620400.

Lundberg, Derek S. et al. 2021. "Host-Associated Microbe PCR (hamPCR) Enables Convenient Measurement of Both Microbial Load and Community Composition." *eLife* 10: e66186. https://doi.org/10.7554/eLife.66186.

Salomon, Jemay, Jérôme Enjalbert, and Timothée Flutre. 2026. "Joint Modeling of Social Genetic Effects in Mono- and Pluri-Specific Groups: Case Study in Intercrops." *Genetics* 234 (1): iyag177. https://doi.org/10.1093/genetics/iyag177.

Sampoux, Jean-Paul, Héloïse Giraud, and Isabelle Litrico. 2020. "Which Recurrent Selection Scheme to Improve Mixtures of Crop Species? Theoretical Expectations." *G3: Genes, Genomes, Genetics* 10 (1): 89–107. https://doi.org/10.1534/g3.119.400809.

Zhou, Xiang, and Matthew Stephens. 2014. "Efficient Multivariate Linear Mixed Model Algorithms for Genome-Wide Association Studies." *Nature Methods* 11: 407–9. https://doi.org/10.1038/nmeth.2848.