Large-scale composite hypothesis testing for detecting multi-virus resistance QTLs in cucumber

Annaïg De Walsche, Franck Gauthier, Alain Charcosset, Tristan Mary-Huard

DIGIT-BIO Meta-program
Axe 1 : Deciphering the functions of living matter

PhD, February 2022 – March 2025
Funding by Digit-Bio (50%) and KWS (50%)

Supervised by :
- Tristan Mary-Huard (MIA Paris-Saclay)
- Alain Charcosset (GQE Le Moulon)

Context

The **joint analysis of results from different experiments** to identify complex patterns or to improve statistical power is a typical objective of data integration. Here we consider the case of a set of markers $m = 1, \dots, n$ whose effect has been tested under different conditions $k = 1, \dots, K$. Each marker m, therefore, consists of a vector of K critical probabilities. The analysis aims **to identify the elements that have an effect in all the conditions** or in a predefined **subset of conditions**. The critical probabilities must then be flexibly combined to explore complex hypotheses (known as composite hypotheses) while controlling the false positive rate.

Mixture model approach

Framework

We consider a set of markers $m = 1, \dots, n$ whose effects has been tested under different conditions $k = 1, \dots, K$.

We note $\mathcal{C} = \{c = (c_1, \dots, c_K), c_k \in \{0,1\}\}$ the set of $\mathbf{2^K}$ **configurations** of the marker effects across the conditions.

	Trait 1	Trait 2	Trait 3
Marker a	effect	effect	effect
Marker b	no effect	no effect	no effect
Marker c	effect	no effect	no effect
Marker d	no effect	effect	no effect
Marker e	no effect	no effect	effect
Marker f	effect	effect	no effect
Marker g	effect	no effect	effect
Marker h	no effect	effect	effect

This set can be split into:
- the **configurations of interest** $\mathcal{C}_1$, and
- the configurations under the null $\mathcal{C}_0$.

For each marker m , we have access to the **p-value** $\boldsymbol{P_{mk}}$ from the association test of marker m in condition k.

We define the **z-score** Z_{mk} as :

$$\boldsymbol{Z_{mk} = -\Phi^{-1}(P_{mk})}$$

where Φ stands for the standard Gaussian cumulative distribution function.

Model

Each marker m is described by a **configuration label** $\boldsymbol{L_m \in \mathcal{C}}$.
And each **z-score profile** $\boldsymbol{Z_m = (Z_{m1}, \dots, Z_{mK})}$ arises from a mixture model with $\mathbf{2^K}$ components defined as follows:

$$\boldsymbol{Z_m \sim \sum_{c \in \mathcal{C}} w_c \psi_c}$$

where :
- ψ_c is the distribution of Z_m conditionally on $L_m = c$
- $w_c = \mathbb{P}(L_m = c)$

We assume that all distributions ψ_c have the following form:

$$\psi_c^{\theta}(Z_m) = \prod_{k:c_k=0} f_0^k(Z_{mk}) \prod_{k:c_k=1} f_1^k(Z_{mk})\, c_\theta(F_{c_1}^1(Z_{m1}), \dots, F_{c_K}^K(Z_{mK}))$$

where :
- f_0^k (resp. f_1^k) are the distributions of Z_{mk} conditionally on $L_{mk} = 0$ (resp. $L_{mk} = 1$)
- F_0^k (resp. F_1^k) are the cumulative distributions of f_0^k (resp. f_1^k)
- $\boldsymbol{c_\theta}$ **is a copula of parameter** $\boldsymbol{\theta}$ accounting for **the dependency structure** between the K z-scores

Inference and testing procedure

2-step inference procedure

Step 1: infer the distributions f_1^k

Each distribution f_1^k is inferred by **analyzing each test series separately**.
We fit the following mixture model on each set of z-score $(Z_{mk})_{1 \le m \le n}$:

$$\boldsymbol{Z_{mk} \sim \pi_0^k f_0^k + (1 - \pi_0^k) f_1^k}$$

where:
- π_0^k is the null proportion relative to the test k
  - $\hat{\pi}_0^k = [n(1-\lambda)]^{-1} |\{m: P_{mk} > \lambda\}|$ with $\lambda = 0.5$
- f_0^k is the distribution of Z_{mk} conditionally on $L_{mk} = 0$
  - by definition, $f_0^k = \phi$ the **standard Gaussian distribution**
- f_1^k is the distribution of Z_{mk} conditionally on $L_{mk} = 1$
  - $\hat{f}_1^k$ is inferred in **non-parametric way** (kernel method)

Step 2: configuration priors w_c and copula parameter θ estimation

The estimates $\hat{f}_1^k$ are directly plugged into the mixture.
The inference of the weight w_c and the copula parameter θ is performed using a standard EM algorithm.

E step : Estimation of latent variables L_{mc}

$$\hat{\tau}_{mc} = \hat{\mathbb{P}}(L_m = c \,|Z_m; \hat{\theta}) = \frac{w_c \psi_c^{\hat{\theta}}(Z_m)}{\sum_{c \in \mathcal{C}} w_c \psi_c^{\hat{\theta}}(Z_m)}$$

M step : Inference of w_c and θ

$$\hat{w}_c = \frac{1}{n} \sum_m \hat{\tau}_{mc}$$

$\hat{\theta}$ is the maximum likelihood estimate

Testing procedure

Let c_m be the (unknown) configuration of the marker m, we perform the following test :

$$H_0: \{c_m \in \mathcal{C}_0\} \quad \text{vs} \quad H_1: \{c_m \in \mathcal{C}_1\}$$

Once marginals and priors are estimated, one has access to the **posteriors probabilities**

$$\boldsymbol{\hat{\tau}_m = \sum_{c \in \mathcal{C}_1} \hat{\mathbb{P}}(L_m = c \,|Z_m)}$$

That can be used as a test statistic, with associated **p-value**

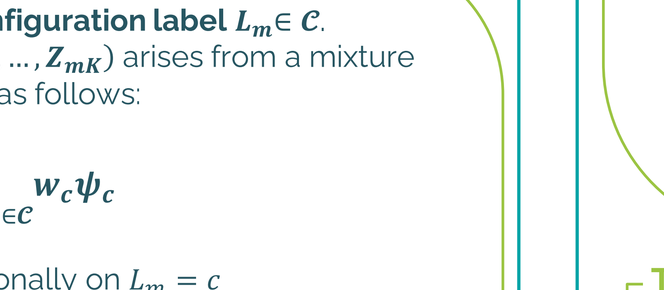
$$\boldsymbol{\widehat{\text{pval}}_m = \frac{1}{nW_0} \sum_{j=1}^{n} \mathbb{1}_{\{\hat{\tau}_j > \hat{\tau}_m\}} (1 - \hat{\tau}_j)}$$

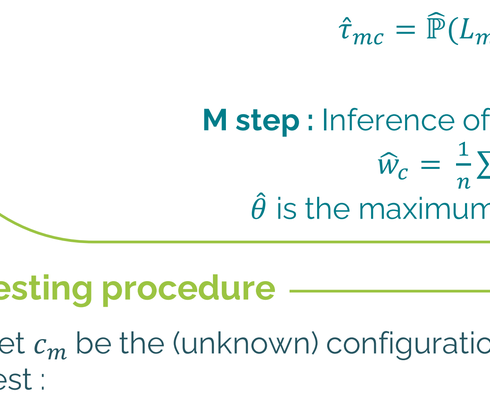
where $W_0 = \sum_{c \in \mathcal{C}_0} \hat{w}_c$

Application

Dataset from [1] consists in :
- 289 cucumber lines/landraces/hybrids
- Resistance to 6 viruses
- Genotyped at 300K SNPs

Which SNPs are associated with resistance to at least two viruses?





Conclusion

We proposed a **composite hypothesis testing procedure** using a multivariate **mixture model**.
Applications on simulated data have been carried out, with promising results both in terms of **false positive control** and **detection**
All the methods presented are available in the R package **qch** available on the CRAN.

Reference:
[1] Severine Monnot et al. Unravelling cucumber resistance to several viruses via genome-wide association studies highlighted resistance hotspots and new QTLs, Horticulture Research.

Acknowledgements: We would like to thank Séverine Monnot and Nathalie Boissot for sharing the GWAS summary statistics.

Annaïg De Walsche
annaig.de-walsche@inrae.fr
Génétique Quantitative et Evolution - Le Moulon
MIA Paris Saclay

Université Paris-Saclay, INRAE, CNRS, AgroParisTech