HAL
open science

VARIABLE SELECTION IN SPARSE MULTIVARIATE GLARMA MODELS: APPLICATION TO GERMINATION CONTROL BY ENVIRONMENT

M Gomtsyan, Céline Lévy-Leduc, Sarah Ouadah, Laure Sansonnet, C Bailly, Loïc Rajjou

To cite this version:

M Gomtsyan, Céline Lévy-Leduc, Sarah Ouadah, Laure Sansonnet, C Bailly, et al.. VARIABLE SELECTION IN SPARSE MULTIVARIATE GLARMA MODELS: APPLICATION TO GERMINATION CONTROL BY ENVIRONMENT. 2022. hal-03905876

HAL Id: hal-03905876

https://agroparistech.hal.science/hal-03905876v1

Preprint submitted on 19 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

VARIABLE SELECTION IN SPARSE MULTIVARIATE GLARMA MODELS: APPLICATION TO GERMINATION CONTROL BY ENVIRONMENT

M. GOMTSYAN, C. LÉVY-LEDUC, S. OUADAH, L. SANSONNET, C. BAILLY, AND L. RAJJOU

Abstract. We propose a novel and efficient iterative two-stage variable selection approach for multivariate sparse GLARMA models, which can be used for modelling multivariate discrete-valued time series. Our approach consists in iteratively combining two steps: the estimation of the autoregressive moving average (ARMA) coefficients of multivariate GLARMA models and the variable selection in the coefficients of the Generalized Linear Model (GLM) part of the model performed by regularized methods. We explain how to implement our approach efficiently. Then we assess the performance of our methodology using synthetic data and compare it with alternative methods. Finally, we illustrate it on RNA-Seq data resulting from polyribosome profiling to determine translational status for all mRNAs in germinating seeds. Our approach, which is implemented in the `MultiGlarmaVarSel` R package and available on the CRAN, is very attractive since it benefits from a low computational load and is able to outperform the other methods for recovering the null and non-null coefficients.

1. Introduction

Seed germination is a complex agronomic trait largely influenced by environmental conditions [44]. In cropping systems related to seed production, climatic variations experienced by the mother plant shape the physiological features of seeds, such as dormancy, longevity and germination vigor [30]. Plants grown under different temperature regimes produce seeds with contrasting germination potential. The molecular factors that may explain these phenotypes are still poorly described. It has been previously demonstrated that the translation of mRNAs is a key and essential process for the success of germination [45]. Studying translation in germinating seeds leads to a better understanding of gene expression regulation providing a direct link between transcriptome and proteome rearrangements [22]. Polysome profiling has been developed to infer the translational status of specific mRNA populations [4,6]. A rapid polysome formation occurs during early germination process. The combined approaches of polysome profiling and RNA-seq provide a unique opportunity to thoroughly investigate the translational dynamics of germinating seeds produced under different temperature regimes to highlight novel molecular mechanisms related to the physiological quality of seeds in response to the environment of the mother plant.

Key words and phrases. multivariate GLARMA, sparsity, variable selection, seed quality, gene expression.

In this paper we consider a novel multivariate count time series model to study the translational dynamics of germinating seeds. A detailed review of the main approaches for modelling multivariate count time series is available in [15]. These approaches can be classified into three model classes described hereafter.

The first class includes integer-valued autoregressive (INAR) models. The first introduction of INAR(1) processes was done by [37] and [2]. Later it was extended to pth order process in [3]. The properties of the multivariate INAR (MINAR) were derived in [20] and [35]. Further studies of MINAR were done by [40] and [41]. However, even in the univariate INAR models, the statistical inference is not straightforward, as explained in [10], and this is all the more true for higher-order INAR models.

The second class are parameter-driven models. Following the first introduction by [9], parameter-driven models are time series driven by an unobserved process. It means that the state vector evolves independently of the past history of the observations. Multivariate state space models are studied in [33] and [32]. Additional developments are found in [43]. Although these models are simple to construct, the parameter estimation is computationally expensive, see [31].

The third class of models, observation-driven models, do not suffer from computational drawback and are an alternative to parameter-driven models. In these models, the state vector depends on past observations and some additional noise. Univariate observation-driven models were first proposed by [9] and further studied by [50]. Different kinds of observation-driven models can be found in the literature: the Generalized Linear Autoregressive Moving Average (GLARMA) models introduced by [13] and further studied in [11], [12], [14] and the (log-)linear Poisson autoregressive models studied in [16], [19] and [18]. Note that GLARMA models cannot be seen as a particular case of the log-linear Poisson autoregressive models. In the past years many studies were conducted in the framework of multivariate observation-driven count time series models, many of which are based on the copula approach. An example is the Multivariate Autoregressive Conditional Double Poisson model [27], based on the double Poisson distribution with the mean vector being a VARMA process. Another model using copula [8] is developed for count time series with a domain $\mathbb{Z}^n$, $n \in \mathbb{N}$. Here the conditional probabilities of the direction of the process (whether the process is negative, positive or equal to zero) is modeled with the autoregressive conditional multinomial model (ACM). In [17], the authors impose a copula function on a vector of related continuous random variables to determine the joint distribution of the count time series. Finally, the model in [28] can be seen as a Poisson branching process model with immigration. It takes as covariates for the mean of each series at time t the counts of other series at time $t-1$.

In our context of analyzing RNA-Seq data from polysome profiling experiments, we are interested in performing variable selection in multivariate count time series. However, this problem is not addressed in the exact framework of our interest so far. There exist methods for variable selection for multivariate Poisson data using spike and slab approach [23]. The method is based on extending the Poisson Lognormal model in [1], which is a parameter-driven model, to the multivariate case and relaxing the mean-equal-variance property of

the Poisson distribution. Another study in [36] performs Bayesian variable selection in multivariate zero-inflated count data.

In this paper, we develop an observation-driven variable selection model, which is an extension of [24] to the multivariate case by considering the following multivariate GLARMA model. Given the past history $\mathcal{F}_{i,j,t-1} = \sigma(Y_{i,j,s}, s \leq t-1)$, we assume that

$$Y_{i,j,t}|\mathcal{F}_{i,j,t-1} \sim \mathcal{P}(\mu^\star_{i,j,t}), \tag{1}$$

where $\mathcal{P}(\mu)$ denotes the Poisson distribution with mean μ, $1 \leq i \leq I$, $1 \leq j \leq n_i$ and $1 \leq t \leq T$. For instance, $Y_{i,j,t}$ can be seen as a random variable modelling RNA-Seq data of the jth replication of gene t obtained in condition i. In (1)

$$\mu^\star_{i,j,t} = \exp(W^\star_{i,j,t}) \quad \text{with} \quad W^\star_{i,j,t} = \eta^\star_{i,t} + Z^\star_{i,j,t}, \tag{2}$$

where

$$Z^\star_{i,j,t} = \sum_{k=1}^{q} \gamma^\star_k E^\star_{i,j,t}, \qquad \text{with } 1 \leq q \leq \infty, \tag{3}$$

and $\eta^\star_{i,t}$, the non random part of $W^\star_{i,j,t}$, does not depend on j.

Let us denote $\boldsymbol{\eta}^\star = (\eta^\star_{1,1}, \ldots, \eta^\star_{I,1}, \eta^\star_{I,2}, \ldots, \eta^\star_{I,T})'$ the vector of coefficients corresponding to the effect of a qualitative variable on the observations. For instance, in the case of RNA-Seq data, $\eta^\star_{i,t}$ can be seen as the effect of condition i (i.e. temperature regime during seed production) on polysome-associated mRNAs t. Assume moreover that $\boldsymbol{\gamma}^\star = (\gamma^\star_1, \ldots, \gamma^\star_q)'$ is such that $\sum_{k\geq 1} |\gamma^\star_k| < \infty$, where u' denotes the transpose of u. Additionally,

$$E^\star_{i,j,t} = \frac{Y_{i,j,t} - \mu^\star_{i,j,t}}{\mu^\star_{i,j,t}} = Y_{i,j,t} \exp\big(-W^\star_{i,j,t}\big) - 1. \tag{4}$$

with $E^\star_{i,j,t} = 0$ for all $t \leq 0$ and $1 \leq q \leq \infty$. When $q = \infty$, $Z^\star_{i,j,t}$ satisfies an ARMA-like recursion in (4), because causal ARMA can be written as MA process of infinite order.

$E^\star_{i,j,t}$ in (4) corresponds to the particular case of working residuals in classical Generalized Linear Models (GLM) usually defined by $E^\star_{i,j,t} = (Y_{i,j,t} - \mu^\star_{i,j,t}){\mu^\star_{i,j,t}}^{-\lambda}$ with $\lambda = 1$. The resulting model defined by Equations (1), (2), (3) and (4) is referred to as multivariate GLARMA model.

The main goal of this paper is to introduce a novel variable selection approach in the deterministic part ($\boldsymbol{\eta}^\star$) of the sparse multivariate GLARMA model that is defined in Equations (1), (2), (3) and (4), where the vector of the $\eta^\star_{i,t}$'s is sparse. Sparsity means that many $\eta^\star_{i,t}$'s are null, and thus just a few coefficients are significant. The novel approach that we propose combines a procedure for estimating the ARMA part coefficients to take into account the dependence that may exist in the data with regularized methods designed for GLM as those proposed by [21] and [26].

The paper is organized as follows. Firstly, we propose a novel two-stage estimation procedure in multivariate GLARMA models in Section 2.1 and Section 2.2. It consists of first estimating the ARMA coefficients and then estimating the $\eta^\star_{i,t}$'s by using a regularized approach. The practical implementation of our approach is given in Section 2.3. Next, in Section 3, we provide numerical experiments to illustrate our method and compare its

performance to alternative approaches on finite sample size data. Additionally, in Section 4, we apply our method to RNA-Seq data from polysome profiling experiments to determine translational status for all mRNAs in germinating seeds.

2. Statistical Inference

Extending the estimation procedure existing in standard univariate GLARMA models described in [11] and [12] to the multivariate case would consist in estimating $\boldsymbol{\delta}^\star = (\boldsymbol{\eta}^{\star\prime}, \boldsymbol{\gamma}^{\star\prime})$, where $\boldsymbol{\eta}^{\star\prime}$ is the vector of coefficients and $\boldsymbol{\gamma}^{\star\prime}$ is the vector of the ARMA part coefficients by $\widehat{\boldsymbol{\delta}}$, which is defined as follows:

$$\widehat{\boldsymbol{\delta}} = \arg\max_{\boldsymbol{\delta}} L(\boldsymbol{\delta}). \tag{5}$$

In (5), L is based on the conditional log-likelihood and is defined by:

$$L(\boldsymbol{\delta}) = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{t=1}^{T}(Y_{i,j,t}W_{i,j,t}(\boldsymbol{\delta}) - \exp(W_{i,j,t}(\boldsymbol{\delta})),$$

where $W_{i,j,t}(\boldsymbol{\delta})$ is defined as in (2)-(4):

$$W_{i,j,t}(\boldsymbol{\delta}) = \eta_{i,t} + \sum_{k=1}^{q}\gamma_k E_{i,j,t}(\boldsymbol{\delta}) \text{ with } E_{i,j,t}(\boldsymbol{\delta}) = Y_{i,j,t}\exp\big(-W_{i,j,t}(\boldsymbol{\delta})\big) - 1. \tag{6}$$

However, this procedure is not designed for dealing with a sparse framework where many components of $\boldsymbol{\eta}^\star$ are null. This is the reason why we propose hereafter a novel two-stage estimation procedure described in the following sections.

2.1. Estimation of $\boldsymbol{\gamma}^\star$.

In our estimation procedure, we use the Newton-Raphson algorithm to obtain $\widehat{\boldsymbol{\gamma}}$ based on the following recursion. For $r \geq 1$, starting from the initial value $\boldsymbol{\gamma}^{(0)} = (\gamma_1^{(0)}, \ldots, \gamma_q^{(0)})'$ and $\boldsymbol{\eta}^{(0)} = (\eta_{1,1}^{(0)}, \ldots, \eta_{I,1}^{(0)}, \eta_{I,2}^{(0)}, \ldots, \eta_{I,T}^{(0)})'$:

$$\boldsymbol{\gamma}^{(r)} = \boldsymbol{\gamma}^{(r-1)} - \frac{\partial^2 L}{\partial\boldsymbol{\gamma}'\partial\boldsymbol{\gamma}}\big(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma}^{(r-1)}\big)^{-1}\frac{\partial L}{\partial\boldsymbol{\gamma}}\big(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma}^{(r-1)}\big). \tag{7}$$

To obtain $\frac{\partial L}{\partial\boldsymbol{\gamma}}$, we shall use that

$$\frac{\partial L}{\partial\boldsymbol{\gamma}}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma}) = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{t=1}^{T}(Y_{i,j,t} - \exp(W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma})))\frac{\partial W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}},$$

where details for computing the first derivative of $W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ are given in Appendix A.1.1.

Concerning the Hessian of L, it can be obtained as follows:

$$\frac{\partial^2 L}{\partial \boldsymbol{\gamma}' \partial \boldsymbol{\gamma}}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \sum_{t=1}^{T} (Y_{i,j,t} - \exp(W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma}))) \frac{\partial^2 W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}' \partial \boldsymbol{\gamma}}$$
$$- \sum_{i=1}^{I} \sum_{j=1}^{n_i} \sum_{t=1}^{T} \exp(W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma})) \frac{\partial W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} \frac{\partial W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}},$$

where details for computing the second derivative of $W_{i,j,t}(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ are given in Appendix A.1.2.

2.2. **Variable selection in $\boldsymbol{\eta}^\star$ estimation.**

2.2.1. *Variable selection criterion.* To perform variable selection in the $\boldsymbol{\eta}^\star$ of Model (2)–(4), namely to obtain a sparse estimator of $\boldsymbol{\eta}^\star$, we shall use a regularized approach inspired by [21] for fitting sparse generalized linear models. It consists in penalizing (with an ℓ_1 penalty) a quadratic approximation to the log-likelihood obtained by a second order Taylor expansion. Using $\boldsymbol{\eta}^{(0)}$ and $\widehat{\boldsymbol{\gamma}}$ defined in Section 2.1, we obtain the quadratic approximation as follows:

$$\begin{aligned}\tilde{L}(\boldsymbol{\eta}) &\coloneqq L(\eta_{1,1}, \ldots, \eta_{I,1}, \eta_{I,2}, \ldots, \eta_{I,T}, \widehat{\boldsymbol{\gamma}}) \\ &= \tilde{L}(\boldsymbol{\eta}^{(0)}) + \frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})(\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)}) \\ &\qquad + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)})' \frac{\partial^2 L}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})(\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)}),\end{aligned}$$

where

$$\frac{\partial L}{\partial \boldsymbol{\eta}} = \Big(\frac{\partial L}{\partial \eta_{1,1}}, \ldots, \frac{\partial L}{\partial \eta_{I,1}}, \frac{\partial L}{\partial \eta_{I,2}}, \ldots, \frac{\partial L}{\partial \eta_{I,T}} \Big)$$

and

$$\frac{\partial^2 L}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'} = \Big(\frac{\partial^2 L}{\partial \eta_{i_0,t_0} \partial \eta_{i_1,t_1}} \Big)_{\substack{1 \le i_0, i_1 \le I \\ 1 \le t_0, t_1 \le T}}.$$

Let $U \Lambda U'$ be the singular values decomposition of the positive semidefinite symmetric matrix $-\frac{\partial^2 L}{\partial \boldsymbol{\eta} \boldsymbol{\eta}'}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})$ and $\boldsymbol{\nu} - \boldsymbol{\nu}^{(0)} = U'(\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)})$. Therefore, the quadratic approximation is

$$\tilde{L}(\boldsymbol{\eta}) = \tilde{L}(\boldsymbol{\eta}^{(0)}) + \frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}}) U(\boldsymbol{\nu} - \boldsymbol{\nu}^{(0)}) - \frac{1}{2}(\boldsymbol{\nu} - \boldsymbol{\nu}^{(0)})' \Lambda (\boldsymbol{\nu} - \boldsymbol{\nu}^{(0)}). \tag{8}$$

In order to obtain a sparse estimator of $\boldsymbol{\eta}^\star$ we use $\widehat{\boldsymbol{\eta}}(\lambda)$ defined by minimizing the following criterion:

$$\widehat{\boldsymbol{\eta}}(\lambda) = \arg\min_{\boldsymbol{\eta}} \{ -\tilde{L}_Q(\boldsymbol{\eta}) + \lambda \|\boldsymbol{\eta}\|_1 \}, \tag{9}$$

for a positive λ, where $\|\eta\|_1 = \sum_{i=1}^{I}\sum_{t=1}^{T}|\eta_{i,t}|$ and $\tilde{L}_Q(\boldsymbol{\eta})$ denotes the quadratic approximation of the log-likelihood. This quadratic approximation is defined by

$$-\tilde{L}_Q(\boldsymbol{\eta}) = \frac{1}{2}\|\mathcal{Y} - \mathcal{X}\boldsymbol{\eta}\|_2^2 \tag{10}$$

with

$$\mathcal{Y} = \Lambda^{1/2}U'\boldsymbol{\eta}^{(0)} + \Lambda^{-1/2}U'\Big(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})\Big)', \quad \mathcal{X} = \Lambda^{1/2}U', \tag{11}$$

where $\|\cdot\|_2$ is the ℓ_2 norm.

2.2.2. *Criterion derivation.* Let us now explain how the expression of $\tilde{L}_Q$ given in (10) was obtained. By (8), we get

$$\begin{aligned}
\tilde{L}(\boldsymbol{\eta}) &= \tilde{L}(\boldsymbol{\eta}^{(0)}) + \sum_{i=1}^{I}\sum_{t=1}^{T}\left(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})U\right)_{i,t}(\nu_{i,t} - \nu_{i,t}^{(0)}) - \frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\lambda_{i,t}(\nu_{i,t} - \nu_{i,t}^{(0)})^2 \\
&= \tilde{L}(\boldsymbol{\eta}^{(0)}) - \frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\lambda_{i,t}\left(\nu_{i,t} - \nu_{i,t}^{(0)} - \frac{1}{\lambda_{i,t}}\left(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})U\right)_{i,t}\right)^2 \\
&\quad + \sum_{i=1}^{I}\sum_{t=1}^{T}\frac{1}{2\lambda_{i,t}}\left(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})U\right)_{i,t}^2,
\end{aligned} \tag{12}$$

where the $\lambda_{i,t}$'s are the diagonal terms of Λ.

Since only the second term of (12) depends on $\boldsymbol{\eta}$,

$$\begin{aligned}
-\tilde{L}_Q(\boldsymbol{\eta}) &= \frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\lambda_{i,t}\left(\nu_{i,t} - \nu_{i,t}^{(0)} - \frac{1}{\lambda_{i,t}}\left(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})U\right)_{i,t}\right)^2 \\
&= \frac{1}{2}\left\|\Lambda^{1/2}\left(\boldsymbol{\nu} - \boldsymbol{\nu}^{(0)} - \Lambda^{-1}\left(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})U\right)'\right)\right\|_2^2 \\
&= \frac{1}{2}\left\|\Lambda^{1/2}U'(\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)}) - \Lambda^{-1/2}U'\left(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})\right)'\right\|_2^2 \\
&= \frac{1}{2}\left\|\Lambda^{1/2}U'(\boldsymbol{\eta}^{(0)} - \boldsymbol{\eta}) + \Lambda^{-1/2}U'\left(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})\right)'\right\|_2^2 \\
&= \frac{1}{2}\|\mathcal{Y} - \mathcal{X}\boldsymbol{\eta}\|_2^2,
\end{aligned}$$

where

$$\mathcal{Y} = \Lambda^{1/2}U'\boldsymbol{\eta}^{(0)} + \Lambda^{-1/2}U'\Big(\frac{\partial L}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}^{(0)}, \widehat{\boldsymbol{\gamma}})\Big)', \quad \mathcal{X} = \Lambda^{1/2}U'.$$

2.2.3. *Stability selection.* To obtain the final estimator $\widehat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}^\star$, we shall consider an approach called stability selection devised by [38], which guarantees the robustness of the selected variables. This approach can be described as follows. The vector $\mathcal{Y}$ defined in (11) is randomly split into several subsamples of size $IT/2$, corresponding to half of the length of $\mathcal{Y}$. The number of subsamples is equal to 1000 in our numerical experiments. For each subsample $\mathcal{Y}^{(s)}$ and the corresponding design matrix $\mathcal{X}^{(s)}$, Criterion (9) is applied with a given λ, where $\mathcal{Y}$ and $\mathcal{X}$ are replaced by $\mathcal{Y}^{(s)}$ and $\mathcal{X}^{(s)}$, respectively. For each subsampling, the indices i and t of the non-null $\widehat{\eta}_{i,t}$ are stored. In the end, we calculate the frequency of index selection, namely the number of times each couple of indices was selected divided by the number of subsamples. For a given threshold, in the final set of selected variables, we keep the ones whose indices have a frequency larger than this threshold. Concerning the choice of λ, we shall consider the smallest element of the grid of λ provided by the R `glmnet` package. It is also possible to use the one obtained by cross-validation (Chapter 7 of [25]). However, based on our experiments, choosing the minimal λ of the grid led to better results.

2.3. **Practical implementation.** In practice, the previous approach can be summarized as follows.

- Initialization. We take for $\boldsymbol{\eta}^{(0)}$ the estimator of $\boldsymbol{\eta}^\star$ obtained by fitting a GLM to the observations $Y_{1,1,1}, \ldots, Y_{I,n_I,T}$ thus ignoring the ARMA part of the model. For $\boldsymbol{\gamma}^{(0)}$, we take the null vector.
- Newton-Raphson algorithm. We use the recursion defined in (7) with the initialization $(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma}^{(0)})$ obtained in the previous step and we stop at the iteration R such that $\|\boldsymbol{\gamma}^{(R)} - \boldsymbol{\gamma}^{(R-1)}\|_\infty < 10^{-6}$.
- Variable selection. To obtain a sparse estimator of $\boldsymbol{\eta}^\star$, we use Criterion (9), where $\boldsymbol{\eta}^{(0)}$ and $\widehat{\boldsymbol{\gamma}}$ appearing in (11) are replaced by $\boldsymbol{\eta}^{(0)}$ and $\boldsymbol{\gamma}^{(R)}$ obtained in the previous steps. We thus get $\widehat{\boldsymbol{\eta}}$ by using the stability selection approach described in Section 2.2.3.

This procedure can be improved by iterating the Newton-Raphson algorithm and Variable selection steps. More precisely, let us denote by $\boldsymbol{\eta}_1^{(0)}$, $\gamma_1^{(R_1)}$ and $\widehat{\boldsymbol{\eta}}_1$ the values of $\boldsymbol{\eta}^{(0)}$, $\gamma^{(R)}$ and $\widehat{\boldsymbol{\eta}}$ obtained in the three steps described above at the first iteration. At the second iteration, $(\boldsymbol{\eta}^{(0)}, \boldsymbol{\gamma}^{(0)})$ appearing in the Newton-Raphson algorithm step is replaced by $(\widehat{\boldsymbol{\eta}}_1, \gamma_1^{(R_1)})$. At the end of this second iteration, $\widehat{\boldsymbol{\eta}}_2$ and $\gamma_2^{(R_2)}$ denote the obtained values of $\widehat{\boldsymbol{\eta}}$ and $\gamma^{(R)}$, respectively. This approach is iterated until the stabilisation of $\gamma_k^{(R_k)}$.

3. Numerical experiments

This section aims at investigating the performance of our method, which is implemented in the R package `MultiGlarmaVarSel` available on the CRAN (Comprehensive R Archive Network). We study it both from a statistical and a numerical point of view, using synthetic data generated from the model defined by (1)–(4), where $n_i = J$ for all i. The different simulation settings that we considered are given in Table 1. In all the experiments we set

the number of non-null coefficients in $\boldsymbol{\eta}^\star$ to 10 and the number of simulations to 50. The non-null values of $\boldsymbol{\eta}^\star$ range from 0.41 to 2.62.

T	J	I	$q^\star$	$\boldsymbol{\gamma}^\star$
50	10	3	1	0.5
50	100	3	1	0.5
200	10	3	1	0.5
200	100	3	1	0.5
50	10	3	2	0.2, 0.5
50	100	3	2	0.2, 0.5
200	10	3	2	0.2, 0.5
200	100	3	2	0.2, 0.5

TABLE 1. Parameters of simulated datasets used in the experiments.

3.1. **Statistical performance.**

3.1.1. *Estimation of* $\boldsymbol{\eta}^\star$.

Support recovery of $\boldsymbol{\eta}^\star$. In this section, we assess the performance of our methodology in terms of support recovery, namely the identification of the non-null coefficients of $\boldsymbol{\eta}^\star$, and of the estimation of $\boldsymbol{\gamma}^\star$.

Figures 1 and 3 display the maximum difference between TPR (True Positive Rates, namely the proportion of non-null coefficients correctly estimated as non-null) and FPR (False Positive Rates, namely the proportion of null coefficients estimated as non-null) for $q^\star = 1$ and $q^\star = 2$ correspondingly. For each simulation, we considered 9 thresholds ranging from 0.1 to 0.9 in the stability selection step. For each threshold, we calculated the maximum difference between TPR and FPR. Then, from the 9 differences, we took the largest one, which is the best result. It means we did not use the same threshold from one simulation to another. We considered five different approaches: our method with $q = 0$, $q = 1$ and $q = 2$, classical LASSO for Poisson distribution, and our method where we took $\boldsymbol{\gamma}^\star$ instead of estimating it. More precisely, classical LASSO for Poisson distribution consists in applying the `glmnet` R package dedicated to Poisson distribution to the $Y_{i,j,t}$'s for each t. We did not compare our method with `glarma` package because it does not support the multivariate setting.

In Figures 1 and 3 the closer the maximum difference between TPR and FPR is to 1, the better is the performance of the method. Our approach with $q = 1$ and $q = 2$ outperforms classical LASSO and the estimation with $q = 0$. We notice that when J is larger, the estimation is better both for $T = 50$ and $T = 200$. Additionally, the performance for the simulation frameworks with $T = 50$ is better than for the ones with $T = 200$. In general, our estimation is close to the one with the true value of $\boldsymbol{\gamma}^\star$.

Figures 2 and 4 display the error bars of TPR and FPR of our method with respect to the threshold for $q = 1$ and $q = 2$, respectively. More precisely, the threshold 0.6 achieves

a satisfactory trade-off between the TPR and the FPR. The best trade-offs are achieved for $T = 50$ and $J = 100$, for both $q = 1$ and $q = 2$.

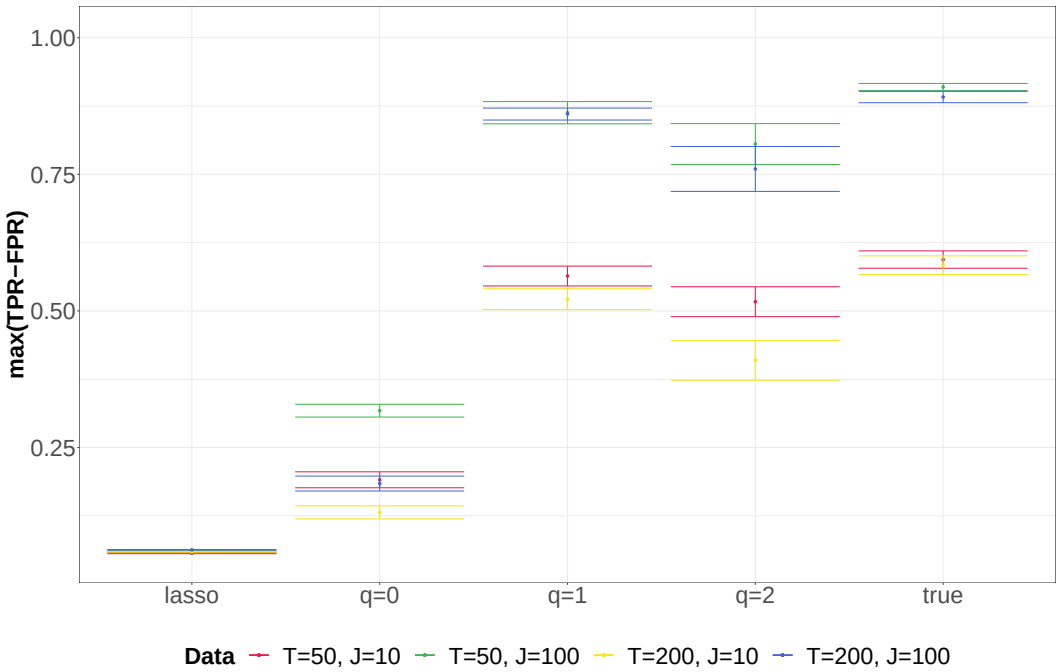

FIGURE 1. Error bars of the maximum difference between TPR and FPR for different thresholds associated to the support recovery of $\boldsymbol{\eta}^\star$ with 5 approaches for 4 simulation frameworks when $I = 3$, $q^\star = 1$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations.

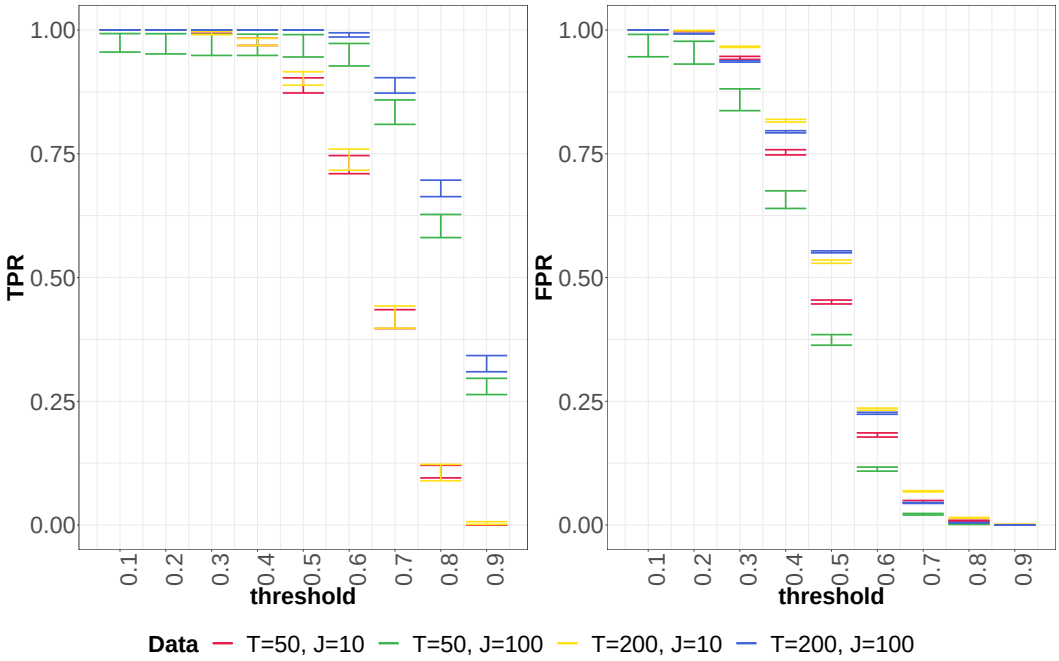

FIGURE 2. Error bars of the TPR and FPR for different thresholds associated to the support recovery of $\boldsymbol{\eta}^\star$ estimated with $q = 1$ for 4 different simulation frameworks with respect to the thresholds when $I = 3$, $q^\star = 1$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations.

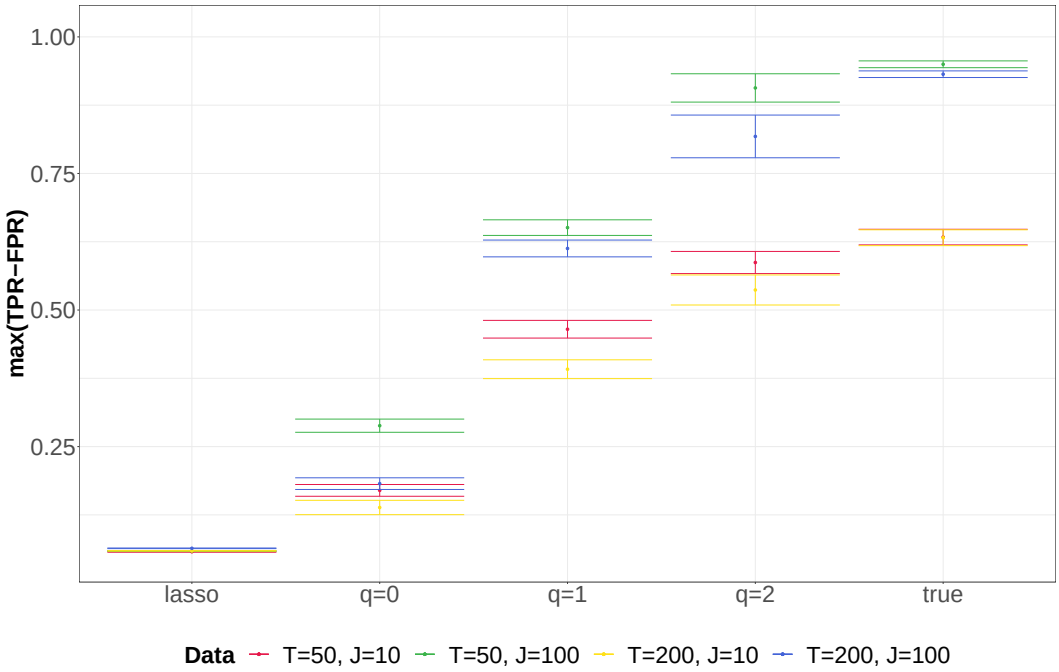

FIGURE 3. Error bars of the maximum difference between TPR and FPR for different thresholds associated to the support recovery of $\boldsymbol{\eta}^\star$ with 5 approaches for 4 simulation frameworks when $I = 3$, $q^\star = 2$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations.

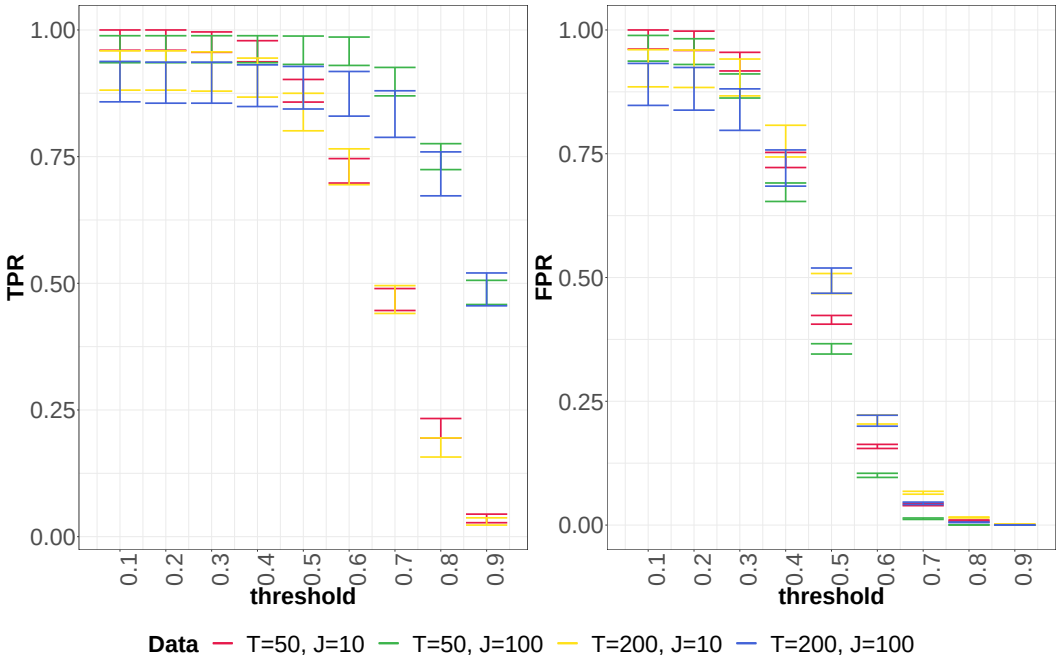

FIGURE 4. Error bars of the TPR and FPR for different thresholds associated to the support recovery of $\boldsymbol{\eta}^\star$ estimated with $q = 2$ for 4 different simulation frameworks with respect to the thresholds when $I = 3$, $q^\star = 2$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations.

Sign consistency of the estimation of $\boldsymbol{\eta}^\star$. In Figures 5 and 6 we illustrate the TPR of sign recovery of $\boldsymbol{\eta}$. For these figures, we looked at the estimation with the threshold of 0.6. The sign recovery is considered as true positive if for negative (positive) it is estimated with a negative (positive) sign and if 0 is estimated as 0. Here again, we can conclude that the best results are obtained for $J = 100$, similar to the support recovery of $\boldsymbol{\eta}^\star$.

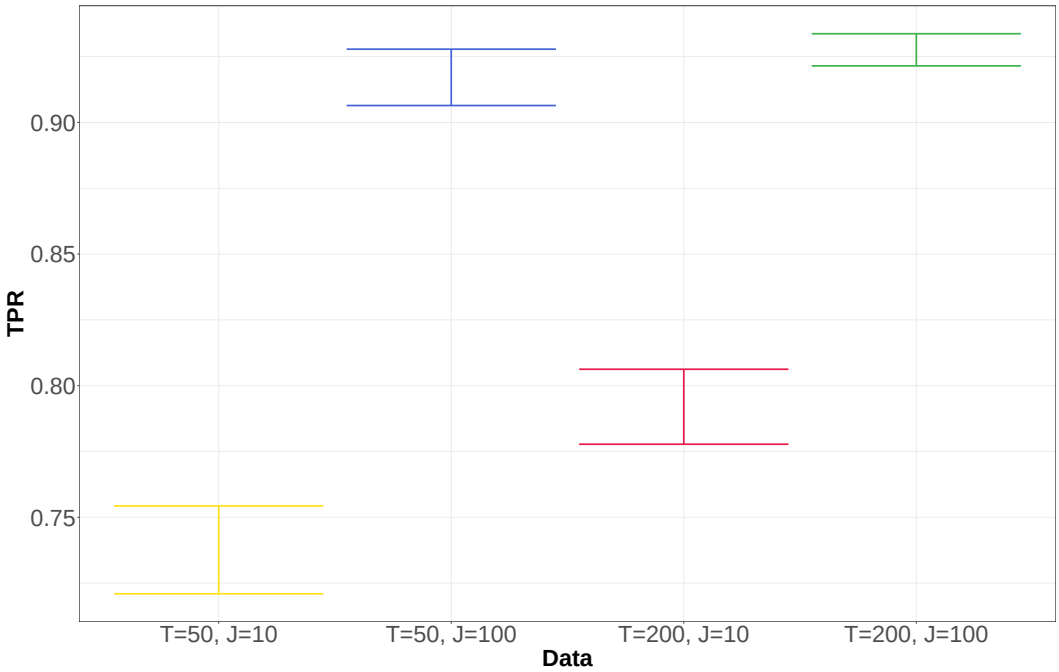

FIGURE 5. Error bars of the TPR of sign recovery of $\boldsymbol{\eta}^\star$ estimated with $q = 1$ for 4 simulation frameworks when $I = 3$, $q^\star = 1$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations.

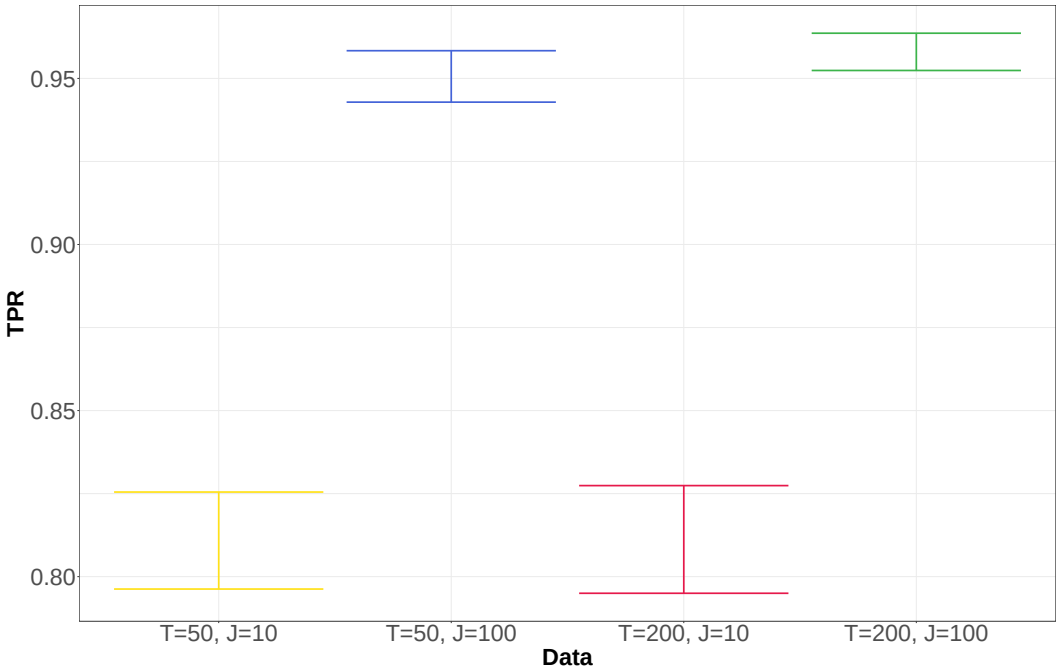

FIGURE 6. Error bars of the TPR of sign recovery of $\boldsymbol{\eta}^\star$ estimated with $q = 2$ for 4 simulation frameworks when $I = 3$, $q^\star = 2$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations.

3.1.2. *Estimation of* $\boldsymbol{\gamma}^\star$. In this section we investigate the performance of the method for the estimation of $\boldsymbol{\gamma}^\star$ for the simulation frameworks of Table 1. In Figures 7 (resp. 8), boxplots for the estimations of $\boldsymbol{\gamma}^\star$ in (3) are displayed for $q^\star = 1$ (resp. $q^\star = 2$). We can see from these figures that when $J = 10$, both for $T = 50$ and $T = 200$, iterating our approach does not improve the results. However, this is not the case for $J = 100$: the estimation of $\boldsymbol{\gamma}^\star$ improves at the second iteration. In the Appendix A.2, we present additional figures for the settings $T = 50$ with $J = 10$ and $J = 100$, and 10 iterations. These plots justify that for a small value of J iterating the method does not improve the estimation, whereas for a large value of J the estimation stabilises and results become better.

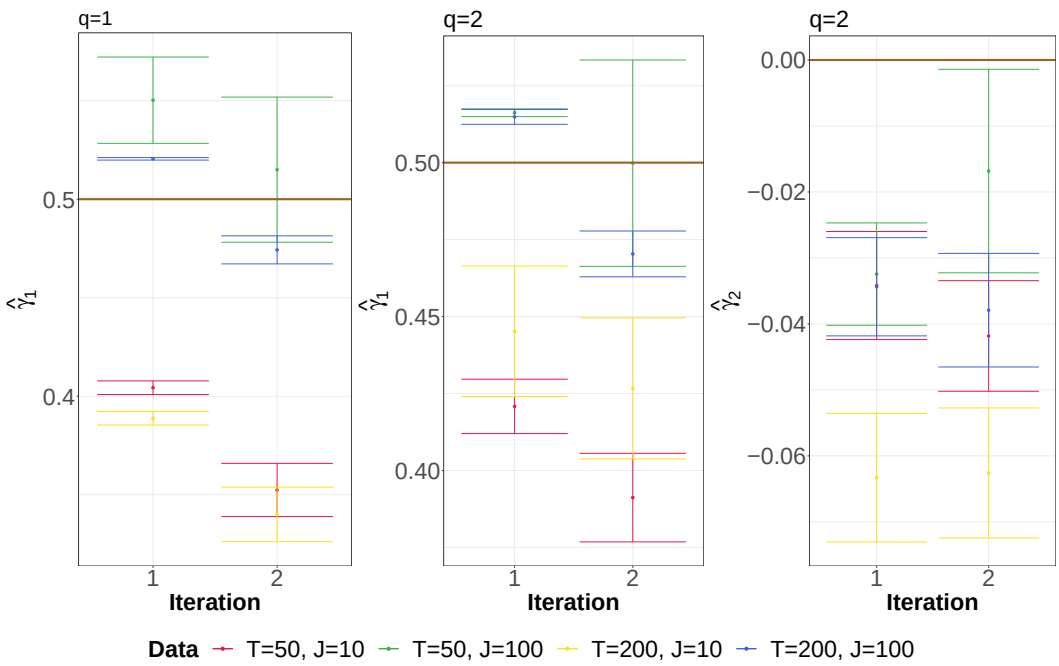

FIGURE 7. Boxplots for the estimations of $\boldsymbol{\gamma}^\star$ in Model (3) for 4 different simulation frameworks when $I = 3$, $q^\star = 1$, $\gamma^\star = 0.5$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations obtained by $q = 1$ and $q = 2$. The horizontal lines correspond to the values of the $\gamma_i^\star$'s.

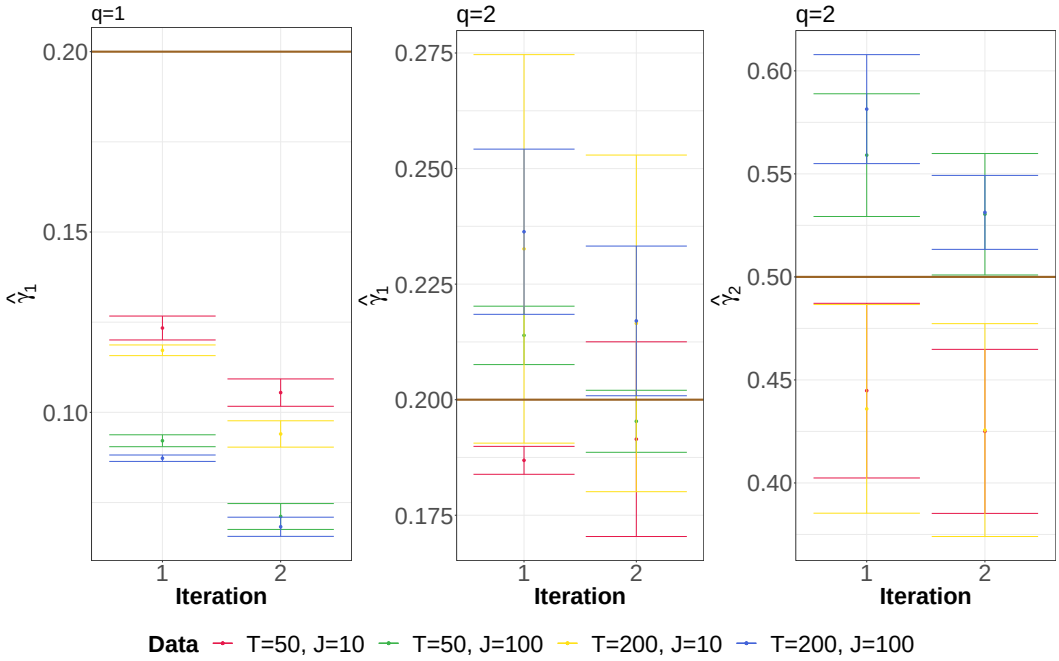

FIGURE 8. Boxplots for the estimations of $\boldsymbol{\gamma}^\star$ in Model (3) for 4 different simulation frameworks when $I = 3$, $q^\star = 2$, $\gamma_1^\star = 0.2$, $\gamma_1^\star = 0.5$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations obtained by $q = 1$ and $q = 2$. The horizontal lines correspond to the values of the $\gamma_i^\star$'s.

3.2. **Numerical performance.** Figure 9 displays the means of the computational times of our approach implemented in the R package `multiGlarmaVarSel` for different simulation frameworks. The timings were obtained on a workstation with 32GB of RAM and Intel Core i7-9700 (3.00GHz) CPU. We can see from this figure that the computational time goes from 10 seconds to 5 minutes to process the data for a given threshold and one iteration, when we increase T from 50 to 200 and when $q = 1$, 2 or 3.

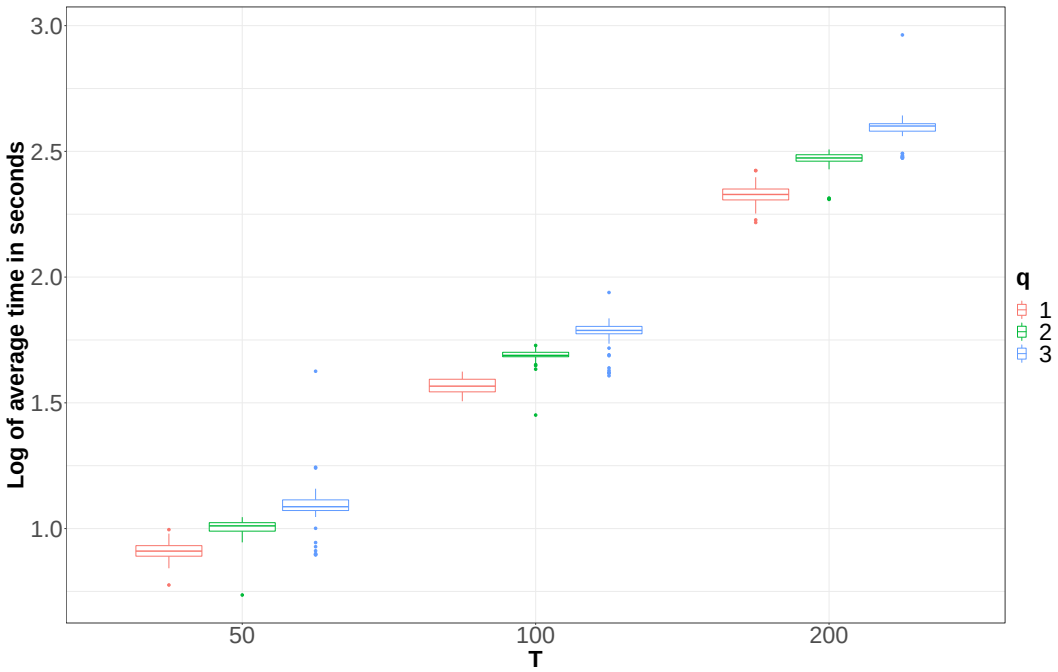

FIGURE 9. Boxplots of the $\log_{10}$ computational times in seconds in the case where $I = 3$, $J = 10$, $q = q^\star$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and different values of T and $q^\star$, a given threshold and one iteration. We performed 50 simulations.

4. Application to RNA-Seq data

4.1. **Biological context and modelling.** In order to address the issue related to the influence of the thermal environment of the mother plant on the germination potential of the progeny at the scale of translational regulation, the model plant, Arabidopsis thaliana (Col-0 accession), was cultivated under three temperature regimes (Low, 14–16°C; Medium, 18–22°C ; Elevated, 25–28°C) under a long-day photoperiod. As described in the introduction, the lower the cultivation temperature, the deeper the dormancy of the harvested seeds [42]. Seeds produced under the three temperature regimes were placed under germination conditions at low temperature (10°C) to avoid thermo-dormancy. After 72 hours of imbibition, before the first radicle protrusion, the seeds were collected for molecular analyses. Polysome profiling of Arabidopsis seeds was performed as described by [6]. The purified polysomal mRNAs and non-polysomal mRNAs were analysed by RNA-sequencing [47].

The data consists of 26725 gene expressions observed in 3 conditions of temperature with 3 replicates for 5 chromosomes and two mRNA populations (polysomal and non-polysomal). Since the gene expresssions are integer values and since there may be some dependence between them we modeled this data by using Model (1)–(4). In this model, $Y_{i,j,t}$ is a random variable describing the expression of the jth replication of gene t in temperature i with $I = 3$, $n_i = J = 3$ for all i and $T = 26725$. Moreover, $\eta^\star_{i,t}$ corresponds to the effect of temperature i on gene t.

In this framework, where I and J are very small, according to the numerical results obtained in Section 3, the value of T has to be reduced to obtain satisfactory statistical performance. This is the reason why we preprocessed the data as follows. For each mRNA population and each of the five chromosomes, we used a one-way ANOVA GLM with Poisson distribution to identify the genes on which the conditions have an influence. We kept the genes for which the p-value of the corresponding t-test is smaller than $1/T_{c,pop}$ where $T_{c,pop}$ is the number of genes present in the chromosome c and in the mRNA population pop ($T_{c,pop}$ ranges from 4074 to 7003). With this filtering, the new values of $T_{c,pop}$ for each mRNA population are given in Table 3 where "Non-poly" (resp. "Poly") refers to non-polysomal (resp. polysomal).

4.2. **Results obtained with the `multiGlarmaVarSel` R package.** This section provides the results obtained by applying our methodology to each of the five chromosomes of each mRNA population. Since I and J are very small in this application, we only focus on $q = 1$. Based on the results obtained in 3.1.2, we only ran one iteration of our procedure.

The estimation of $\gamma^\star_1$ for the polysomal and non-polysomal mRNA populations and the different chromosomes are given in Table 2. We can see from this table that the estimations are similar for the two populations except for chromosomes 3 and 4.

Chromosome	$\widehat{\gamma}_1$ for non-poly population	$\widehat{\gamma}_1$ for poly population
1	-0.00369	0.06234
2	0.05617	0.04312
3	0.03314	0.79662
4	0.21809	0.09473
5	0.00574	0.00159

TABLE 2. Estimation of $\gamma_1^\star$ for the 5 chromosomes and the two mRNA populations.

Table 3 provides the number of genes selected by our procedure in the two mRNA populations for the different chromosomes. We can see from this table that the number of selected genes that are common in the two mRNA populations ranges from 1 to 9 and is the highest for the first chromosome.

Chromosome	$T_{c,\text{Non-poly}}$	Selected genes in non-poly population	$T_{c,\text{Poly}}$	Selected genes in poly population	Intersection
1	70	59	25	20	9
2	59	44	16	12	6
3	37	31	4	1	1
4	41	25	18	12	4
5	43	39	9	7	1

TABLE 3. Number of genes selected by our procedure with $q = 1$ in the two mRNA populations as well as those that are common in the both (Intersection column).

Figure 10 displays the average gene expression values of the 3 replications for each temperature condition (Low, Medium, Elevated). The genes displayed in this figure are obtained by our selection procedure and are common to the two populations (polysomal and non-polysomal). We can see from this figure that the temperature conditions may have a different impact on the expression of the genes. This is the case, for instance, for AT1G48130, AT2G33830 and AT1G14950, on which the low temperature has a positive effect on their expression. Figures 14, 15, 16, 17 and 18 in Appendix A.3 display the average gene expression values of all the selected genes for non-polysomal and polysomal populations.

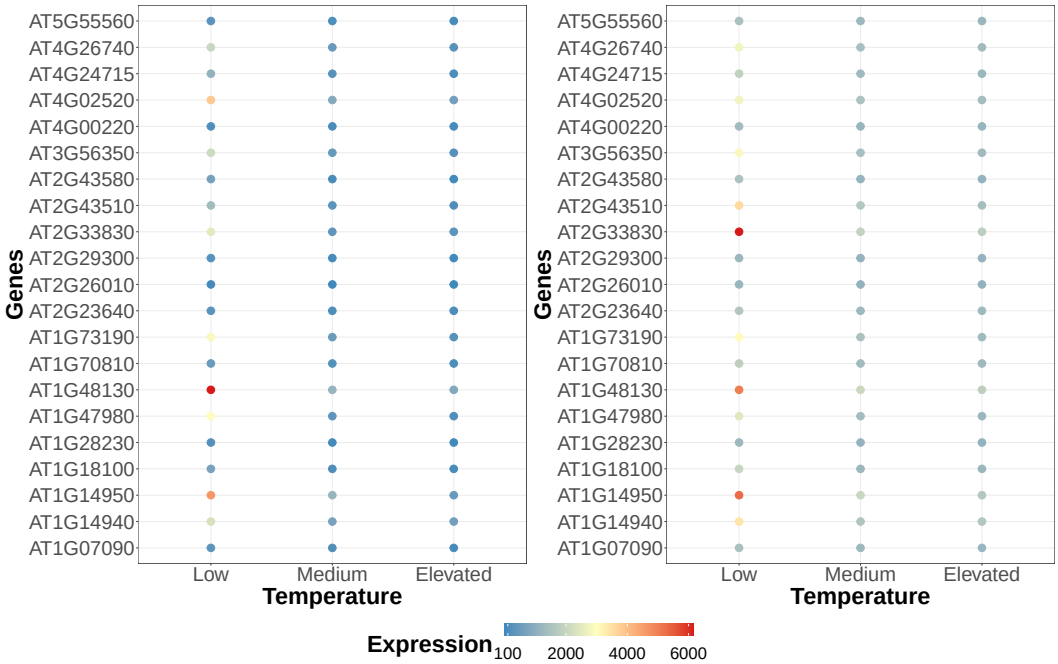

FIGURE 10. Average gene expression values of 3 replications at each temperature condition (Low, Medium, Elevated) for the non-polysomal (resp. polysomal) population on the left (resp. right). The genes displayed in this figure are obtained by our selection procedure and are common to both populations.

4.3. **Biological results.** The results allowed the selection of 198 non-polysomal mRNAs and 52 polysomal mRNAs accumulated differentially in the germinating seeds under the influence of the growth temperature of the mother plant. Overall it concerns 229 genes, and 21 are shared in both populations of mRNA. It is worth noting that several genes previously described to be involved in the control of seed dormancy and germination were selected by the present statistical method, such as RDO5 (AT4G11040, [49]), DRM2 (AT2G33830, [29]), DOG1-like 3 (AT4G18690, [7]), MFT (AT1G18100, [48]), XERICO (AT2G04240, [34]) or HAI3 (AT2G29380, [39]). Thus, these genes also seem to be involved in the modulation of germination potential induced by the thermal environment of the mother plant. A gene ontology (GO) analysis from the 229 genes revealed that the top 5 of the biological processes affected were the response to stress (GO:0006950), response to oxygen-containing compound (GO:1901700), defence response (GO:0006952), response to hormone (GO:0009725) and signal transduction (GO:0007165) (Figure 11). This is a pioneering observation showing that the environment of the mother plant not only influences the germinative potential of offspring seeds through hormonal and redox regulation [5,46], but also the ability of germinating seeds to cope with biotic and abiotic stresses. Interestingly, 23 transcription factors were selected by the statistical approach

(Table 4). These genes could represent key regulators of the modulation of seed physiological quality in response to various types of biotic and environmental stress during seed production and/or during germination. These results open the door for further research addressing the question of the control of mRNA metabolism during seed germination, notably concerning the selectivity of translational control. The germinating seed is undoubtedly a relevant biological model for exploring the precise mechanisms of combined transcriptional and translational regulation related to gene expression ending with the production of functional protein.

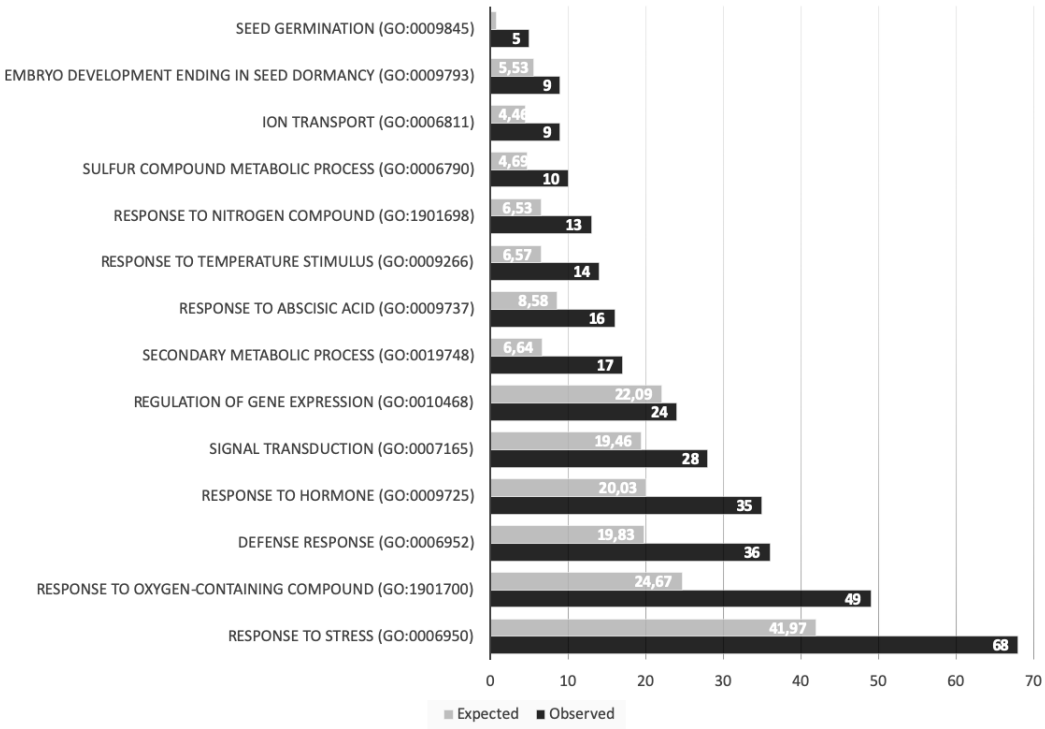

FIGURE 11. Gene ontology (GO) term enrichment analysis of the selected 229 genes based on polysomal-associated mRNA differentially accumulated in germinating seeds produced beforehand under different temperature regimes. Data from PANTHER overrepresentation test (http://www.geneontology.org); Arabidopsis thaliana GO database released March 22nd 2022 (DOI: 10.5281/zenodo.6399963). Black bars: observed gene number in the selection; grey bars: expected gene number from the reference Arabidopsis genome.

TF Family Name	AGI	Gene TAIR Curator Summary	Gene TAIR Short Description	Gene TAIR Aliases
AP2-EREBP	AT1G28360	Encodes a member of the ERF (ethylene response factor) subfamily B-1 of ERF/AP2 transcription factor family (ERF12). The protein contains one AP2 domain.	ERF domain protein 12	ATERF12, ERF12
AP2-EREBP	AT1G75490	Encodes a member of the DREB subfamily A-2 of ERF/AP2 transcription factor family. The protein contains one AP2 domain.	Integrase-type DNA-binding superfamily protein	DREB2D, ERF049
AP2-EREBP	AT3G16770	Encodes a member of the ERF (ethylene response factor) subfamily B-2 of the plant specific ERF/AP2 transcription factor family (RAP2.3). The protein contains one AP2 domain. Overexpression of this gene in tobacco BY-2 cells confers resistance to H2O2 and heat stresses. Overexpression in Arabidopsis causes upregulation of PDF1.2 and GST6. It is part of the ethylene signalling pathway and is predicted to act downstream of EIN2 and CTR1, but not under EIN3.	Ethylene-responsive element binding protein	ATEBP, EBP, ERF72, RAP2.3
AP2-EREBP	AT5G53290	Encodes a member of the ERF (ethylene response factor) subfamily B-5 of ERF/AP2 transcription factor family. The protein contains one AP2 domain. CRF proteins relocalize to the nucleus in response to cytokinin.	Cytokinin response factor 3	CRF3
AP2-EREBP	AT5G65510	Encodes one of three PLETHORA transcription factors required to maintain high levels of PIN1 expression at the periphery of the meristem and modulate local auxin production in the central region of the SAM which underlies phyllotactic transitions.	AINTEGUMENTA-like 7	AIL7, PLT7
C2C2-Gata	AT3G51080	Encodes a member of the GATA factor family of zinc finger transcription factors.	GATA transcription factor 6	GATA6
C2H2	AT2G41940	Encodes a zinc finger protein containing only a single zinc finger. Involved in GA and cytokinin signal integration.	Zinc finger protein 8	ZFP8
C2H2	AT3G07940	Calcium-dependent ARF-type GTPase activating protein family	Calcium-dependent ARF-type GTPase activating protein family	AtGAP
C2H2	AT5G07500	Encodes an embryo-specific zinc finger transcription factor required for heart-stage embryo formation.	Zinc finger C-x8-C-x5-C-x3-H type family protein	AtTZF6, PEI1, TZF6
C2H2	AT5G43540	Encodes a protein containing a zinc finger, C2H2-type domain.	C2H2 and C2HC zinc fingers superfamily protein	ZF-C2H2-type

C3H	AT2G04240	Encodes a small protein with an N-terminal trans-membrane domain and a RING-H2 zinc finger motif located at the C-terminus. Gene expression is induced by salt and osmotic stress. Transcript levels are induced by DELLA proteins and repressed by gibberellic acid. Involved in ABA metabolism.	RING/U-box superfamily protein	XERICO
CCAAT-HAP2	AT3G14020	Encodes a nuclear factor Y A (NF-YA), a highly conserved transcription factor presented in all eukaryotic organisms	Nuclear factor Y, subunit A6	NF-YA6
CCAAT-HAP5	AT5G27910	Encodes a nuclear factor Y C (NF-YC), a highly conserved transcription factor presented in all eukaryotic organisms	Nuclear factor Y, subunit C8	NF-YC8
G2-like	AT2G20570	Encodes GLK1, Golden2-like 1, one of a pair of partially redundant nuclear transcription factors that regulate chloroplast development in a cell-autonomous manner. GLK2, Golden2-like 2, is encoded by At5g44190. GLK1 and GLK2 regulate the expression of the photosynthetic apparatus.	GBF's pro-rich	ATGLK1, GLK1, GPRI1
G2-like	AT4G04580	Encodes a protein containing Myb, DNA-binding domain	Homeodomain-like superfamily protein	MYB-TF
G2-like	AT5G16560	Encodes a KANADI protein (KAN) that regulates organ polarity in Arabidopsis. KAN encodes a nuclear-localised protein in the GARP family of putative transcription factors. Together with KAN2, this gene appears to be involved in the development of the carpel and the outer integument of the ovule. Along with KAN2 and KAN4, KAN1 appears to be required for proper regulation of PIN1 in early embryogenesis.	Homeodomain-like superfamily protein	KAN, KAN1
Homeobox	AT1G52150	Member of the class III HD-ZIP protein family. Contains homeodomain and leucine zipper domain. Critical for vascular development and negatively regulates vascular cell differentiation.	Homeobox-leucine zipper family protein / lipid-binding START domain-containing protein	ATHB-15, ATHB15, CNA, ICU4
MADS	AT5G51870	Encodes a MADS-box transcription factor involved in floral transition.	AGAMOUS-like 71	AGL71
MADS	AT5G65070	Encodes MADS-box containing FLC paralog. Five splice variants have been identified but not characterised with respect to expression patterns and/or differing function. Overexpression of the gene in the Landsberg ecotype leads to a delay in flowering, transcript levels of MAF4 are reduced after a 6 week vernalization.	K-box region and MADS-box transcription factor family protein	AGL69, FCL4, MAF4

NAC	AT1G52890	Encodes a NAC transcription factor whose expression is induced by drought, high salt, and abscisic acid. This gene binds to ERD1 promoter in vitro.	NAC domain containing protein 19	ANAC019, ANAC19, NAC019
NAC	AT1G69490	Encodes a member of the NAC transcription factor gene family. It is expressed in floral primordia and upregulated by AP3 and PI. Its expression is associated with leaf senescence. The mRNA is cell-to-cell mobile.	NAC-like, activated by AP3/PI	ANAC029, ATNAP, NAP
NAC	AT2G27300	NTL8 is a membrane-associated NAC transcription factor that binds both TRY and TCL1. Overexpression results in fewer trichomes.	NTM1-like 8	ANAC040, NTL8
NLP	AT3G59580	NIN-LIKE PROTEIN 9	Plant regulator RWP-RK family protein	NLP9

TABLE 4. List of 23 transcription factors selected by statistical analysis based on polysomal-associated mRNA differentially accumulated in germinating seeds produced beforehand under different temperature regimes.

APPENDIX A. APPENDIX

A.1. Computation of the first and second derivatives of W_t defined in (6).

A.1.1. *Computation of the first derivatives of W_t.* By the definition of W_t given in (6), we have

$$\frac{\partial W_{i,j,t}}{\partial \boldsymbol{\delta}} = \frac{\partial \eta_{i,t}}{\partial \boldsymbol{\delta}} + \frac{\partial Z_{i,j,t}}{\partial \boldsymbol{\delta}}.$$

For all $i_0 \in \{1, \dots, I\}$ and $t_0 \in \{1, \dots, T\}$ we have

$$\begin{aligned}
\frac{\partial W_{i,j,t}}{\partial \eta_{i_0,t_0}} &= \frac{\partial}{\partial \eta_{i_0,t_0}}\Big(\eta_{i,t} + Z_{i,j,t}\Big) = \frac{\partial \eta_{i,t}}{\partial \eta_{i_0,t_0}} + \sum_{k=1}^{q\wedge(t-1)} \gamma_k \frac{\partial E_{i,j,t-k}}{\partial \eta_{i_0,t_0}} \\
&= \frac{\partial \eta_{i,t}}{\partial \eta_{i_0,t_0}} - \sum_{k=1}^{q\wedge(t-1)} \gamma_k Y_{i,j,t-k} \frac{\partial W_{i,j,t-k}}{\partial \eta_{i_0,t_0}} \exp(-\eta_{i,t-k} - Z_{i,j,t-k}) \\
&= \frac{\partial \eta_{i,t}}{\partial \eta_{i_0,t_0}} - \sum_{j=k}^{q\wedge(t-1)} \gamma_k (1 + E_{i,j,t-k}) \frac{\partial W_{i,j,t-k}}{\partial \eta_{i_0,t_0}},
\end{aligned}$$

where $E_{i,j,t} = 0$ for any $t \leq 0$.

For all $q_0 \in \{1, \dots, q\}$

$$\begin{aligned}\frac{\partial W_{i,j,t}}{\partial \gamma_{q_0}} &= \frac{\partial}{\partial \gamma_{q_0}}\Big(\eta_{i,t} + Z_{i,j,t}\Big) = \frac{\partial \eta_{i,t}}{\partial \gamma_{q_0}} + \frac{\partial}{\partial \gamma_{q_0}} \sum_{k=1}^{q} \gamma_k E^{(i)}_{j,t-k} = E^{(i)}_{j,t-q_0} + \sum_{k=1}^{q\wedge(t-1)} \gamma_k \frac{\partial E_{i,j,t-k}}{\partial \gamma_{q_0}} \\ &= E_{i,j,t-q_0} - \sum_{k=1}^{q\wedge(t-1)} \gamma_k Y_{i,j,t-k} \frac{\partial W_{i,j,t-k}}{\partial \gamma_{q_0}} \exp(-W_{i,j,t-k}) \\ &= E_{i,j,t-q_0} - \sum_{k=1}^{q\wedge(t-1)} \gamma_k (1 + E_{i,j,t-k}) \frac{\partial W_{i,j,t-k}}{\partial \gamma_{q_0}},\end{aligned}$$

where we used the fact that $E_{i,j,t-q_0} = 0$ for any $t \leq 0$.

We obtain the first derivatives of $W_{i,j,t}$ from the following recursive expressions. For all $i_0 \in \{1, \dots, I\}$ and $t_0 \in \{1, \dots, T\}$

$$\begin{aligned}\frac{\partial W_{i,j,1}}{\partial \eta_{i_0,t_0}} &= \frac{\partial \eta_{i,1}}{\partial \eta_{i_0,t_0}} = \begin{cases} 1, & \text{if } i = i_0 \quad \text{and } t_0 = 1 \\ 0, & \text{otherwise} \end{cases}, \\ \frac{\partial W_{i,j,2}}{\partial \eta_{i_0,t_0}} &= \frac{\partial \eta_{i,2}}{\partial \eta_{i_0,t_0}} - \gamma_1 (1 + E_{i,j,1}) \frac{\partial W_{i,j,1}}{\partial \eta_{i_0,t_0}} \\ &= \begin{cases} 1, & \text{if } i = i_0 \quad \text{and } t_0 = 2 \\ -\gamma_1(1 + E_{i,j,1}), & \text{if } i = i_0 \quad \text{and } t_1 = 1 \\ 0, & \text{otherwise} \end{cases},\end{aligned}$$

In the same way, for all $q_0 \in \{1, \dots, q\}$ we have

$$\begin{aligned}\frac{\partial W_{i,j,1}}{\partial \gamma_{q_0}} &= 0, \\ \frac{\partial W_{i,j,2}}{\partial \gamma_{q_0}} &= E_{i,j,2-q_0}, \\ \frac{\partial W_{i,j,3}}{\partial \gamma_{q_0}} &= E_{i,j,3-q_0} - \gamma_1 (1 + E_{i,j,2}) \frac{\partial W_{i,j,2}}{\partial \gamma_{q_0}},\end{aligned}$$

and so on. Note that

$$\begin{aligned}W_{i,j,1} &= \eta_{i,1} + Z_{i,j,1} = \eta_{i,1} + \sum_{k=1}^{q} \gamma_k E_{i,j,1-k} = \eta_{i,1}, \\ E_{i,j,1} &= Y_{i,j,1} \exp(-W_{i,j,1}) - 1 = Y_{i,j,1} \exp(-\eta_{i,1}) - 1, \\ W_{i,j,2} &= \eta_{i,2} + Z_{i,j,2} = \eta_{i,2} + \sum_{k=1}^{q} \gamma_k E_{i,j,2-k} = \eta_{i,2} + \gamma_1 E_{i,j,1}, \\ E_{i,j,2} &= Y_{i,j,2} \exp(-W_{i,j,2}) - 1 = Y_{i,j,2} \exp(-\eta_{i,2} - \gamma_1 E_{i,j,1}) - 1.\end{aligned}$$

A.1.2. *Computation of the second derivatives of W_t.* For all $i_0, i_1 \in \{0, \dots, I\}$ and $t_0, t_1 \in \{1, \dots, T\}$

$$\begin{aligned}\frac{\partial^2 W_{i,j,t}}{\partial \eta_{i_0,t_0}\partial \eta_{i_1,t_1}} &= \frac{\partial}{\partial \eta_{i_1,t_1}}\left\{\frac{\partial \eta_{i,t}}{\partial \eta_{i_0,t_0}} - \sum_{k=1}^{q\wedge(t-1)} \gamma_k(1+E_{i,j,t-k})\frac{\partial W_{i,j,t-k}}{\partial \eta_{i_0,t_0}}\right\}\\ &= -\sum_{k=1}^{q\wedge(t-1)} \gamma_k \frac{\partial E_{i,j,t-k}}{\partial \eta_{i_1,t_1}}\frac{\partial W_{i,j,t-k}}{\partial \eta_{i_0,t_0}} - \sum_{k=1}^{q\wedge(t-1)} \gamma_k(1+E_{i,j,t-k})\frac{\partial^2 W_{i,j,t-k}}{\partial \eta_{i_0,t_0}\partial \eta_{i_1,t_1}}\\ &= \sum_{k=1}^{q\wedge(t-1)} \gamma_k(1+E_{i,j,t-k})\frac{\partial W_{i,j,t-k}}{\partial \eta_{i_0,t_0}}\frac{\partial W_{i,j,t-k}}{\partial \eta_{i_1,t_1}} - \sum_{k=1}^{q\wedge(t-1)} \gamma_k(1+E_{i,j,t-k})\frac{\partial^2 W_{i,j,t-k}}{\partial \eta_{i_0,t_0}\partial \eta_{i_1,t_1}}.\end{aligned}$$

For all $q_0, q_1 \in \{1, \dots, q\}$

$$\begin{aligned}\frac{\partial^2 W_{i,j,t}}{\partial \gamma_{q_0}\partial \gamma_{q_1}} &= \frac{\partial E_{i,j,t-q_0}}{\partial \gamma_{q_1}} - (1+E_{i,j,t-q_1})\frac{\partial W_{i,j,t-q_1}}{\partial \gamma_{q_0}} - \sum_{k=1}^{q\wedge(t-1)} \gamma_k\Bigg\{\frac{\partial W_{i,j,t-k}}{\partial \gamma_{q_0}}\frac{\partial E_{i,j,t-k}}{\partial \gamma_{q_1}}\\ &\quad + (1+E_{i,j,t-k})\frac{\partial^2 W_{i,j,t-k}}{\partial \gamma_{q_0}\partial \gamma_{q_1}}\Bigg\} = -(1+E_{i,j,t-q_0})\frac{\partial W_{j,t-q_0}}{\partial \gamma_{q_1}} - (1+E_{i,j,t-q_1})\frac{\partial W_{i,j,t-q_1}}{\partial \gamma_{q_0}}\\ &\quad - \sum_{k=1}^{q\wedge(t-1)} \gamma_k\Bigg\{-(1+E_{i,j,t-k})\frac{\partial W_{i,j,t-k}}{\partial \gamma_{q_0}}\frac{\partial W_{i,j,t-k}}{\partial \gamma_{q_1}} + (1+E_{i,j,t-k})\frac{\partial^2 W_{i,j,t-k}}{\partial \gamma_{q_0}\partial \gamma_{q_1}}\Bigg\}.\end{aligned}$$

To obtain the second derivatives of W_t we use the following recursive expressions for all $i_0, i_1 \in \{0, \dots, I\}$ and $t_0, t_1 \in \{1, \dots, T\}$

$$\begin{aligned}&\frac{\partial^2 W_{i,j,1}}{\partial \eta_{i_0,t_0}\partial \eta_{i_1,t_1}} = 0,\\ &\frac{\partial^2 W_{i,j,2}}{\partial \eta_{i_0,t_0}\partial \eta_{i_1,t_1}} = \begin{cases}\gamma_1(1+E_{i,j,1}), & \text{if } i = i_0 = i_1 \quad \text{and } t_0 = t_1 = 1\\ 0, & \text{otherwise}\end{cases}.\end{aligned}$$

We also have that for all $q_0, q_1 \in \{1, \dots, q\}$

$$\begin{aligned}\frac{\partial^2 W_{i,j,1}}{\partial \gamma_{q_0}\partial \gamma_{q_1}} &= 0,\\ \frac{\partial^2 W_{i,j,2}}{\partial \gamma_{q_0}\partial \gamma_{q_1}} &= 0,\end{aligned}$$

and so on.

A.2. **Additional numerical experiments.**

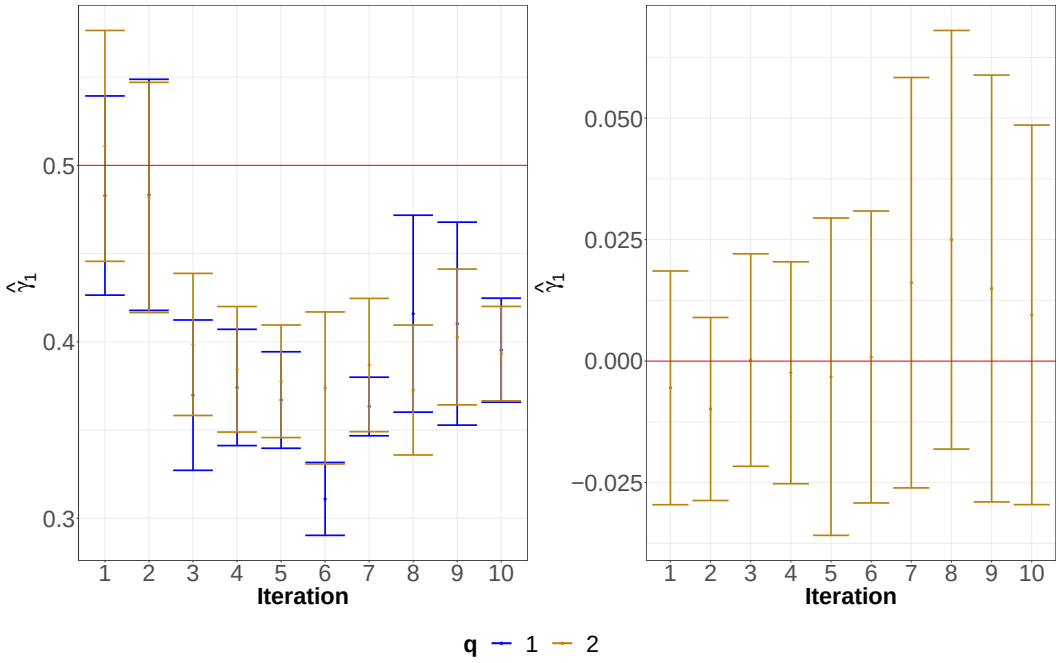

FIGURE 12. Error bars for the estimations of $\boldsymbol{\gamma}^\star$ in Model (2) for $I = 3$, $T = 50$, $J = 10$, $q^\star = 1$, $\gamma^\star = 0.5$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations obtained by $q = 1$. The horizontal lines correspond to the values of the $\gamma_i^\star$'s.

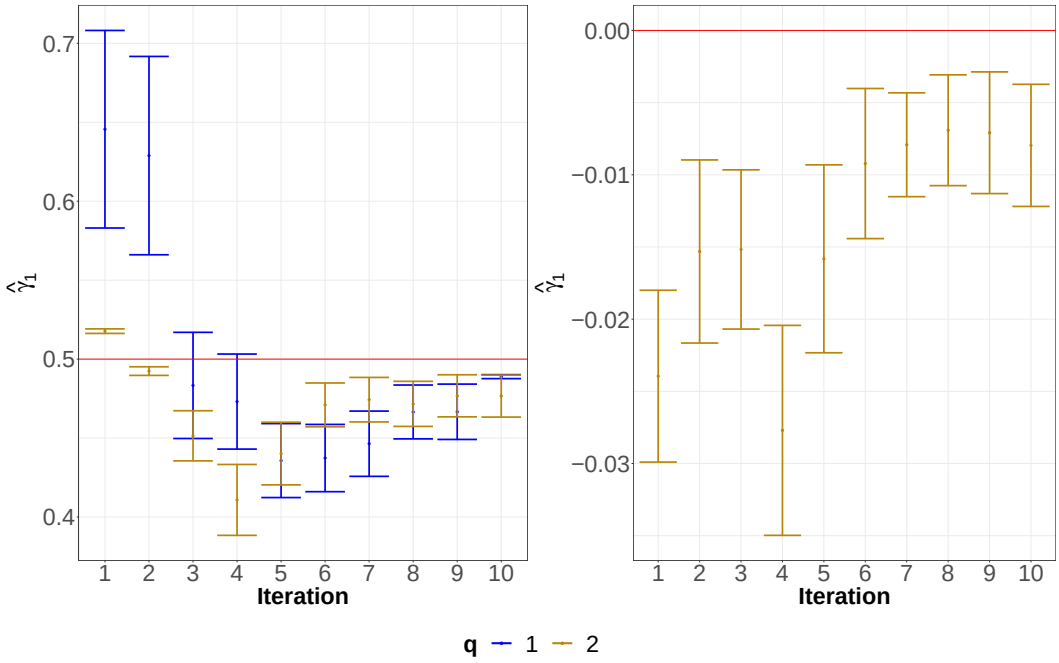

FIGURE 13. Error bars for the estimations of $\boldsymbol{\gamma}^\star$ in Model (2) for $I = 3$, $T = 50$, $J = 100$, $q^\star = 1$, $\gamma^\star = 0.5$, 10 non-null coefficients in $\boldsymbol{\eta}^\star$, and 50 simulations obtained by $q = 1$. The horizontal lines correspond to the values of the $\gamma_i^\star$'s.

Gene \ Temp.pop	Low.npoly	Medium.npoly	Elevated.npoly	Low.poly	Medium.poly	Elevated.poly
AT1G28230	240	52.33	21.33	268.67	56	25.33
AT1G18100	696	162.33	98.67	1385.33	249	219
AT1G48130	6151.67	1179.33	841	6170.67	1511.33	1162.33
AT1G47980	2947	346.67	176.33	2114.33	418.33	198
AT1G14940	2277.67	689.67	598	3661	776.67	832
AT1G14950	4630.67	1180.33	476	6617	1465	883
AT1G07090	320.33	189	81.67	557.33	286.33	174.33
AT1G70810	499	221	137.33	1161.33	410.67	305
AT1G73190	2860.33	502	236.33	3090	608.33	335.33
AT2G33830	2498.33	379.33	305	7977.33	1307	1110
AT2G26010	50.67	7.67	1.33	150.67	17.33	3.67
AT2G43510	1350.33	305.33	162.33	3943.67	887.67	514
AT2G29300	248.33	70.33	32.33	253.67	62.33	34
AT2G43580	664	91	31.33	694.33	88	37.33
AT2G23640	285.67	140	122.33	830	262	328
AT3G56350	2105.67	442	223.67	2785.67	542	345
AT4G24715	1097.33	263.33	116.67	1249.67	318.67	202
AT4G02520	3852	866.67	621	2575.33	656.33	450.67
AT4G00220	180.33	98.33	66	367.33	122.33	108.33
AT4G26740	2026	442.67	266	2606.67	526	393.67
AT5G55560	297.67	138	72	603	266.67	158.67

TABLE 5. Data used for displaying Figure 10 at the different temperature (Low, Medium, ELevated) where ".npoly" (resp. ".poly") are the values corresponding to the non-polysomal (resp. polysomal) population.

A.3. **Additional results for the application section.**

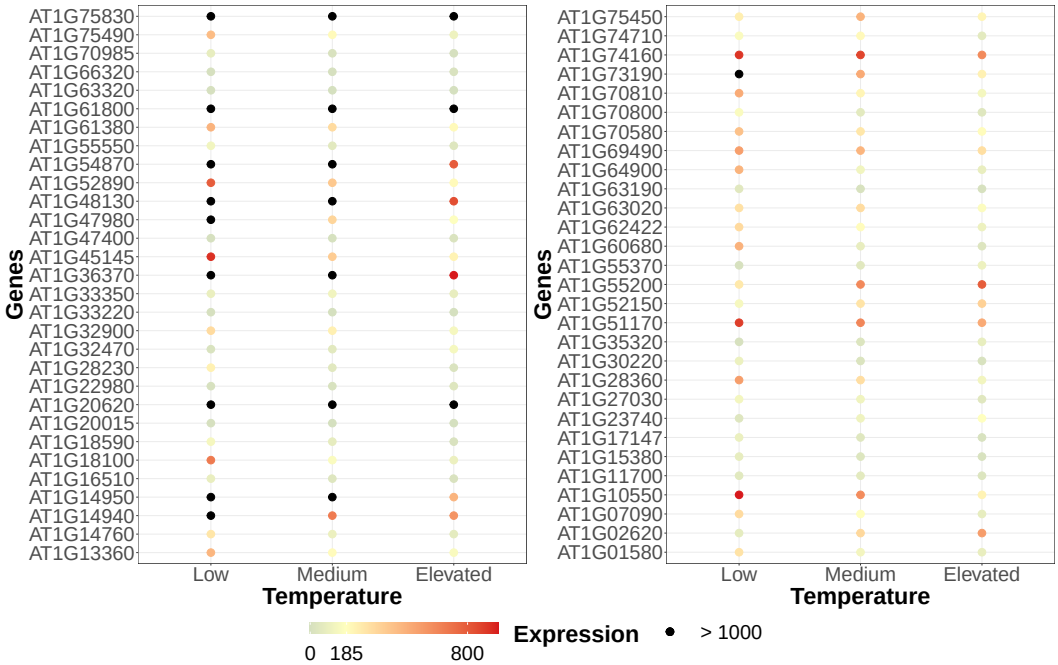

FIGURE 14. Average gene expression values of 3 replications at each temperature condition (Low, Medium, Elevated) for the chromosome 1 in non-polysomal population. The genes displayed in this figure are obtained by our selection procedure.

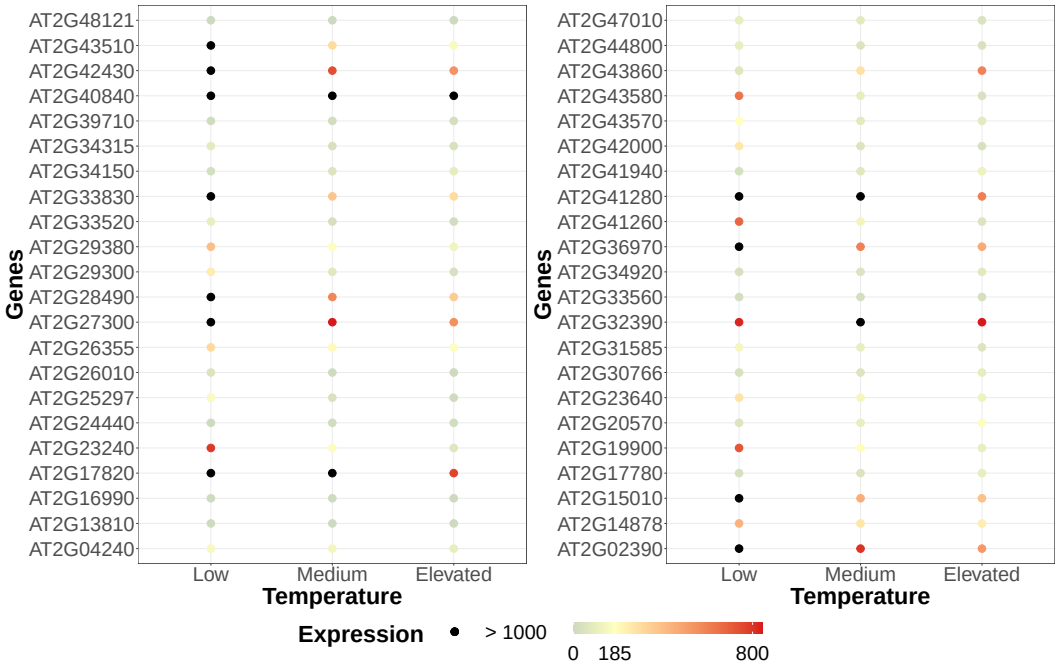

FIGURE 15. Average gene expression values of 3 replications at each temperature condition (Low, Medium, Elevated) for the chromosome 2 in non-polysomal population. The genes displayed in this figure are obtained by our selection procedure.

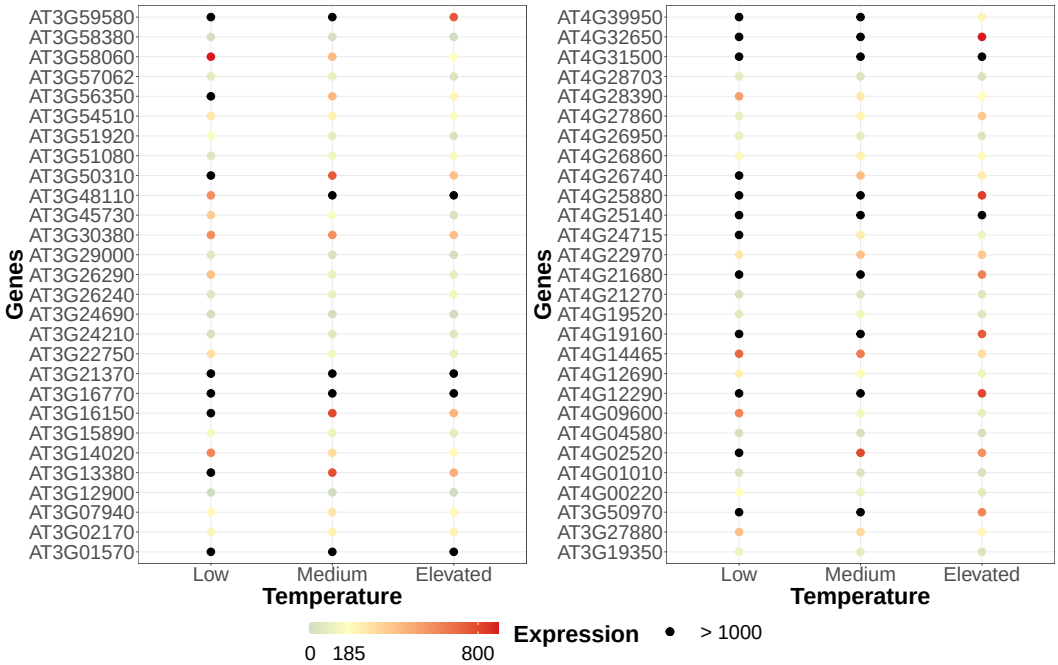

FIGURE 16. Average gene expression values of 3 replications at each temperature condition (Low, Medium, Elevated) for the chromosomes 3 and 4 in non-polysomal population. The genes displayed in this figure are obtained by our selection procedure.

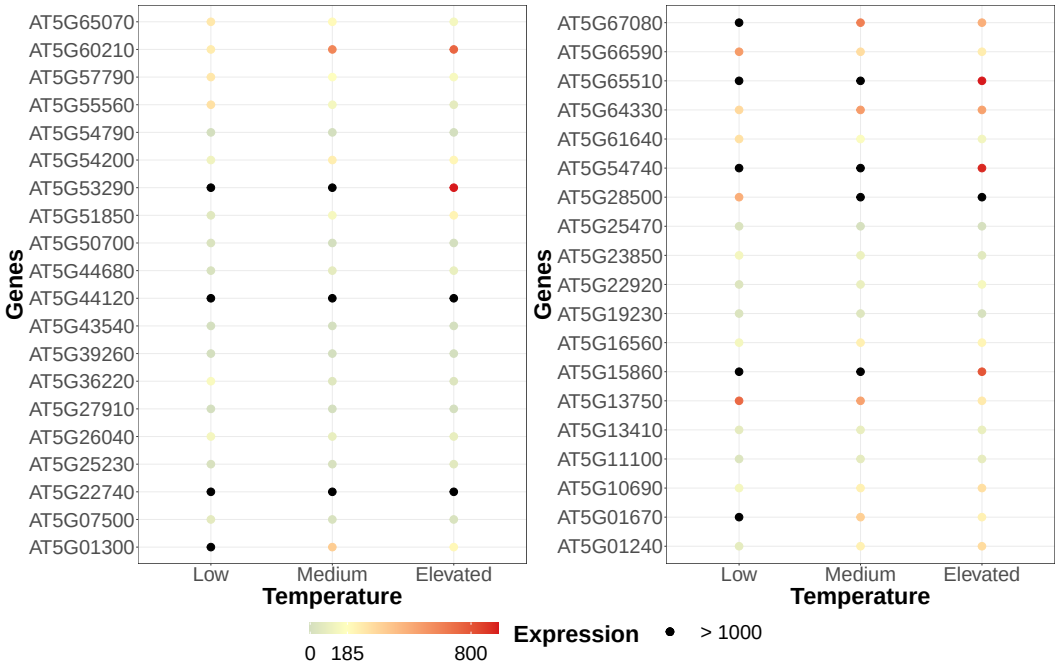

FIGURE 17. Average gene expression values of 3 replications at each temperature condition (Low, Medium, Elevated) for the chromosome 5 in non-polysomal population. The genes displayed in this figure are obtained by our selection procedure.

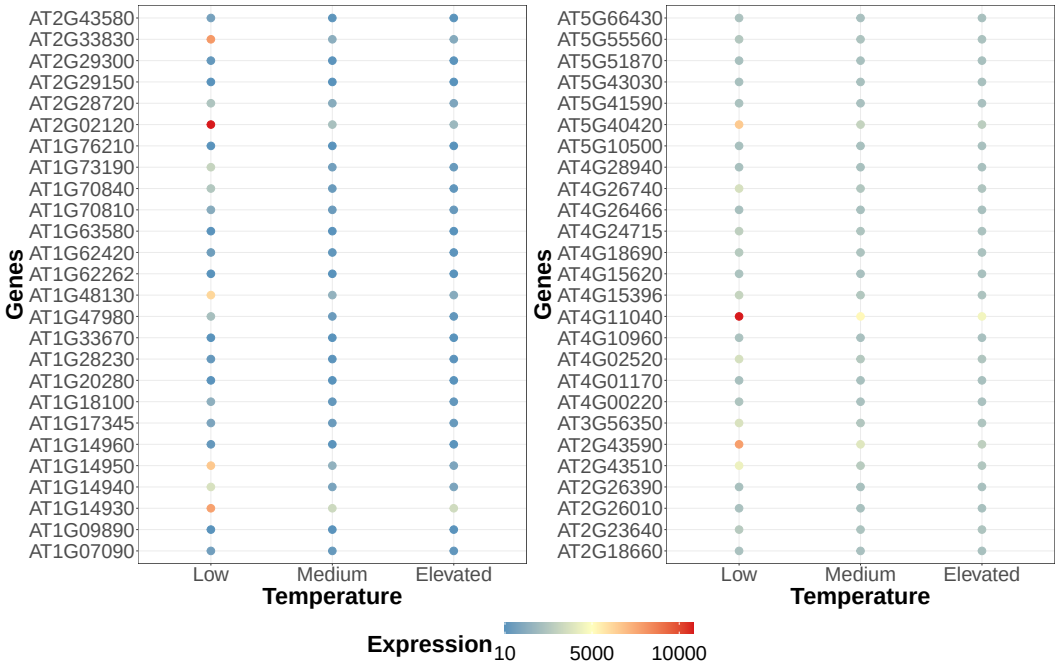

FIGURE 18. Average gene expression values of 3 replications at each temperature condition (Low, Medium, Elevated) for the polysomal population. The genes displayed in this figure are obtained by our selection procedure.

References

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

[2] Al-Osh, M. A. and A. A. Alzaid (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis 8*(3), 261–275.

[3] Alzaid, A. A. and M. Al-Osh (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability 27*(2), 314–324.

[4] Bai, B., A. Peviani, S. van der Horst, M. Gamm, B. Snel, L. Bentsink, and J. Hanson (2017). Extensive translational regulation during seed germination revealed by polysomal profiling. *New Phytologist 214*(1), 233–244.

[5] Bailly, C. (2019, 10). The signalling role of ROS in the regulation of seed germination and dormancy. *Biochemical Journal 476*, 3019–3032.

[6] Basbouss-Serhal, I., L. Soubigou-Taconnat, C. Bailly, and J. Leymarie (2015). Germination potential of dormant and nondormant arabidopsis seeds is driven by distinct recruitment of messenger rnas to polysomes. *Plant Physiology 168*(3), 1049–1065.

[7] Bentsink, L., J. Jowett, C. J. Hanhart, and M. Koornneef (2006). Cloning of DOG1, a quantitative trait locus controlling seed dormancy in arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America 103*(45), 17042–17047.

[8] Bien, K., I. Nolte, and W. Pohlmeier (2011, November). An inflated multivariate integer count hurdle model: an application to bid and ask quote dynamics. *Journal of Applied Econometrics 26*(4), 669–707.

[9] Cox, D. R., G. Gudmundsson, G. Lindgren, L. Bondesson, E. Harsaae, P. Laake, K. Juselius, and S. L. Lauritzen (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics 8*(2), 93–115.

[10] Davis, R., K. Fokianos, S. Holan, H. Joe, J. Livsey, R. Lund, V. Pipiras, and N. Ravishanker (2021, 03). Count time series: A methodological review. *Journal of the American Statistical Association 116*, 1–50.

[11] Davis, R. A., W. T. M. Dunsmuir, and S. B. Streett (2003). Observation-driven models for Poisson counts. *Biometrika 90*(4), 777–790.

[12] Davis, R. A., W. T. M. Dunsmuir, and S. B. Streett (2005). Maximum likelihood estimation for an observation driven model for Poisson counts. *Methodology and Computing in Applied Probability 7*(2), 149–159.

[13] Davis, R. A., W. T. M. Dunsmuir, and Y. Wang (1999). Modeling time series of count data. *Statistics Textbooks and Monographs 158*, 63–114.

[14] Dunsmuir, W. T. M. (2015). *Generalized Linear Autoregressive Moving Average Models*, Chapter 3, pp. 51–76. CRC Press.

[15] Fokianos, K. (2021). Multivariate count time series modelling. *Econometrics and Statistics*.

[16] Fokianos, K., A. Rahbek, and D. Tjøstheim (2009). Poisson autoregression. *Journal of the American Statistical Association 104*(488), 1430–1439.

[17] Fokianos, K., B. Støve, D. Tjøstheim, and P. Doukhan (2020). Multivariate count autoregression. *Bernoulli 26*, 471–499.

[18] Fokianos, K. and D. Tjøstheim (2012). Nonlinear poisson autoregression. *Annals of the Institute of Statistical Mathematics 64*(6), 1205–1225.

[19] Fokianos, K. and D. Tjøstheim (2011). Log-linear poisson autoregression. *Journal of Multivariate Analysis 102*(3), 563 – 578.

[20] Franke, J. and T. S. Rao. (1995). Multivariate first-order integer values autoregressions. Technical report, Departement of Mathemtics, UMIST.

[21] Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

[22] Galland, M., R. Huguet, E. Arc, G. Cueff, D. Job, and L. Rajjou (2014). Dynamic proteomics emphasizes the emportance of selective mRNA translation and protein turnover during arabidopsis seed germination. *Molecular & Cellular Proteomics 13*(1), 252–268.

[23] Giese, G., D. P. S. Zepeda, J. Beacham, and C. V. Cruz (2021). Modeling nematode population dynamics using a multivariate poisson model with spike and slab variable selection. *Journal of Applied Statistics*, 1–20.

[24] Gomtsyan, M., C. Lévy-Leduc, S. Ouadah, L. Sansonnet, and T. Blein (2022). Variable selection in sparse glarma models. *Statistics 56*(4), 755–784.

[25] Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[26] Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

[27] Heinen, A. and E. Rengifo (2007). Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance 14*(4), 564–583.

[28] Held, L. and M. Hofmann (2005). A statistical framework for the analysis of multivariate infectious disease surveillance data. *Statistical Modelling 5*, 187–199.

[29] Iwasaki, M., L. Hyvärinen, U. Piskurewicz, and L. Lopez-Molina (2019). Non-canonical RNA-directed DNA methylation participates in maternal and environmental control of seed dormancy. *eLife 8*, e37434.

[30] Iwasaki, M., S. Penfield, and L. Lopez-Molina (2022). Parental and environmental control of seed dormancy in arabidopsis thaliana. *Annual Review of Plant Biology 73*(1), 355–378. PMID: 35138879.

[31] Jung, R. C. and R. Liesenfeld (2001). Estimating time series models for count data using efficient importance sampling. *AStA Advances in Statistical Analysis 4*(85), 387–407.

[32] Jung, R. C., R. Liesenfeld, and J.-F. Richard (2011). Dynamic factor models for multivariate count data: An application to stock-market trading activity. *Journal of Business & Economic Statistics 29*(1), 73–85.

[33] Jørgensen, B., S. Lundbye-Christensen, P. X.-K. Song, and L. Sun (1996). State-space models for multivariate longitudinal data of mixed types. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 24*(3), 385–402.

[34] Ko, J.-H., S. H. Yang, and K.-H. Han (2006). Upregulation of an arabidopsis RING-H2 gene, XERICO, confers drought tolerance through increased abscisic acid biosynthesis. *The Plant Journal 47*(3), 343–355.

[35] Latour, A. (1997). The multivariate GINAR(p) process. *Advances in Applied Probability 29*(1), 228–248.

[36] Lee, K. H., B. A. Coull, A.-B. Moscicki, B. J. Paster, and J. R. Starr (2018, December). Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data. *Biostatistics 21*(3), 499–517.

[37] McKenzie, E. (1985). Some simple models for discrete variate time series. *Journal of the American Water Resources Association 21*(4), 645–650.

[38] Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

[39] Nishimura, N., W. Tsuchiya, J. J. Moresco, Y. Hayashi, K. Satoh, N. Kaiwa, T. Irisa, T. Kinoshita, J. I. Schroeder, J. R. Yates, T. Hirayama, and T. Yamazaki (2018). Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme. *Nature Communications 9*(1).

[40] Pedeli, X. and D. Karlis (2013a, 03). On composite likelihood estimation of a multivariate INAR(1) model. *Journal of Time Series Analysis 34*, 206–220.

[41] Pedeli, X. and D. Karlis (2013b, 11). Some properties of multivariate INAR(1) processes. *Computational Statistics and Data Analysis 67*, 213–225.

[42] Penfield, S. and D. R. MacGregor (2017). Effects of environmental variation during seed production on seed dormancy and germination. *Journal of Experimental Botany 68*(4), 819–825.

[43] Ravishanker, N., V. Serhiyenko, and M. R. Willig (2014). Hierarchical dynamic models for multivariate times series of counts. *Statistics and Its Interface 7*, 559–570.

[44] Reed, R. C., K. J. Bradford, and I. Khanday (2022). Seed germination and vigor: ensuring crop sustainability in a changing climate. *Heredity 128*(6), 450–459.

[45] Sano, N., L. Rajjou, and H. M. North (2020). Lost in translation: Physiological roles of stored mRNAs in seed germination. *Plants 9*(3).

[46] Shu, K., X. dong Liu, Q. Xie, and Z. hua He (2016). Two faces of one seed: hormonal regulation of dormancy and germination. *Molecular Plant 9*(1), 34–45.

[47] Stark, R., M. Grzelak, and J. Hadfield (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics 20*(11), 631–656.

[48] Xi, W., C. Liu, X. Hou, and H. Yu (2010). MOTHER OF FT AND TFL1 regulates seed germination through a negative feedback loop modulating ABA signaling in arabidopsis. *The Plant Cell 22*(6), 1733–1748.

[49] Xiang, Y., K. Nakabayashi, J. Ding, F. He, L. Bentsink, and W. J. Soppe (2014). Reduced dormancy5 encodes a protein phosphatase 2c that is required for seed dormancy in arabidopsis. *The Plant Cell 26*(11), 4362–4375.

[50] Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics 44*(4), 1019–1031.

Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120 Palaiseau, France

UMR7622 CNRS-Sorbonne Université; Laboratoire de Biologie du Développement, Biologie des Semences, Institut de Biologie Paris-Seine, Paris, France. Boite 24, 4 place Jussieu, Paris 75005, France

Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay, 78026, Versailles, France