

Variable selection in a specific regression for count time series

Marina Gomtsyan



StatMathAppli
4 September, 2025

Context and Motivation

Count time series

- Record of the number of occurrences of events over time
- Nonnegative and integer-valued
- Examples: daily records of COVID cases, number of crimes, transactions in stocks, and RNA-Seq time series
- **Statistical model:** Generalised Linear Autoregressive Moving Average (GLARMA)

The goal: Propose efficient variable selection approach in sparse GLARMA models for overdispersed data, *i.e.* the variance is much larger than the mean

Negative Binomial GLARMA Model

$$Y_t | \mathcal{F}_{t-1} \sim \text{NB}(\mu_t^*, \alpha^*), \text{ with } \mathcal{F}_{t-1} = \sigma(Y_s, s \leq t-1),$$

with

$$W_t^* := \log(\mu_t^*) = \beta_0^* + \sum_{i=1}^p \beta_i^* x_{t,i} + \sum_{j=1}^q \gamma_j^* E_{t-j}^*,$$

where $1 \leq q \leq \infty$ and $1 \leq t \leq n$

- $x_{t,1}, \dots, x_{t,p}$ are the covariates at time t
- $\beta^* = (\beta_0^*, \beta_1^*, \dots, \beta_p^*)^T$ the vector of regressor coefficients
- $\gamma^* = (\gamma_1^*, \dots, \gamma_q^*)^T$ such that $\sum_{k \geq 1} |\gamma_k^*| < \infty$
- $E_t^* = \frac{Y_t - \mu_t^*}{\mu_t^* + \mu_t^{*2}/\alpha^*}$, with $E_t^* = 0$ for all $t \leq 0$

Classical Estimation of the Parameters in GLARMA Models (*Davis, Dunsmuir & Streett, 2005*)

Classical approach: estimation of

$\boldsymbol{\delta}^* = (\beta_0^*, \beta_1^*, \dots, \beta_p^*, \gamma_1^*, \dots, \gamma_q^*)^T$ with conditional maximum likelihood

$$\hat{\boldsymbol{\delta}} = \arg \max_{\boldsymbol{\delta}} L(\boldsymbol{\delta}, \alpha),$$

where

$$L(\boldsymbol{\delta}, \alpha) = \sum_{t=1}^n \left(\log \Gamma(\alpha + Y_t) - \log \Gamma(Y_t + 1) - \log \Gamma(\alpha) \right. \\ \left. + \alpha \log \alpha + Y_t W_t(\boldsymbol{\delta}, \alpha) - (\alpha + Y_t) \log(\alpha + \exp(W_t(\boldsymbol{\delta}, \alpha))) \right)$$

- In the sparse framework, with many components of $\boldsymbol{\beta}^*$ being null, this procedure provides poor estimation results

Our Estimation Procedure in Sparse GLARMA Models

Estimation of γ^* :

$$\hat{\gamma} = \arg \max_{\gamma} L(\beta^{(0)T}, \gamma^T, \alpha^{(0)}),$$

for a given initial value $\beta^{(0)} = (\beta_0^{(0)}, \dots, \beta_p^{(0)})^T$ and $\alpha^{(0)}$

To obtain $\hat{\gamma}$, use Newton-Raphson algorithm with initial value $\gamma^{(0)} = (\gamma_0^{(0)}, \dots, \gamma_q^{(0)})^T$ for $r \geq 1$

$$\gamma^{(r)} = \gamma^{(r-1)} - \frac{\partial^2 L}{\partial \gamma^T \partial \gamma}(\beta^{(0)T}, \gamma^{(r-1)}, \alpha^{(0)})^{-1} \frac{\partial L}{\partial \gamma}(\beta^{(0)T}, \gamma^{(r-1)}, \alpha^{(0)})$$

Variable Selection Step

Variable selection: To obtain a sparse estimator of β^* , we propose using

$$\hat{\beta}(\lambda) = \arg \min_{\beta} \left\{ -\tilde{L}_Q(\beta) + \lambda \|\beta\|_1 \right\},$$

where $\lambda > 0$

$\tilde{L}_Q$ is the quadratic approximation of L obtained by second order Taylor approximation:

$$-\tilde{L}(Q) = \frac{1}{2} \|\mathcal{Y} - \mathcal{X}\beta\|_2^2,$$

where

$$\mathcal{Y} = \Lambda^{1/2} U^T \beta^{(0)} + \Lambda^{-1/2} U^T \left(\frac{\partial L}{\partial \beta}(\beta^{(0)}, \hat{\gamma}, \alpha^{(0)}) \right)^T, \quad \mathcal{X} = \Lambda^{1/2} U^T$$

- $U\Lambda U'$ is the singular value decomposition of the positive semidefinite symmetric matrix $-\frac{\partial^2 L}{\partial \beta \partial \beta'}(\beta^{(0)}, \hat{\gamma}, \alpha^{(0)})$

The Choice of λ : Stability Selection

(Meinshausen & Bühlmann, 2010)

- Identifies a set of “stable” variables that are selected with high probability
- The vector $\mathcal{Y}$ is randomly split into several subsamples of size $(p + 1)/2$
- For each subsampling, we apply the LASSO criterion for a given λ and store the indices i of the non null $\hat{\beta}_i$
- For a given threshold, we keep in the final set of selected variables only the ones appearing a number of times larger than this threshold
- Concerning the choice of λ , we consider the smallest element of the grid of λ provided by the R `glmnet` package
- Need to find a threshold

Practical Implementation

1. **Initialisation.** As the estimator of β^* , take $\beta^{(0)}$, which is obtained by fitting a GLM to the observations $Y_1, \dots, Y_n$. For $\alpha^{(0)}$, we take the ML estimate of α^* of the same GLM model. As for $\gamma^{(0)}$, take a vector of zeros
2. **Newton-Raphson algorithm for estimation of γ^* .** As initial points, take $\beta^{(0)}$, $\alpha^{(0)}$ and $\gamma^{(0)}$. Stop at the iteration R , such that $\|\gamma^{(R)} - \gamma^{(R-1)}\|_\infty < 10^{-6}$
3. **Variable selection.** Use the LASSO criterion and for that replace $\beta^{(0)}$, $\alpha^{(0)}$ and $\hat{\gamma}$ in the formula of $\mathcal{Y}$ by $\beta^{(0)}$, $\alpha^{(0)}$ and $\gamma^{(R)}$. Then, use one of the stability selection approaches to get $\hat{\beta}$
4. **Reestimation.** We fit a GLM to the observations $Y_1, \dots, Y_n$ and the design matrix X , in which we leave only the columns corresponding to the indices i that we got in the previous step. We obtain $\hat{\beta}$ and $\hat{\alpha}$ as the final estimates

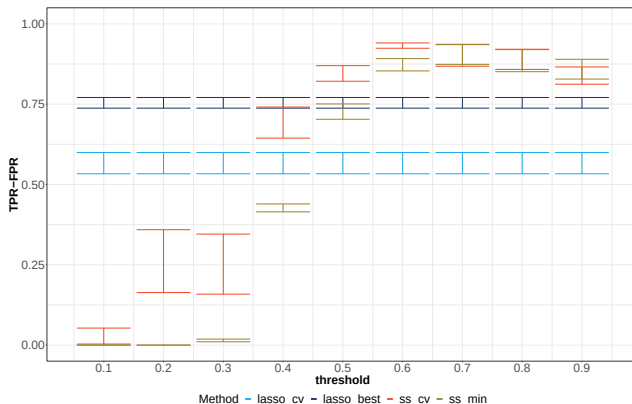
Support Recovery of β^* 

Figure: Error bars of the difference between the TPR and FPR when $n = 1000$, $q = 2$, $p = 100$, $\alpha^* = 2$, and a 5% sparsity level (10 simulations)

Impact of the Value of n and q on the Recovery of β^*

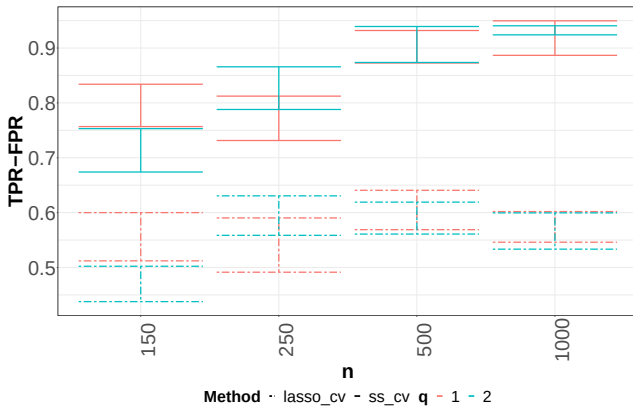


Figure: Error bars of the difference between the TPR and FPR for different values of n and q , $p = 100$, $\alpha^* = 2$, and a 5% sparsity level (10 simulations)

Impact of the Value of n on the Estimation of γ^*

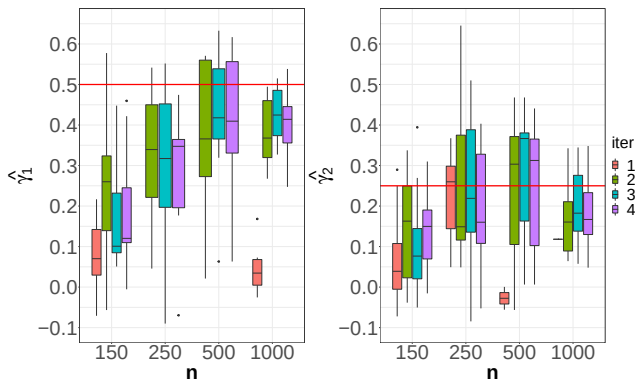


Figure: Boxplots for the estimations of γ^* for $q = 2$, $p = 100$, $\alpha^* = 2$, and a 5% sparsity level (10 simulations)

Impact of the Value of n on the Estimation of α^*

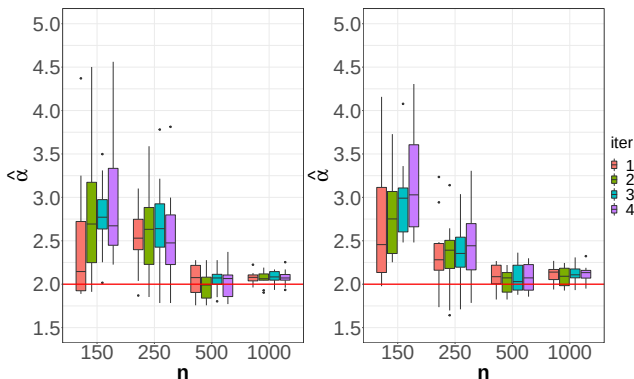


Figure: Boxplots for the estimations of α^* for $p = 100$, $\alpha^* = 2$, a 5% sparsity level, and $q = 1$ (left), $q = 2$ (right) (10 simulations)

Study of the Kinetics of Transcriptomic data

- In RNA-seq time series gene expression levels are measured at different time points
- Can be used to understand the temporal dependence existing in the gene expression
- Eukaryotic genomes of some plants are transcribed outside of protein-coding genes, named **non-coding RNAs**
 - **Coding RNAs** code for proteins
 - Among them, **long non-coding RNAs (lncRNAs)** are a heterogeneous group of RNA molecules, transcribed from non-coding genes, that regulate genome expression
- **Goal:** Identify the lncRNAs, the expression of which affects the expression of coding genes, by using the temporal evolution of the expression of both coding genes and lncRNAs
- Model plant: *Arabidopsis thaliana*

Means and the variances of RNA-Seq time series

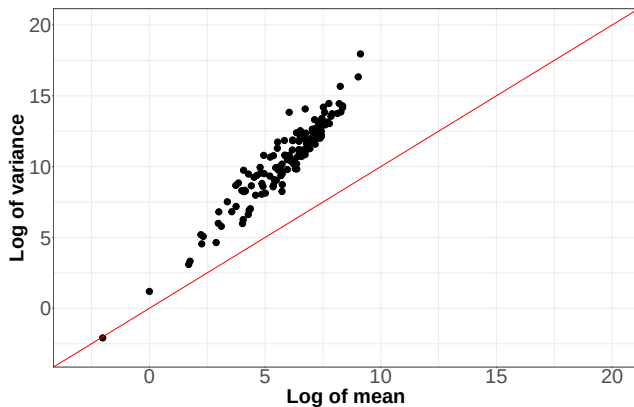


Figure: Scatter plot of the means and the variances of 145 RNA-Seq time series

Results

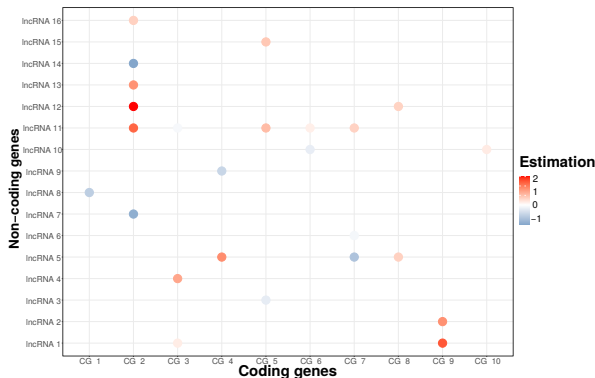


Figure: Estimation of β^* for explaining the values of 10 coding genes (Y_t) by some of the lncRNAs ($x_{t,i}$), where $n = 15$ and $p = 95$. A sample of 10 coding genes is illustrated. 37 out of 95 lncRNAs were selected, whereas the Poisson model selected 93 out of 95

Thank you for your attention!

- Paper

Gomtsyan, M. (2024). Variable selection in a specific regression time series of counts. *arXiv:2307.00929*

- R package

M. Gomtsyan *NBtsVarSel: Variable Selection in a Specific Regression Time Series of Counts* (2023)

- Acknowledgement

I would like to thank Céline Lévy-Leduc, Sarah Ouadah and Laure Sansonnet for their valuable comments