Package 'PAGFL'

February 17, 2024

Title Joint Estimation and Identification of Latent Groups in Panel Data Models

Version 1.0.1

Maintainer Paul Haimerl <paul.haimerl@maastrichtuniversity.nl>

Description In panel data analysis, unobservable group structures are a common challenge. Disregarding group-level heterogeneity by assuming an entirely homogeneous panel can introduce bias. Conversely, estimating individual coefficients for each cross-sectional unit is inefficient and may lead to high uncertainty.
This package addresses this issue by implementing the pairwise adaptive group fused Lasso (PAGFL) by Mehrabani (2023) <doi:10.1016/j.jeconom.2022.12.002>. PAGFL is an efficient methodology to identify latent group structures and estimate group-specific coefficients simultaneously.

License AGPL (>= 3)

Encoding UTF-8

RoxygenNote 7.2.3

LinkingTo Rcpp, RcppArmadillo

Imports Rcpp, pbapply

BugReports https://github.com/Paul-Haimerl/PAGFL/issues

URL https://github.com/Paul-Haimerl/PAGFL

NeedsCompilation yes

Author Paul Haimerl [aut, cre] (<https://orcid.org/0000-0003-3198-8317>), Ali Mehrabani [ctb] (<https://orcid.org/0000-0002-1848-5582>)

Repository CRAN

Date/Publication 2024-02-17 11:10:05 UTC

R topics documented:

PAGFL 2
sim_DGP 5

Index **8**

PAGFL *Apply the Pairwise Adaptive Group Fused Lasso*

Description

The pairwise adaptive group fused lasso (PAGFL) by Mehrabani (2023) jointly estimates the latent group structure and group-specific slope parameters in a panel data model. It can handle static and dynamic panels, either with or without endogenous regressors.

Usage

```
PAGFL(
  y,
  X,
  n_periods,
  lambda,
  method = "PLS",
  Z = NULL,
  min_group_frac = 0.05,
  bias_correc = FALSE,
  kappa = 2,
  max_iter = 2000,
  tol_convergence = 0.001,
  tol_group = sqrt(p/(sqrt(N * n_periods) * log(log(N * n_periods)))),
  rho = 0.07 * log(N * n_periods)/sqrt(N * n_periods),
  varrho = max(sqrt(5 * N * n_periods * p)/log(N * n_periods * p) - 7, 1),
  verbose = TRUE
)
```

Arguments

y	a $NT \times 1$ vector or data.frame of the dependent variable, with $\boldsymbol{y} = (y_1, \ldots, y_N)'$, $y_i = (y_{i1}, \ldots, y_{iT})'$ and the scalar y_{it}.
X	a $NT \times p$ matrix or data.frame of explanatory variables, with $\boldsymbol{X} = (x_1, \ldots, x_N)'$, $x_i = (x_{i1}, \ldots, x_{iT})'$ and the $p \times 1$ vector x_{it}.
n_periods	the number of observed time periods T.
lambda	the tuning parameter governing the strength of the penalty term. Either a single λ or a vector of candidate values can be passed. If a vector is supplied, a BIC-type information criterion selects the best fitting parameter value.
method	the estimation method. Options are
	'PLS' for using the penalized least squares (*PLS*) algorithm. We recommend *PLS* in case of (weakly) exogenous regressors (Mehrabani, 2023, sec. 2.2).
	'PGMM' for using the penalized Generalized Method of Moments (*PGMM*). *PGMM* is required when instrumenting endogenous regressors (Mehrabani, 2023, sec. 2.3). A matrix Z contains the necessary exogenous instruments.

Default is 'PLS'.

Z — a $NT \times q$ matrix of exogenous instruments, where $q \geq p$, $\boldsymbol{Z} = (z_1, \ldots, z_N)'$, $z_i = (z_{i1}, \ldots, z_{iT})'$ and z_{it} is a $q \times 1$ vector. Z is only required when method = 'PGMM' is selected. When using 'PLS', either pass NULL or any matrix $\boldsymbol{Z}$ is disregarded. Default is NULL.

min_group_frac — the minimum group size as a fraction of the total number of individuals N. In case a group falls short of this threshold, a hierarchical classifier allocates its members to the remaining groups. Default is 0.05.

bias_correc — logical. If TRUE, a Split-panel Jackknife bias correction following Dhaene and Jochmans (2015) is applied to the slope parameters. We recommend using this correction when facing a dynamic panel. Default is FALSE.

kappa — the weight placed on the adaptive penalty weights. Default is 2.

max_iter — the maximum number of iterations for the *ADMM* estimation algorithm. Default is 2000.

tol_convergence — the tolerance limit for the stopping criterion of the iterative *ADMM* estimation algorithm. Default is 0.001.

tol_group — the tolerance limit for within-group differences. Two individuals are placed in the same group if the Frobenius norm of their coefficient parameter difference is below this parameter. If left unspecified, the heuristic $\sqrt{\frac{p}{\sqrt{NT}\log(\log(NT))}}$ is used. We recommend the default.

rho — the tuning parameter balancing the fitness and penalty terms in the information criterion that determines the penalty parameter λ. If left unspecified, the heuristic $\rho = 0.07\frac{\sqrt{NT}\log(NT)}{NT}$ of Mehrabani (2023, sec. 6) is used. We recommend the default.

varrho — the non-negative Lagrangian *ADMM* penalty parameter. For *PLS*, the ϱ value is trivial. However, for *PGMM*, small values lead to slow convergence of the algorithm. If left unspecified, the default heuristic $\varrho = \max(\frac{\sqrt{5NTp}}{\log(NTp)} - 7, 1)$ is used.

verbose — logical. If TRUE, a progression bar is printed when iterating over candidate λ values and helpful warning messages are shown. Default is TRUE.

Details

The *PLS* method minimizes the following criterion:

$$\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \beta_i'\tilde{x}_{it})^2 + \frac{\lambda}{N}\sum_{1\leq i}\sum_{i<j\leq N}\dot{w}_{ij}\|\beta_i - \beta_j\|,$$

where $\tilde{y}_{it}$ is the de-meaned dependent variable, $\tilde{x}_{it}$ represents a vector of de-meaned weakly exogenous explanatory variables, λ is the penalty tuning parameter and $\dot{w}_{ij}$ reflects adaptive penalty weights (see Mehrabani, 2023, eq. 2.6). $\|\cdot\|$ denotes the Frobenius norm. The adaptive weights $\dot{w}_{ij}$ are obtained by a preliminary least squares estimation. The solution $\hat{\beta}$ is computed via an iterative alternating direction method of multipliers (*ADMM*) algorithm (see Mehrabani, 2023, sec. 5.1).

PGMM employs a set of instruments Z to control for endogenous regressors. Using *PGMM*, $\boldsymbol{\beta} = (\beta_1', \dots, \beta_N')'$ is estimated by minimizing:

$$\sum_{i=1}^{N} \left[\frac{1}{N} \sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i' \Delta x_{it}) \right]' W_i \left[\frac{1}{T} \sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i' \Delta x_{it}) \right] + \frac{\lambda}{N} \sum_{1 \leq i} \sum_{i < j \leq N} \ddot{w}_{ij} \|\beta_i - \beta_j\|.$$

$\ddot{w}_{ij}$ are obtained by an initial *GMM* estimation. Δ gives the first differences operator $\Delta y_{it} = y_{it} - y_{it-1}$. W_i represents a data-driven $q \times q$ weight matrix. I refer to Mehrabani (2023, eq. 2.10) for more details. $\boldsymbol{\beta}$ is again estimated employing an efficient *ADMM* algorithm (Mehrabani, 2023, sec. 5.2).

Two individuals are assigned to the same group if $\|\hat{\beta}_i - \hat{\beta}_j\| \leq \epsilon_{\text{tol}}$, where ϵ_{tol} is given by `tol_group`.

We suggest identifying a suitable λ parameter by passing a logarithmically spaced grid of candidate values with a lower limit of 0 and an upper limit that leads to a fully homogenous panel. A BIC-type information criterion then selects the best fitting λ value.

Value

A list holding

`IC`	the BIC-type information criterion.
`lambda`	the penalization parameter. If multiple λ values were passed, the parameter yielding the lowest IC.
`alpha_hat`	a $K \times p$ matrix of the post-Lasso group-specific parameter estimates.
`K_hat`	the estimated total number of groups.
`groups_hat`	a vector of estimated group memberships.
`iter`	the number of executed algorithm iterations.
`convergence`	logical. If `TRUE`, convergence was achieved. If `FALSE`, `max_iter` was reached.

Author(s)

Paul Haimerl

References

Dhaene, G., & Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3), 991-1030. doi:10.1093/restud/rdv007.

Mehrabani, A. (2023). Estimation and identification of latent group structures in panel data. *Journal of Econometrics*, 235(2), 1464-1482. doi:10.1016/j.jeconom.2022.12.002.

Examples

```
# Simulate a panel with a group structure
sim <- sim_DGP(N = 50, n_periods = 80, p = 2, n_groups = 3)
y <- sim$y
X <- sim$X

# Run the PAGFL procedure for a set of candidate tuning parameter values
```

```
lambda_set <- exp(log(10) * seq(log10(1e-4), log10(10), length.out = 10))
estim <- PAGFL(y = y, X = X, n_periods = 80, lambda = lambda_set, method = 'PLS')
print(estim)
```

`sim_DGP` *Simulate a Panel With a Latent Group Structure*

Description

Construct a static or dynamic, exogenous or endogenous panel data set subject to a latent group structure with optional $AR(1)$ or $GARCH(1,1)$ innovations.

Usage

```
sim_DGP(
  N = 50,
  n_periods = 40,
  p = 2,
  n_groups = 3,
  group_proportions = NULL,
  error_spec = NULL,
  dyn_panel = FALSE,
  q = NULL,
  alpha_0 = NULL
)
```

Arguments

`N`	the number of cross-sectional units. Default is 50.
`n_periods`	the number of simulated time periods T. Default is 40.
`p`	the number of explanatory variables. Default is 2.
`n_groups`	the number of latent groups K. Default is 3.
`group_proportions`	a numeric vector of length `n_groups` indicating the fraction of N each group will contain. If `NULL`, all groups are of size $\frac{N}{K}$. Default is `NULL`.
`error_spec`	the error specification used. Options are `NULL` for iid errors. `'AR'` for an $AR(1)$ error process with an autoregressive coefficient of 0.5. `'GARCH'` for a $GARCH(1,1)$ error process with a 0.05 constant, a 0.05 ARCH and a 0.9 GARCH coefficient. Default is `NULL`.
`dyn_panel`	Logical. If `TRUE`, the panel includes one stationary autoregressive lag of the dependent variable (see sec. Details for information on the AR coefficient). Default is `FALSE`.

`q` the number of exogenous instruments when a panel with endogenous regressors is to be simulated. If panel data set with exogenous regressors is supposed to be generated, pass `NULL`. Default is `NULL`.

`alpha_0` an optional pre-specified $K \times p$ parameter matrix. If `NULL`, the coefficients are drawn randomly (see sec. Details). If `dyn_panel = TRUE`, the first column represents the stationary AR coefficient. Default is `NULL`.

Details

The scalar dependent variable y_{it} is driven by the following panel data model

$$y_{it} = \eta_i + \beta_i' x_{it} + u_{it}, \quad i = \{1, \dots, N\}, \quad t = \{1, \dots, T\}.$$

η_i represents individual fixed effects and $x_{it} = (x_{it,1}, \dots, x_{it,p})$ a $p \times 1$ vector of regressors. The individual slope coefficient vectors β_i are subject to a latent group structure $\beta_i = \sum_{k=1}^{K} \alpha_k \mathbf{1}\{i \in G_k\}$. As a consequence, the group-level coefficients $\boldsymbol{\alpha} = (\alpha_1', \dots, \alpha_K')'$ follow the partition $\boldsymbol{G}$ of N cross-sectional units $\boldsymbol{G} = (G_1, \dots, G_K)$ such that $\cup_{k=1}^{K} = \{1, \dots, N\}$ and $G_k \cap G_l = \emptyset,\ \alpha_k \neq \alpha_l$ for any two groups $k \neq l$ (Mehrabani, 2023, sec. 2.1).

If a panel data set with exogenous regressors is generated (set `q = NULL`), the p predictors are simulated as:

$$x_{it,j} = 0.2\eta_i + e_{it,j}, \quad \eta_i, e_{it,j} \sim i.i.d.N(0,1), \quad j = \{1, \dots, p\},$$

where $e_{it,j}$ denotes a series of innovations. η_i and e_i are independent of each other.

In case `alpha_0 = NULL`, the group-level slope parameters α_k are drawn from $\sim U[-2, 2]$.

If a dynamic panel is specified (`dyn_panel = TRUE`), the AR coefficients β_i^{AR} are drawn from a uniform distribution with support $(-1, 1)$ and $x_{it,j} = e_{it,j}$. The individual fixed effects enter the dependent variable via $(1 - \beta_i^{\text{AR}})\eta_i$ to account for the autoregressive dependency. I refer to Mehrabani (2023, sec 6) for details.

When specifying an endogenous panel (set `q` to $q \geq p$), $e_{it,j}$ correlate with the cross-sectional innovations u_{it} by a magnitude of 0.5 to produce endogenous regressors with $\mathrm{E}(u|X) \neq 0$. However, the endogenous regressors can be accounted for by exploiting the q instruments in $\boldsymbol{Z}$, for which $\mathrm{E}(u|Z) = 0$ holds. The instruments and the first stage coefficients are generated in the same fashion as $\boldsymbol{X}$ and $\boldsymbol{\alpha}$ when `q = NULL`, respectively.

The function nests, among other, the DGPs employed in the simulation study of Mehrabani (2023, sec. 6).

Value

A list holding

`alpha` the $K \times p$ matrix of group-specific slope parameters. In case of `dyn_panel = TRUE`, the first column holds the AR coefficient.

`groups` a vector indicating the group memberships.

`y` a $NT \times 1$ vector of the dependent variable, with $\boldsymbol{y} = (y_1, \dots, y_N)'$, $y_i = (y_{i1}, \dots, y_{iT})'$ and the scalar y_{it}.

`X` a $NT \times p$ matrix of explanatory variables, with $\boldsymbol{X} = (x_1, \dots, x_N)'$, $x_i = (x_{i1}, \dots, x_{iT})'$ and the $p \times 1$ vector x_{it}.

Z — a $NT \times q$ matrix of instruments , where $q \geq p$, $\boldsymbol{Z} = (z_1, \ldots, z_N)'$, $z_i = (z_{i1}, \ldots, z_{iT})'$ and z_{it} is a $q \times 1$ vector. In case a panel with exogenous regressors is generated (q = NULL), $\boldsymbol{Z}$ equals NULL.

Author(s)

Paul Haimerl

References

Mehrabani, A. (2023). Estimation and identification of latent group structures in panel data. *Journal of Econometrics*, 235(2), 1464-1482. doi:10.1016/j.jeconom.2022.12.002.

Examples

```
# Simulate DGP 1 from Mehrabani (2023, sec. 6)
alpha_0_DGP1 <- matrix(c(0.4, 1, 1.6, 1.6, 1, 0.4), ncol = 2)
DGP1 <- sim_DGP(
  N = 50, n_periods = 20, p = 2, n_groups = 3,
  group_proportions = c(.4, .3, .3), alpha_0 = alpha_0_DGP1
)

# Simulate DGP 6 from Mehrabani (2023, sec. 6)
alpha_0_DGP6 <- cbind(
  c(0.8, 0.6, 0.4, 0.2, -0.2, -0.4, -0.6, -0.8),
  c(-4, -3, -2, -1, 1, 2, 3, 4),
  c(4, 3, 2, 1, -1, -2, -3, -4)
)
```

Index

PAGFL, 2

sim_DGP, 5