

Simultaneous semi-parametric estimation of clustering and regression

Matthieu Marbac¹, Mohammed Sedki², Christophe Biernacki³,
Vincent Vandewalle⁴

¹ Univ. Rennes, Ensai, CNRS, CREST - UMR 9194

² Univ. Paris-Sud and Inserm

³ Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé

⁴ Inria, Univ. Lille, CHU Lille, ULR 2694 - METRICS

Fix group effect regression with missing group

- Regression with fixed group effect
- Group variable missing
- Group-related variables available

Example: High blood pressure (Berney et al., 2018)

- Link with the level of physical activity
- Level of physical activity not easily available
- Mean time spent doing various physical activity available

Specific mixture of regressions (Wang et al., 1996; Hunter and Young, 2012; Wu and Yao, 2016)

- Only intercepts vary among components
- Mixture weights related to other variables
- Quantization of the variables related to the group membership (Charlier et al., 2015)

The usual two-step approach

Two-step approach (Auray et al., 2015; Ando and Bai, 2016; Zhang et al., 2019)

- 1 Clustering based on group-related variables
- 2 Regression fitted with estimated group variable as covariate

The usual two-step approach

Two-step approach (Auray et al., 2015; Ando and Bai, 2016; Zhang et al., 2019)

- 1 Clustering based on group-related variables
- 2 Regression fitted with estimated group variable as covariate

Remarks

- Usual in epidemiology, economics
- ⚠ Suboptimal procedures
 - Step 2: regression model with measurement errors
 - Information about the group in the target variable \Rightarrow target variable relevant for clustering

Latent variables models

- Latent class, latent profile analysis (Guo et al., 2006; Kim et al., 2016)
- No focus on a particular variable, parametric framework
- Latent variable mixed effect model: arbitrary covariate effects, direct modelling of covariate on the latent variable

Concomitant variables framework (Dayton and Macready, 1988; Grün and Leisch, 2008; Vaňkátová and Fišerová, 2017)

- Variables to locally adjust the weights of the mixture of regressions
- More focused on regression than on clustering

Cluster-weighted models(Mazza et al., 2018)

- Joint mixture model on the response and the covariates
- All the covariates are used in the clustering

Simultaneously estimate clustering and regression

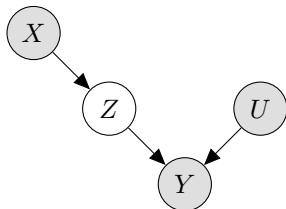
- Full parametric or semi-parametric framework
- Semi-parametric framework circumvents biased estimators

Focus on semi-parametric approach (Hunter et al., 2011)

- Component densities as a product of univariate densities (Chauveau et al., 2015; Zhu and Hunter, 2016; Zheng and Wu, 2019)
- Identifiable if densities are linearly independent and at least 3 variables used for clustering (Allman et al., 2009)
- Inference by maximum smoothed likelihood (Levine et al., 2011)
- Allow to consider robust losses in regression and flexible clustering

Notations

- $(V^\top, X^\top, Y)^\top$: random variables
- $V = (U^\top, Z^\top)^\top$: vector of covariates
- $Y \in \mathbb{R}$: the variable to predict
- $Z = (Z_1, \dots, Z_K)^\top \in \mathcal{Z}$ a categorical variable with K levels



Example on high blood pressure

- Y : diastolic blood pressure
- U : gender, age, alcohol consumption, obesity and sleep quality
- X : is the set of covariates measuring the level of physical activity (practice of that activity, hours spent watching TV, hours spent on the computer, etc.)
- Z : group of subjects with similar physical activity behaviours

Regression model

$$Y = V^T \beta + \varepsilon \text{ with } \mathbb{E}[\rho(\varepsilon)|V] = 0$$

Regression model

$$Y = V^\top \beta + \varepsilon \text{ with } \mathbb{E}[\rho(\varepsilon)|V] = 0$$

	$\mathcal{L}(t)$	$\rho(t)$
Mean	t^2	$2t$
τ -quantile	$ t + (2\tau - 1)t$	$2(\tau - \mathbf{1}_{\{t \leq 0\}})$
τ -expectile	$ \tau - \mathbf{1}\{t \leq 0\} t^2$	$2t((1 - \tau)\mathbf{1}\{t \leq 0\} + \tau\mathbf{1}\{t > 0\})$

Regression model

$$Y = V^{\top} \beta + \varepsilon \text{ with } \mathbb{E}[\rho(\varepsilon)|V] = 0$$

Clustering model

$$f(x; \vartheta) = \sum_{k=1}^K \pi_k f_k(x)$$

Joint predictive clustering model

Regression model

$$Y = V^\top \beta + \varepsilon \text{ with } \mathbb{E}[\rho(\varepsilon)|V] = 0$$

Clustering model

$$f(x; \vartheta) = \sum_{k=1}^K \pi_k f_k(x)$$

	parametric	semi-parametric
$f_k(x)$	$f_k(x; \theta_k)$	$\prod_{j=1}^{d_X} f_{kj}(x_j)$

Regression model

$$Y = V^\top \beta + \varepsilon \text{ with } \mathbb{E}[\rho(\varepsilon)|V] = 0$$

Clustering model

$$f(x; \vartheta) = \sum_{k=1}^K \pi_k f_k(x)$$

Joint clustering and regression model

$$f(x, y|u; \theta) = \sum_{k=1}^K \pi_k f_k(x) f_\varepsilon(y - u^\top \gamma - \delta_k),$$

where, for $k = 1, \dots, K$

$$\mathbb{E}[\rho(Y - U^\top \gamma - \delta_k)|U, Z_k = 1] = 0$$

Lemma 1

Let a model with identifiable parameters defined in preview slide, noting

$$r_k^{X,Y}(x, y) := \mathbb{P}(Z_k = 1|x, y) = \frac{\pi_k f_k(x) f_\varepsilon(y - u^\top \gamma - \delta_k)}{\sum_{\ell=1}^K \pi_\ell f_\ell(x) f_\varepsilon(y - U^\top \gamma - \delta_\ell)},$$

then $\beta = (\delta^\top, \gamma^\top)^\top$ is the single parameter satisfying

$$\forall k = 1, \dots, K, \mathbb{E}[r_k^{X,Y}(X, Y) \rho(Y - u^\top \gamma - \delta_k) | U, X] = 0. \quad (1)$$

Lemma 1

Let a model with identifiable parameters defined in preview slide, noting

$$r_k^{X,Y}(x, y) := \mathbb{P}(Z_k = 1|x, y) = \frac{\pi_k f_k(x) f_\varepsilon(y - u^\top \gamma - \delta_k)}{\sum_{\ell=1}^K \pi_\ell f_\ell(x) f_\varepsilon(y - U^\top \gamma - \delta_\ell)},$$

then $\beta = (\delta^\top, \gamma^\top)^\top$ is the single parameter satisfying

$$\forall k = 1, \dots, K, \mathbb{E}[r_k^{X,Y}(X, Y) \rho(Y - u^\top \gamma - \delta_k) | U, X] = 0. \quad (1)$$

Remark

- Moment equation verified on the joint model
- Justify the need for a simultaneous approach

The two step approach

- 1 **Clustering step** Perform a clustering of \mathbb{X} to obtain an estimated hard classification rule $\hat{r}^X : \mathbb{R}^{d_x} \rightarrow \mathcal{Z}$ or an estimated fuzzy classification rule $\hat{r}^X : \mathbb{R}^{d_x} \rightarrow \tilde{\mathcal{Z}}_K$ where $\tilde{\mathcal{Z}}_K$ is the simplex of size K .

The two step approach

- 1 **Clustering step** Perform a clustering of \mathbb{X} to obtain an estimated hard classification rule $\hat{r}^X : \mathbb{R}^{d_X} \rightarrow \mathcal{Z}$ or an estimated fuzzy classification rule $\hat{r}^X : \mathbb{R}^{d_X} \rightarrow \tilde{\mathcal{Z}}_K$ where $\tilde{\mathcal{Z}}_K$ is the simplex of size K .
- 2 **Regression step** Estimation of the regression parameters given the estimator of the group memberships

$$\hat{\beta}^{\hat{r}^X} := (\hat{\gamma}^{\hat{r}^X \top}, \hat{\delta}^{\hat{r}^X \top})^\top = \arg \min_{\beta} \sum_{i=1}^n \sum_{k=1}^K \hat{r}_k^X(x_i) \mathcal{L}(y_i - u_i^\top \gamma - \delta_k)$$

where $\hat{r}_k^X(x_i)$ is the element k of vector $\hat{r}^X(x_i)$.

Lemma 2

Let an identifiable model, then

- 1 Assume that \hat{r}^X converges to the best classification rule based on X , denoted by \tilde{r}^X . Then \hat{r}^X is asymptotically suboptimal since

$$\mathbb{E} \left[\sum_{k=1}^K \tilde{r}_k^X(X) Z_k \right] < \mathbb{E} \left[\sum_{k=1}^K r_k^{X,Y}(X, Y) Z_k \right].$$

- 2 Let f_ε defines a random variable with finite variance and U to have a covariance matrix with non-zero eigenvalues. Then, considering the quadratic loss and a consistent fuzzy classification rule \hat{r}^X , the estimator $\hat{\gamma}^{\hat{r}^X}$ is asymptotically unbiased but the estimator $\hat{\delta}^{\hat{r}^X}$ is asymptotically biased since

$$\lim_{n \rightarrow \infty} \text{bias}(\hat{\delta}_k^{\hat{r}^X}) = \frac{\sum_{\ell=1}^K \Delta_{k\ell} \delta_\ell}{\sum_{h=1}^K \Delta_{kh}} - \delta_k,$$

where $\Delta_{k\ell} = \mathbb{E}[r_k^X(X) r_\ell^X(X)]$.

Semi-parametric model

$$f(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k f_k(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k \prod_{j=1}^{d_X} f_{kj}(x_j) f_\varepsilon(y - u^\top \gamma - \delta_k).$$

Let $w := (x^\top, y)^\top$, thus $f(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k f_k(w \mid u; \theta)$.

Semi-parametric model

$$f(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k f_k(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k \prod_{j=1}^{d_X} f_{kj}(x_j) f_\varepsilon(y - u^\top \gamma - \delta_k).$$

Let $w := (x^\top, y)^\top$, thus $f(y, x \mid u; \theta) = \sum_{k=1}^K \pi_k f_k(w \mid u; \theta)$.

Sufficient condition for identifiability: marginal distribution of X is identifiable and thus a sufficient condition is to consider linearly independent densities f_{kj} 's and $d_X \geq 3$ (Allman et al., 2009).

Semi-parametric model

$$f(y, x | u; \theta) = \sum_{k=1}^K \pi_k f_k(y, x | u; \theta) = \sum_{k=1}^K \pi_k \prod_{j=1}^{d_X} f_{kj}(x_j) f_\varepsilon(y - u^\top \gamma - \delta_k).$$

Let $w := (x^\top, y)^\top$, thus $f(y, x | u; \theta) = \sum_{k=1}^K \pi_k f_k(w | u; \theta)$.

Smoothed log-likelihood (Levine et al., 2011)

Let $(\mathcal{N}f_k)(w | u; \theta) = \exp \left\{ \int K_h(w - \tilde{w}) \ln f_k(\tilde{w} | u; \theta) d\tilde{w} \right\}$ with $K_h(a) = \prod_{j=1}^d K_h(a_j)$ and $K_h(a_j)$ the rescaled kernel function. We want to maximize the smoothed log-likelihood

$$\ell(\theta) = \sum_{i=1}^n \ln \left(\sum_{k=1}^K \pi_k (\mathcal{N}f_k)(w_i | u_i; \theta) \right)$$

subject to $\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{\pi_k f_k(w_i | u_i; \theta)}{\sum_{\ell=1}^K \pi_\ell f_\ell(w_i | u_i; \theta)} \rho(y_i - u_i^\top \gamma - \delta_k) = 0$.

Minorize-Maximize algorithm

Let an initial value $\theta^{[0]}$, an iteration $[r]$ is defined by

- **Minorize step:**

$$t_{ik}^{[r-1]} = \frac{\pi_k^{[r-1]} \left(\mathcal{N} f_k^{[r-1]} \right) (w_i | u_i; \theta^{[r-1]})}{\sum_{\ell=1}^K \pi_\ell^{[r-1]} \left(\mathcal{N} f_\ell^{[r-1]} \right) (w_i | u_i; \theta^{[r-1]})}.$$

- **Maximize step:**

- 1 Updating the parametric elements

$$\pi_k^{[r]} = \frac{1}{n} \sum_i t_{ik}^{[r-1]} \text{ and } \beta^{[r]} = \arg \min_{\beta} \sum_{i,k} t_{ik}^{[r-1]} \mathcal{L}(y_i - u_i^\top \gamma - \delta_k).$$

- 2 Updating the nonparametric elements

$$f_{kj}(a) = \frac{1}{n\pi_k^{[r]}} \sum_i t_{ik}^{[r-1]} K_h(x_{ij} - a)$$
$$f_\varepsilon(a) = \frac{1}{n} \sum_{i,k} t_{ik}^{[r-1]} K_h(y_i - u_i^\top \gamma^{[r]} - \delta_k^{[r]} - a).$$

Minorize-Maximize algorithm

Let an initial value $\theta^{[0]}$, an iteration $[r]$ is defined by

- **Minorize step:**

$$t_{ik}^{[r-1]} = \frac{\pi_k^{[r-1]} \left(\mathcal{N} f_k^{[r-1]} \right) (w_i | u_i; \theta^{[r-1]})}{\sum_{\ell=1}^K \pi_\ell^{[r-1]} \left(\mathcal{N} f_\ell^{[r-1]} \right) (w_i | u_i; \theta^{[r-1]})}.$$

- **Maximize step:**

- 1 Updating the parametric elements

$$\pi_k^{[r]} = \frac{1}{n} \sum_i t_{ik}^{[r-1]} \text{ and } \beta^{[r]} = \arg \min_{\beta} \sum_{i,k} t_{ik}^{[r-1]} \mathcal{L}(y_i - u_i^\top \gamma - \delta_k).$$

- 2 Updating the nonparametric elements

$$f_{kj}(a) = \frac{1}{n\pi_k^{[r]}} \sum_i t_{ik}^{[r-1]} K_h(x_{ij} - a)$$
$$f_\varepsilon(a) = \frac{1}{n} \sum_{i,k} t_{ik}^{[r-1]} K_h(y_i - u_i^\top \gamma^{[r]} - \delta_k^{[r]} - a).$$

The MM algorithm increases the value of the smoothed likelihood at each iteration.

Simulated data

- $Z_i \sim \mathcal{M}(\pi_1, \pi_2)$, $\pi_1 = \pi_2 = 0.5$
- $X_{ij} = (-1)^{Z_{i1}} \xi + \eta_{ij}$, η_{ij} i.i.d, $j = 1, \dots, 4$
- $U_i \sim \mathcal{N}_2(0, \mathbf{I}_2)$
- $Y_i = (Z_{i1}, Z_{i2}, U_{i1}, U_{i2})\beta + \varepsilon_i$, $\beta = (-1, 1, 1, 1)^\top$, ε_i i.i.d

Simulation setting

- Different distributions for η_{ij} and ε_i
- 100 samples are generated for each case
- ξ control the class overlap: tuned to have a misclassification of 0.10

Evaluation metrics

- **Regression:** MSE of the coefficient of the regression β
- **Clustering:** adjusted rand index (ARI) between the true and the estimated partition

Semi-parametric approach applied with bandwidth $h = n^{-1/5}$.

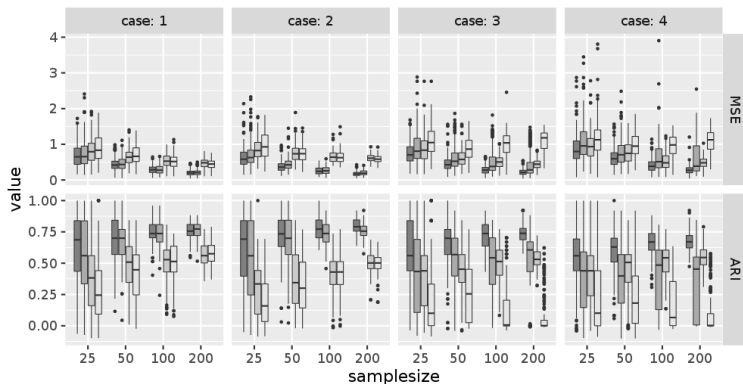
Simultaneous vs two-step \times parametric vs nonparametric

Quadratic loss

Mixture \ Noise	Well-specified	Miss-specified
Well-specified	Case 1: $\eta_{ij} \sim \mathcal{N}(0, 1)$ $\varepsilon_i \sim \mathcal{N}(0, 1)$	Case 2: $\eta_{ij} \sim \mathcal{N}(0, 1)$ $\varepsilon_i = \tau_i - 1, \tau_i \sim \mathcal{E}(1)$
Miss-specified	Case 3: $\eta_{ij} \sim \mathcal{T}_3$ $\varepsilon_i \sim \mathcal{N}(0, 1)$	Case 4: $\eta_{ij} \sim \mathcal{T}_3$ $\varepsilon_i \sim \mathcal{T}_3$

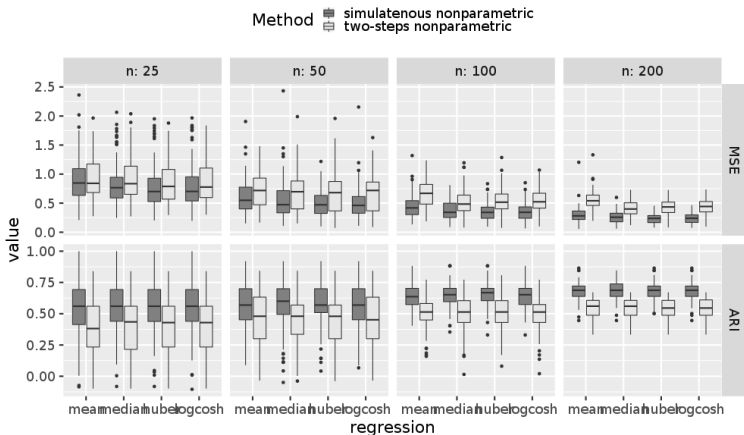
Method

- simultaneous nonparametric
- two-steps nonparametric
- simultaneous parametric
- two-steps parametric



Robust regression

- Different losses: mean, median, Huber with parameter 1 and logcosh
- $\eta_{ij} \sim \mathcal{T}_3$ and $\varepsilon_i \sim \mathcal{T}_3$



National Health and Nutrition Examination Survey of 2011-2012 ¹

- $n = 2626$ subjects between 18 and 60 years old.
- Target variable (Y): *diastolic blood pressure* in mmHg
- 7 covariates (U): gender, age, alcohol, obesity, sleep, smoke, cholesterol
- 7 variables (X) related the physical activity
 - 5 binary: indicated whether the subject has a vigorous work activity
 - 2 categorical: time spent watching TV, time spent using a computer (7 levels each)

Experimental design

- 67% learning / 33% test
- Smoothing with a Gaussian kernel
- Bandwidth $h = n^{-1/5}$.

¹<https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes>

Results

	<i>regquadUZ-K3</i>	<i>regmedUZ-K3</i>	<i>reglogchUZ-K3</i>
MSEP test	122.55	122.44	122.48
Gender (woman)	-3.07	-3.22	-3.13
Age	0.24	0.23	0.23
Alcohol (no)	-1.47	-2.48	-2.48
Obesity (yes)	3.00	2.72	2.72
Sleep	-0.48	-0.39	-0.31
Smoke (no)	0.09	0.42	0.51
Cholesterol	0.03	0.03	0.03
Class 1	60.28	59.04	58.58
Class 2	60.52	59.75	59.48
Class 3	61.09	60.14	59.62

Results

	<i>regquadUZ-K3</i>	<i>regmedUZ-K3</i>	<i>reglogchUZ-K3</i>
MSEP test	122.55	122.44	122.48
Gender (woman)	-3.07	-3.22	-3.13
Age	0.24	0.23	0.23
Alcohol (no)	-1.47	-2.48	-2.48
Obesity (yes)	3.00	2.72	2.72
Sleep	-0.48	-0.39	-0.31
Smoke (no)	0.09	0.42	0.51
Cholesterol	0.03	0.03	0.03
Class 1	60.28	59.04	58.58
Class 2	60.52	59.75	59.48
Class 3	61.09	60.14	59.62

- Class 1 ($\pi_1 = 0.15$): recreational physical activities, traveling by foot or by bike, no physical activity at work, spending few hours by watching screens
- Class 2 ($\pi_2 = 0.44$): few physical activities
- Class 3 ($\pi_3 = 0.44$): high physical activities at work

	<i>regquadUZ-K3</i>	<i>regmedUZ-K3</i>	<i>reglogchUZ-K3</i>
MSEP test	122.55	122.44	122.48
Gender (woman)	-3.07	-3.22	-3.13
Age	0.24	0.23	0.23
Alcohol (no)	-1.47	-2.48	-2.48
Obesity (yes)	3.00	2.72	2.72
Sleep	-0.48	-0.39	-0.31
Smoke (no)	0.09	0.42	0.51
Cholesterol	0.03	0.03	0.03
Class 1	60.28	59.04	58.58
Class 2	60.52	59.75	59.48
Class 3	61.09	60.14	59.62

- Class 1 ($\pi_1 = 0.15$): recreational physical activities, traveling by foot or by bike, no physical activity at work, spending few hours by watching screens
- Class 2 ($\pi_2 = 0.44$): few physical activities
- Class 3 ($\pi_3 = 0.44$): high physical activities at work

Conclusion: Moderate physical activities (recreational activities, traveling by bike or foot, not spending many hours by watching screens) prevents the high blood pressure.

Conclusion

- Alternative to the two-step approach
- Simultaneous clustering and regression: improves the accuracy of the regression parameters and of the partition
- Applicable to a wide range of regression models
- Semi-parametric approach avoids bias in the estimation

Conclusion







- Alternative to the two-step approach
- Simultaneous clustering and regression: improves the accuracy of the regression parameters and of the partition
- Applicable to a wide range of regression models
- Semi-parametric approach avoids bias in the estimation

Perspectives

- Take into account prediction quality for selecting the number of components and bandwidth
- Investigate selecting the variables in both regression and clustering
- Consider location scale symmetric distributions (Hunter et al., 2007) or an independent component analysis structure (Zhu and Hunter, 2019).
- Extension possible to several latent categorical variables explained by different sub-groups of variables of X (Marbac and Vandewalle, 2019)

Thank you!

Question?

-  Allman, Elizabeth S, Catherine Matias, John A Rhodes, et al. (2009). “Identifiability of parameters in latent structure models with many observed variables”. In: *The Annals of Statistics* 37.6A, pp. 3099–3132.
-  Ando, Tomohiro and Jushan Bai (2016). “Panel data models with grouped factor structure under unknown group membership”. In: *Journal of Applied Econometrics* 31.1, pp. 163–191.
-  Auray, Stéphane, Nicolas Klutchnikoff, Laurent Rouviere, et al. (2015). “On clustering procedures and nonparametric mixture estimation”. In: *Electronic journal of statistics* 9.1, pp. 266–297.
-  Berney, M, M Burnier, and G Wuerzner (2018). “Isolated diastolic hypertension: do we still have to care about it?”. In: *Revue medicale suisse* 14.618, pp. 1607–1610.
-  Charlier, Isabelle, Davy Paindaveine, and Jérôme Saracco (2015). “Conditional quantile estimation through optimal quantization”. In: *Journal of Statistical Planning and Inference* 156, pp. 14 –30.
-  Chauveau, Didier, David R Hunter, Michael Levine, et al. (2015). “Semi-parametric estimation for conditional independence multivariate finite mixture models”. In: *Statistics Surveys* 9, pp. 1–31.

-  Dayton, C Mitchell and George B Macready (1988). “Concomitant-Variable Latent-Class Models”. en. In: *Journal of the American Statistical Association* 83.401, pp. 173–178.
-  Grün, Bettina and Friedrich Leisch (2008). “FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters”. en. In: *Journal of Statistical Software* 28.4.
-  Guo, Jia, Melanie Wall, and Yasuo Amemiya (2006). “Latent class regression on latent factors”. In: *Biostatistics* 7.1, pp. 145–163.
-  Hunter, David R, Donald St P Richards, and James L Rosenberger (2011). *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P Hettmansperger, the Pennsylvania State University, USA, 23-24 May 2008*. World Scientific.
-  Hunter, David R, Shaoli Wang, and Thomas P Hettmansperger (2007). “Inference for mixtures of symmetric distributions”. In: *The Annals of Statistics*, pp. 224–251.
-  Hunter, David R and Derek S Young (2012). “Semiparametric mixtures of regressions”. In: *Journal of Nonparametric Statistics* 24.1, pp. 19–38.



Kim, Minjung, Jeroen Vermunt, Zsuzsa Bakk, Thomas Jaki, and M Lee Van Horn (2016). “Modeling predictors of latent classes in regression mixture models”. In: *Structural Equation Modeling: A Multidisciplinary Journal* 23.4, pp. 601–614.



Levine, Michael, David R Hunter, and Didier Chauveau (2011). “Maximum smoothed likelihood for multivariate mixtures”. In: *Biometrika*, pp. 403–416.



Marbac, Matthieu and Vincent Vandewalle (2019). “A tractable multi-partitions clustering”. In: *Computational Statistics & Data Analysis* 132. Special Issue on Biostatistics, pp. 167 –179.



Mazza, Angelo, Antonio Punzo, and Salvatore Ingrassia (2018). “flexCWM: a flexible framework for cluster-weighted models”. In: *J Stat Softw* 86.2, pp. 1–30.



Vaňkátová, Kristýna and Eva Fišerová (2017). “The Evaluation of a Concomitant Variable Behaviour in a Mixture of Regression Models”. en. In: *Statistika* 97.4, p. 16.

-  Wang, Peiming, Martin L Puterman, Iain Cockburn, and Nhu Le (1996). “Mixed Poisson regression models with covariate dependent rates”. In: *Biometrics*, pp. 381–400.
-  Wu, Qiang and Weixin Yao (2016). “Mixtures of quantile regressions”. In: *Computational Statistics & Data Analysis* 93, pp. 162–176.
-  Zhang, Yingying, Huixia Judy Wang, and Zhongyi Zhu (2019). “Quantile-regression-based clustering for panel data”. In: *Journal of Econometrics*.
-  Zheng, Chaowen and Yichao Wu (2019). “Nonparametric estimation of multivariate mixtures”. In: *Journal of the American Statistical Association*, pp. 1–16.
-  Zhu, Xiaotian and David R Hunter (2016). “Theoretical grounding for estimation in conditional independence multivariate finite mixture models”. In: *Journal of Nonparametric Statistics* 28.4, pp. 683–701.
-  – (2019). “Clustering via finite nonparametric ICA mixture models”. In: *Advances in Data Analysis and Classification* 13.1, pp. 65–87.