A Regression Model on Quantiles Extracted from Scans of Asthmatic Patients

Marie-Felicia Beclin [1] & Pierre Lafaye de Micheaux [2] & Nicolas Molinari[3]

[1] *IDESP, Université de Montpellier, PreMeDICaL, Inria-Inserm, France, marie-felicia.beclin@umontpellier.fr*
[2] *UNSW Sydney lafaye@unsw.edu.au*
[3] *IDESP, Université de Montpellier, PreMeDICaL, Inria-Inserm, France, nicolas.molinari@inserm.fr*

Résumé. Nous nous intéressons à l'évaluation de l'efficacité du Benralizumab, un médicament utilisé pour traiter l'asthme, en utilisant des scans tomographiques capturés pendant l'expiration et l'inspiration avant et après un an de traitement. L'hypothèse médicale de travail postule que les patients dont l'état s'est amélioré présenteront des images de scanners thoraciques en expiration améliorées après le traitement. C'est-à-dire que le patient expire mieux et donc son poumon se vide plus en expiration. Cela se manifeste par des valeurs d'unité Hounsfield plus élevées. Il y a alors un déplacement vers la droite dans l'histogramme construit à partir de l'image post-traitement par rapport à celui pré-traitement.

Irpino et Verde[1] ont adopté la méthode classique de la régression linéaire de manière à pouvoir l'appliquer aux fonctions quantiles plutôt qu'aux observations réelles. Nous généralisons leur approche et obtenons les lois des estimateurs des paramètres du modèle via une approche de maximum de vraisemblance. À partir de l'espace S_Q des fonctions quantiles, nous définissons des polynômes de quantiles. Nous nous penchons ensuite sur le cas particulier linéaire. Nous définissons alors explicitement les estimateurs pas maximum de vraisemblance.

Le modèle a été implémenté en Python et appliqué à un ensemble de données réelles de 40 patients traités par Benralizumab.

L'approche décrite ci-dessus présente certaines limites, notamment la perte d'informations spatiales et l'hypothèse de relations linéaires entre les distributions de voxels. Des recherches supplémentaires sont nécessaires pour développer une méthode de régression plus générale, telle que celle proposée par Chen[2] et Ghodrati et Panaretos[3]. Cependant, notre approche a l'avantage d'être simple, facile à utiliser et comprise par les praticiens. Le recalage des images en inspiration et en expiration[4] permet une correspondance voxel par voxel. Nous pouvons alors généraliser cette approche précédente. Des recherches en cours visent à prédire des histogrammes 2D post-traitement à partir de scanners pendant l'inspiration et l'expiration après enregistrement, ainsi que des histogrammes pré-traitement correspondants, tout en incluant des covariables scalaires.

Mots-clés. Régression de distribution sur distribution, Fonctions quantiles, Biomarqueur dérivé de l'imagerie, Prédiction de traitement, Histogrammes.

Abstract. We are interested in evaluating the efficacy of Benralizumab, a medication to treat asthma, by using tomography scans captured during expiration and inspiration

before and after one year of treatment. The medical working hypothesis posits that patients with improved conditions will exhibit enhanced expiration scans after treatment, which is manifested by higher Hounsfield unit values. This results in a shift to the right in the histogram built from post-treatment image compared to the pre-treatment one.

Irpino and Verde[1] mimicked the classical linear regression method so that it can be applied to quantile functions instead of real-valued observations. We generalize their approach and obtain confidence intervals and laws of the estimators in the model via a maximum likelihood approach. From the space S_Q of quantile functions, we define quantile polynomials. We then focus on the specific linear case. We explicitly define the maximum likelihood estimators.

The model was implemented in a Python code and applied to a real data set of 40 patients treated by Benarlizumab.

The approach described above has some limitations, including the loss of spatial information and the assumption of linear relationships between voxel distributions. Further investigation is needed to develop a more general distribution-on-distribution regression method, such as the works by Chen[2] and Ghodrati and Panaretos[3]. But our approach has the advantage of being simple, easy to use and understood by practitioners. The registration of images in inspiration and expiration[4] allows for a voxel-to-voxel correspondence. We can then generalize our previous approach. Ongoing research aims to predict post-processing 2D histograms from CT scans during inspiration and expiration after recording, as well as corresponding pre-processing histograms, all while including scalar covariates.

Keywords. Distribution on Distribution Regression, Quantiles Functions, Imaging-derived biomarker, Treatment prediction, Histograms.

1 Des images aux Quantile

1.1 Données cliniques

Tous les patients de l'étude ont été traités par Benralizumab sur une période de 48 semaines. Le médicament a été administré par voie sous-cutanée à une dose de 30 mg par injection. Les données récoltées sont diverses : données cliniques ; mesure de l'examen fonctionnel respiratoire et enfin des images de scanner thoracique 3D en inspiration et en expiration. Une image scanner se compose d'environ 600 images 2-D de dimensions 512×512. Les données brutes sont stockées dans des fichiers au format DICOM et, lors du prétraitement, chaque scan est converti en un tableau numérique de dimensions $600 \times 512 \times 512$ pixels. Les valeurs des données des voxels sont exprimées en valeurs de Hounsfield (HU). Les poumons sont ensuite segmentés en appliquant un algorithme de segmentation par seuillage[4] ou par réseau neuronal [5]. Pour chaque patient i, N_i voxels sont sélectionnés par segmentation en fonction de la taille du poumon du patient. La valeur d'unité Hounsfield la plus basse possible, -1024 HU, correspond à de l'air. Nous utilisons les histogrammes des valeurs d'unité Hounsfield avant et après le traitement comme méthode pour identifier les répondeurs cliniques au Benralizumab.

Cependant, il est important de noter que par cette méthode, nous sacrifions des informations sur la distribution spatiale des voxels.

1.2 Les fonctions quantiles

Pour une image 3D en niveaux de gris , on passe à l'histogramme des valeurs.

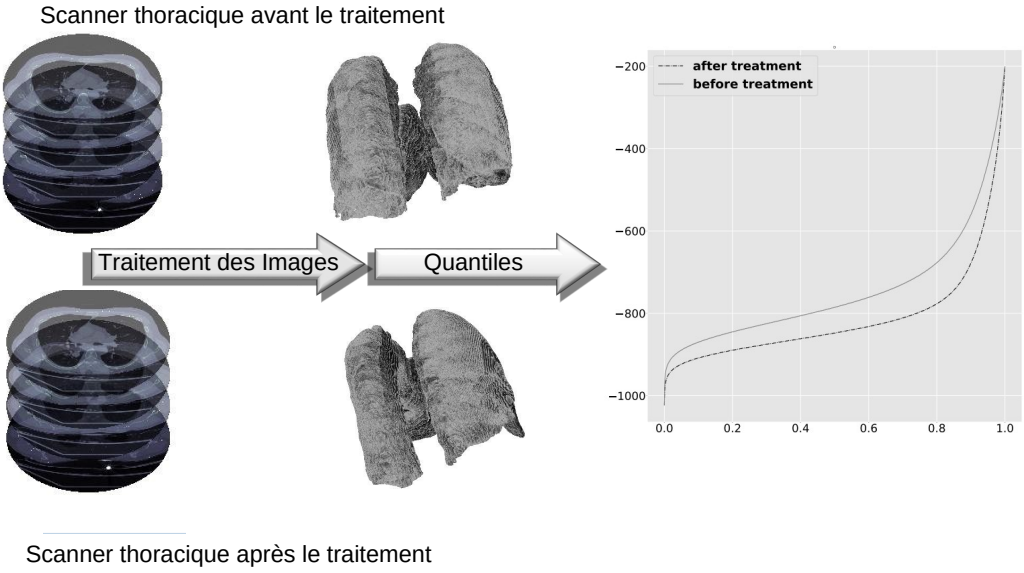

Figure 1: Illustration du processus d'extraction des données quantiles à partir des images de scanners thoraciques

La fonction quantile q^X associée à une variable aléatoire est définie pour tout $p \in]0,1[$:

$$q^X(p) \quad = \quad \inf_{t\in\mathbb{R}}\{t, F_X(t) \geq p\} \text{ avec } F^X \text{ la fonction de répartition de } X. \tag{1}$$

Nous noterons $x_q = \int_0^1 q^X(p)dp$. Ainsi, nous introduisons l'espace des fonctions quantiles, de carré intégrable, doté de la distance suivante : $d(q^X, q^Y) = \sqrt{\int_0^1 (q^X(t) - q^Y(t))^2 dt}$. Les données sur lesquelles nous appliquons notre nouveau modèle de régression sont donc n paires de fonctions quantiles $(q_1^X, q_1^Y), \ldots, (q_n^X, q_n^Y)$ extraites des scans pris sur n sujets avant et après le traitement.

1.3 Le modèle

Polynômes de Quantiles Avant d'introduire notre modèle statistique, nous devons introduire les quantiles polynomiaux. Nous dénoterons l'espace des fonctions quantiles par S_q. Pour réduire cet espace et obtenir un espace de travail plus "confortable", nous travaillerons

sur les quantiles polynomiaux du quantile $\mathcal{N}(0,1)$, notée q. Pour cela, il faut se restreindre aux polynômes P tels que $P \circ q$ est une fonction quantile.

Supposons que nous disposions d'une loi de probabilité $\mathcal{QP}(\boldsymbol{\theta}, L)$ sur cet espace. Nous pouvons alors définir $E_1, \dots E_n$ les éléments aléatoires quantiles suivant $\mathcal{QP}(\boldsymbol{\theta}, L)$. On note $E_k = \sum_{k=1}^{L} A_{i,k} q^i$.

Soit $\forall i \in \{1, 2, \dots, n\}$, $Q_i^Y : \Omega \to \mathrm{S}_q$ qui satisfait :

$$Q_i^Y(p) = \alpha + b_1 x_{q_i} + b_2 q_i^{cx}(p) + E_i(p), \text{ avec } b_2 > 0 \tag{2}$$

Puisque q^{cx} est une fonction quantile de la loi $N(0, \sigma^2)$ $q^{cx} = \sigma q$. Donc, $\alpha + b_1 x_{q_i} + b_2 q_i^{cx}(p) + E_i(p)$ peut être écrit comme $S \circ q$, avec S un polynôme.

$$Q_i^Y \quad = \quad \alpha + b_1 x_{q_i} + b_2 \sigma_i q + \sum_{k=0}^{L} A_{i,k} q^i = \sum_{k=0}^{L} B_{i,k} q^k \tag{3}$$

À l'instar d'une régression linéaire, nous montrons alors que pour tout i, $Q_i^Y \sim \mathcal{QP}(\boldsymbol{\theta_i})$. Ainsi, en réussissant à définir la loi $\mathcal{QP}(\boldsymbol{\theta_i})$, nous pourrons avoir une expression explicite de la vraisemblance.

Le cas $L = 1$ Pour le cas $L = 1$, nous pouvons expliciter une loi sur l'espace des polynômes de quantile et donc créer un modèle avec des solutions explicites.

$$S_{QP_1} := \{r = m + \sigma q\} = \left\{q :]0,1[\to \mathbb{R}; \quad \exists m \in \mathbb{R}, \exists s^2 > 0 \text{ tel que } q \text{ quantile de } \mathcal{N}\left(m, s^2\right)\right\}$$

Définissons Q un élément aléatoire sur cet espace de quantile. Soit $\mu \in \mathbb{R}$ et $\sigma^2, \beta > 0$,

$$\begin{array}{rcl} Q := Q_{\mu,\sigma^2,\beta} : \Omega & \to & S_{q_{\mathcal{N}}} \\ \omega & \mapsto & q_X := Q(\omega), \text{ quantile de X } \sim \mathcal{N}(u, v^2), \end{array} \tag{4}$$

avec $u = U(\omega)$ (respectivement $v = V(\omega)$) est une réalisation de la variable $U : \Omega \to \mathbb{R}$ (respectivement $V : \Omega \to \mathbb{R}^+$) et $U \sim \mathcal{N}\left(\mu, \sigma^2\right)$ (respectivement $V \sim \mathcal{E}xp\left(\beta^{-1}\right)$) avec U et V indépendants.

Q est une variable aléatoire dont on peut définir une fonction de densité. On note $Q\mathcal{N}(0, \sigma^2, \beta^{-1})$ la loi de Q.

Soient $q_1^X, q_2^X, \dots, q_n^X$ des fonctions quantiles déterministes prédictives de S_{QP_1} et $\epsilon_1, \dots, \epsilon_n$ des éléments aléatoires quantiles normaux distribués comme $Q\mathcal{N}(0, \sigma^2, \beta^{-1})$.

Soient $Q_1^Y, \dots, Q_n^Y$ i.i.d. d'un élément aléatoire quantile $Q^Y : \Omega \to S_Q$.

Le modèle statistique est défini comme suit :

$$\forall i \in 1, 2, \dots, n, \forall p \in]0,1[\quad Q_i^Y(p) = \alpha + b_1 x_{q_i} + b_2 q_i^{cx}(p) + \epsilon_i(p), \tag{5}$$

où les coefficients réels inconnus α, b_1 et b_2 seront estimés via une approche basée sur le maximum de vraisemblance.

La méthode décrite peut être modifiée avec différentes hypothèses, mais la condition importante est que les quantiles ont tous la même forme. Si ce prérequis est respecté, le procédé peut être adapté facilement. Dans le cas d'une fonction de quantile de forme normale, la distance de Wasserstein entre les distributions dépend uniquement de la moyenne et de la variance, ce qui facilite l'explication de la vraisemblance et permet d'obtenir des expressions explicites pour chaque estimateur.

Les estimateurs obtenus sont les suivants :

$$\widehat{b}_1 = \left(\sum_{i=1}^{n} y_{q_i} x_{q_i} - n\bar{y}\bar{x}\right)\left(\sum_{i=1}^{n} x_{q_i}^2 - n\bar{x}^2\right)^{-1}, \qquad \widehat{\beta} = \frac{1}{n-1}\sum_{i=1}^{n}\left(s_i - \min_{j\in\{1,\dots,n\}}\left(\frac{s_j}{v_j}\right)v_i\right),$$

$$\widehat{\alpha} = \bar{y} - \widehat{b}_1\bar{x}, \qquad \widehat{\sigma}^2 = (n-2)^{-1}\textstyle\sum_{i=1}^{n}(y_{q_i} - \widehat{\alpha} - \widehat{b}_1 x_{q_i})^2,$$

$$\widehat{b}_2 = \tfrac{n}{n-1}\min_{i\in\{1,\dots,n\}}\left(\tfrac{s_i}{v_i}\right) - \tfrac{\bar{s}}{(n-1)\bar{v}} \text{ avec } \bar{s} = n^{-1}\textstyle\sum_{i=1}^{n} s_i,$$

Qualité des estimations Nos résultats nous permettent d'expliciter l'intervalle de confiance pour chaque estimateur, ainsi qu'une distance de Cook et une quantité $PseudoR^2$.

2 Travail en cours

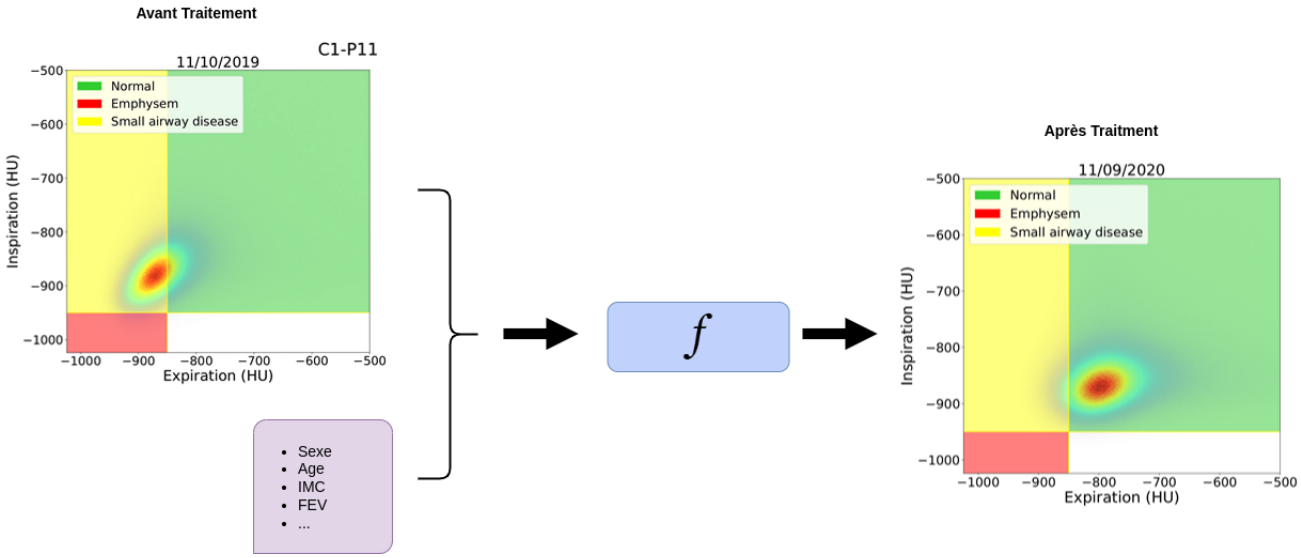

Figure 2: Parametric response map avant et après traitement

Le modèle précédent est utilisé sur les histogrammes de scanner CT en expiration, ce qui est pertinent car l'asthme est une maladie de l'expiration. Cependant, l'interaction entre

l'inspiration et l'expiration n'est pas exploitée. C'est pourquoi nous procédons à un recalage intra-patient de l'image en expiration sur celle en inspiration pour obtenir une correspondance voxel à voxel. Nous utilisons un recalage par B-spline[7].

Ainsi, nous pouvons calculer un histogramme 2D (Figure 2 [1]). Cette "Parametric response map" est introduite par Galbàn[8]. Cela crée un ensemble de données unique. Chaque patient est représenté par une distribution de support bidimensionnel et par des variables cliniques. La question est donc de savoir comment trouver un estimateur basé sur une distribution 2D et une covariable dans l'espace euclidien.

3 Bibliographie

1. Irpino, Verde (2015). **Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance.** *Advances in Data Analysis and Classification*, v.9, n.1,p.81–106.

2. Chen,Lin, Müller (2023). **Wasserstein Regression.** *Journal of the American Statistical Association* v.118, n.542, p.869–882.

3. Ghodrati L., Panaretos V. (2022) **Distribution-on-distribution regression via optimal transport maps.** *Biometrika*, v.109, p.957–974.

4. Coselmon, M.M., Balter, J.M., McShan, D.L. Kessler, M.L. (2004), **Mutual information based CT registration of the lung at exhale and inhale breathing states using thin-plate splines.** *Medical Physics* , 31: 2942-2948.

5. Heuberger, J., Geissbuhler, A., Muller, H. (2005). **Lung CT segmentation for image retrieval using the Insight Toolkit (ITK).** *Medical Imaging and Telemedicine*, 30.

6. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G. (2020). **Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem.** *European Radiology Experimental* , 4(1), 1-13.

7. Galbán CJ, Han MK, Boes JL, Chughtai KA, Meyer CR, Johnson TD, Galbán S, Rehemtulla A, Kazerooni EA, Martinez FJ, Ross BD. **Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression.** *Nature Medecine*. 2012 Nov;18(11):1711-5.

[1] Emphysem = Emphysème; Small airway disease = maladie des petites voies aériennes