Deriving statistical significance maps for support vector regression using medical imaging data

Bilwaj Gaonkar, Aristeidis Sotiras, Christos Davatzikos
Section of Biomedical Image Analysis, Center for Biomedical Image Computing and Analytics,
Department of Radiology, University of Pennsylvania, Philadelphia, USA
Bilwaj.Gaonkar@uphs.upenn.edu, Aristeidis.Sotiras@uphs.upenn.edu, Christos.Davatzikos@uphs.upenn.edu

***Abstract*—Regression analysis involves predicting a continuous variable using imaging data. The Support Vector Regression (SVR) algorithm has previously been used in addressing regression analysis in neuroimaging. However, identifying the regions of the image that the SVR uses to model the dependence of a target variable remains an open problem. It is an important issue when one wants to biologically interpret the meaning of a pattern that predicts the variable(s) of interest, and therefore to understand normal or pathological process. One possible approach to the identification of these regions is the use of permutation testing. Permutation testing involves 1) generation of a large set of 'null SVR models' using randomly permuted sets of target variables, and 2) comparison of the SVR model trained using the original labels to the set of null models. These permutation tests often require prohibitively long computational time. Recent work in support vector classification shows that it is possible to analytically approximate the results of permutation testing in medical image analysis. We propose an analogous approach to approximate permutation testing based analysis for support vector regression with medical imaging data. In this paper we present 1) the theory behind our approximation, and 2) experimental results using two real datasets.**

***Keywords*-Permutation testing; Support Vector Regression;**

I. Introduction

Regression analysis involves prediction of continuous clinical variables using medical images [1], [2], [3], [4], [5]. Multivariate pattern analysis (MVPA) techniques such as SVR directly address the image based regression paradigm. Most MVPA algorithms including SVR train a model by observing image data with known target variables. Target variables associated with a hitherto unseen test image can be estimated using the trained model.

The SVR algorithm offers predictions of continuous clinical variables from images. However, it provides no direct mechanism to assess which image regions are most significant in predicting the target variables. This question is relevant in clinical studies and is crucial to the clinicians who want to biologically understand imaging patterns and form new hypotheses. Traditionally, mass univariate Voxel Based Analysis (VBA) is used to find regions associated with continuous clinical variables. Such analysis associates a statistical significance test with every voxel in the image by regressing the voxel intensity directly with the target variable. While this provides ease of interpretability, such analysis (unlike MVPA) will miss multivariate associations in data. This motivates the need for a multivariate alternative to VBA that can interpret the model trained by an MVPA method such as a SVR. In the pattern classification paradigm, permutation tests using support vector classifiers (SVC) provide a multivariate alternative to VBA. We present an extension of this permutation testing procedure to the regression paradigm using SVRs.

A major problem with SVR/SVC based permutation testing applied to medical imaging data is the computational time and resources required for the actual implementation of these tests. However, recent work [6] showed that an analytical short cut exists for SVC based permutation testing that reduces the time and resource requirements by several orders of magnitude. This paper describes the theory behind such an analytical approximation that applies in case of SVR based permutation testing.

The remainder of the paper is organized as follows: in Section 2 the intuition behind permutation testing for regression analysis is presented. Following, we detail the theory behind the analytical approximation of permutation testing. Section 3 presents the experimental results on two brain imaging datasets. The paper concludes in Section 4 with a discussion.

II. Method

A. Support Vector Regression: Background

Let us first explain how the SVR [7] algorithm is used in the context of predicting continuous clinical variables from images.

1) Training: In order to train an SVR, we stack preprocessed training image data into a matrix $X \in R^{m \times p}$ whose rows $\mathbf{x}_i$ index individuals in the population, and columns index image voxels. A continuous target variable $y_i \in R$ is associated with every $\mathbf{x}_i$ in the training dataset. Then, the ϵ-SVR solves the following optimization problem:

$$\mathbf{w}^*, b^* = min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \quad (1)$$
$$\text{s.t. } \mathbf{w}^T\mathbf{x}_i + b - y_i \leq \epsilon,\ y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \epsilon, \forall i \in \{1, \ldots, m\}$$

The solution fits a tube of width ϵ to the data [8]. When the number of samples is higher than the number of samples

978-0-7695-5061-9/13 $26.00 © 2013 IEEE
DOI 10.1109/PRNI.2013.13

($n > p$), it is not always possible to find a tube of width ϵ that contains all the data. In the medical image analysis setting, the dimensionality is always much greater than the sample size ($p > n$). Hence, it is always possible to fit a p-dimensional ϵ-tube through all the datapoints.

2) Testing: The SVR model is encoded by the pair $\{\mathbf{w}^*, b^*\}$. For a new test subject whose vectorized image is represented by $\mathbf{x}_{test}$, the prediction y_{test} made by the SVR algorithm is $y_{test} = \mathbf{w}^{*T}\mathbf{x}_{test} + b^*$.

B. Permutation testing for support vector regression

The dimensionality of the model vector $\mathbf{w}^*$, trained by the SVR, is the number of voxels in the image. Thus, every component of the vector $\mathbf{w}^*$ can be mapped to a voxel in the image domain. This mapping associates an image with the SVR model. Henceforth, we call this image a $\mathbf{w}$-map. It would be desirable to directly use this image for making inferences about which regions are significantly involved in making predictions. However, these weights: 1) can be biased to be large by the simple scaling/translation operations on the underlying voxel intensities; 2) provide no measure of statistical significance of a specific feature/voxel in the image. Thus, a more rigorous method for interpreting the SVR model is required.

Permutation testing is one such method. The concept of permutation testing for SVRs in 2D space is illustrated in Fig. 1. In permutation testing, the target variables y_i are permuted randomly. For each random permutation, an SVR is used to compute $\mathbf{w}^*_{rp}$. After many thousands of permutations, we can generate an approximation to the null distribution of every component of $w_j \rightarrow D^j_{null}$ where $j \in \{1, \ldots, p\}$. Finally, the original labels are used to train $\mathbf{w}^*$. Comparing the components w^*_j with D^j_{null} gives us a p-value associated with every voxel. It is important to note that the null distribution at any voxel depends on the null distribution at all other voxels. This dependence is also true for the components of $\mathbf{w}$ themselves. Hence, each component-wise test is based on data from all image voxels and is not univariate in the VBA sense. Furthermore, this interdependence has the potential to alleviate multiple comparisons problems associated with VBA.

C. The analytical approximation of permutation testing

The main problem with the procedure detailed above is that it requires multiple runs of the SVR algorithm to approximate the underlying null distribution. This results in high computational demands. Massively parallel cluster computing is often used to perform these tests. In comparison to this, a typical run of VBA finishes in a few seconds on a typical computer. To close the gap, we propose an analytical approximation to SVR based permutation testing which runs in time comparable to VBA analysis while producing results that are comparable to empirical permutation testing.

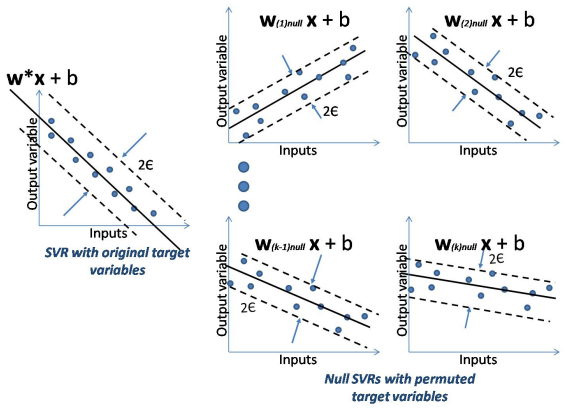


Figure 1: Concept of permutation testing in support vector regression. Comparison of $\mathbf{w}^*$ to the null distribution generated by $\{\mathbf{w}_{(1)null}, ..., \mathbf{w}_{(k)null}\}$ is used for inference.

The fundamental assumption behind the analytical approximation is that in high dimension, low sample size data, for most random permutations, the vast majority of the samples lie at the edges of the tube and are thus Support Vectors. This assumption is motivated by a similar assumption made in [6] with respect to support vector classification. Observations with real data confirm this phenomenon (Fig. 2). This assumption does not typically hold for the model trained with the actual targets. This is because there is enough structure in the data to learn from it. However, since most permutations are random the only way the algorithm can find a tube compatible with the entire dataset is by storing all of the data and its labels as support vectors. Under this assumption, for most permutations, the solution to (1) can be approximated by the solution to:

$$\mathbf{w}^*, b^* = min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \qquad (2)$$
$$\mathbf{w}^T\mathbf{x}_i + b - y_i = \epsilon \text{ OR } y_i - \mathbf{w}^T\mathbf{x}_i - b = \epsilon, \forall i \in \{1, \ldots, m\}$$

Now note that one of the two constraints has to hold for every sample for every permutation. For a permutation, a sample can either adhere to one constraint or another. Thus, for a particular permutation the optimization given by (2) can be solved using the Lagrange multiplier theory to yield exactly as it was done in [6].

$$\mathcal{L}(\mathbf{w}, b) = ||\mathbf{w}||_2^2 + \lambda^T((\mathbf{X}\mathbf{w} + \mathbf{J}b) - (\mathbf{y} + \mathbf{L}))$$

where the constraint vector $\mathbf{L} \in R^m$ has components $L_i = \pm\epsilon$ and and the vector $\mathbf{J} \in R^m$ with $J_i = +1$. Note that the constraint vector for one permutation will differ from that of the next based on which exact components are positive or negative. Setting $\frac{\partial}{\partial \mathbf{w}}\mathcal{L}(\mathbf{w}, b) = 0$ and $\frac{\partial}{\partial \lambda}\mathcal{L}(\mathbf{w}, b) = 0$ and solving for $\mathbf{w}$ yields:

$$\mathbf{w} = \mathbf{C}(\mathbf{y} + \mathbf{L}), \qquad (3)$$

where $\mathbf{C}$ denotes the matrix:

$$\mathbf{C} \doteq \mathbf{X^T}(\mathbf{XX^T})^{-1} + \mathbf{X^T}(\mathbf{XX^T})^{-1}\mathbf{J}(-\mathbf{J^T}(\mathbf{XX^T})^{-1}\mathbf{J})^{-1}\mathbf{J^T}(\mathbf{XX^T})^{-1}.$$

However, every permutation is associated with it's own vector $\mathbf{L}$. Over a large number of permutations, we can expect either constraint in (2) to hold with equal probability for each sample. Thus we may write: $P(L_i = +\epsilon) = 1/2$, $P(L_i = -\epsilon) = 1/2$. Note that (3) can also be written in its component form as:

$$w_j = \sum_{i=1}^{m} C_{ij}(y_i + L_i). \tag{4}$$

Because every element C_{ij} is fully determined by the data matrix $\mathbf{X}$, we can treat them as constants. By taking expectations on both sides of (4), we obtain:

$$E(w_j)\sum_{i=1}^{m} C_{ij}E(y_i + L_i) = E(y_i)\sum_{i=1}^{m} C_{ij}.$$

Note that $E(L_i) = 0$ and that $E(y_i)$ does not change with i allowing us to pull it outside the summation sign. To explicitly acknowledge this invariance, we henceforth denote $E(y_i)$ simply as $E(y)$. Similarly, the variance of w_j can be predicted by taking variances on both sides:

$$Var(w_j) = \sum_{i=1}^{m} C_{ij}^2(Var(y_i)+Var(L_i)) = (Var(y_i)+\epsilon^2)\sum_{i=1}^{m} C_{ij}^2.$$

Note again that the term $Var(y_i) + \epsilon^2$ is invariant with respect to i. Henceforth, we simply denote this term as $Var(y) + \epsilon^2$. Thus, we write:

$$E(w_j) = E(y)\sum_{i=1}^{m} C_{ij} \quad Var(w_j) = (Var(y) + \epsilon^2)\sum_{i=1}^{m} C_{ij}^2.$$

Regarding the distribution of w_j, it can be shown to be normal using the Lyapunov Central Limit Theorem (CLT). To see this, define $z_i^j = C_{ij}(y_i + L_i)$. The variable z_i^j is linearly dependent on $y_i + L_i$. We can infer the expectation and variance of z_i^j from y_j as:

$$E(z_i^j) = C_{ij}E(y) \quad Var(z_i^j) = C_{ij}^2(Var(y) + \epsilon^2).$$

Note that z_i^j are independent but not identically distributed, and w_j are linear combinations of z_j^i. Then, according to the Lyapunov CLT, w_j is distributed normally if:

$$\lim_{m\to\infty} \frac{1}{\left[\sqrt{\sum_{i=1}^{m} Var(z_i^j)}\right]^{2+\delta}} \sum_{k=1}^{m} E\big[\,|z_k^j - \mu_k|^{2+\delta}\,\big] = 0,\ \delta > 0. \tag{5}$$

For $\delta = 1$, we have:

$$E\big[\,|z_k^j - \mu_k|^{2+\delta}\,\big] = E\left[|C_{kj}y_k - C_{kj}E(y_k)|^{2+\delta}\right] = C_{kj}^3 E\left[|y_k - E(y_k)|^3\right].$$

Again we note that $E\left[|y_k - E(y_k)|^3\right]$ is independent of k and henceforth denote it simply as $E\left[|y - E(y)|^3\right]$. Then, we can write the limit in (5) as:

$$\lim_{m\to\infty} \frac{E\left[|y - E(y)|^3\right]\sum_{k=1}^{m} C_{kj}^3}{\left[\sqrt{(Var(y) + \epsilon^2)\sum_{i=1}^{m} C_{ij}^2}\right]^3} = K\sum_{k=1}^{m}\left(\sqrt{\lim_{m\to\infty}\frac{C_{kj}^2}{\sum_{i=1}^{m} C_{ij}^2}}\right)^3 = 0, \tag{6}$$

where K is a constant independent of the sample indices k and i, defined as: $K = \frac{E\left[|y-E(y)|^3\right]}{\left[\sqrt{(Var(y)+\epsilon^2)}\right]^3}$. Because (6) will tend to zero in the limit, we have normality of w_j by the Lyapunov CLT.

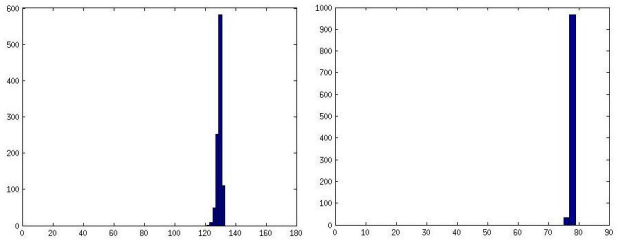


Figure 2: Most samples are support vectors for most permutations for SVRs. Human dataset (left) and mouse dataset (right).

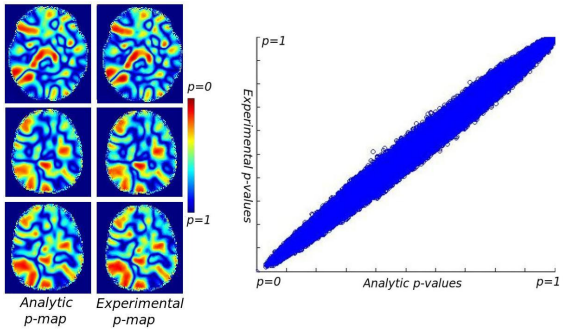
III. Experiments and Results

In order to validate the theory proposed above, we performed two experiments using imaging data. In the following, we discuss these experiments.

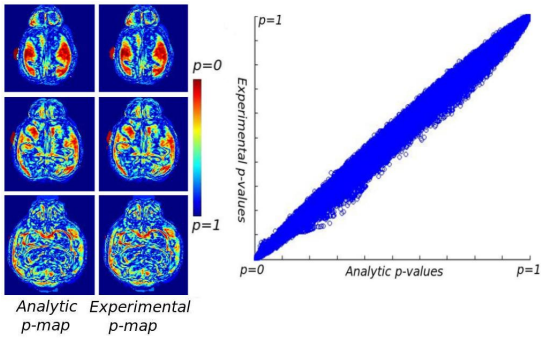
Human brain data: For this experiment, we used a dataset of 132 T1 images corresponding to normal subjects of age between 10 and 20 years. The experiment was done using Grey matter (GM), white matter (WM) and ventricular (CSF) tissue density maps (TDMs) that were generated after preprocessing of the raw images. TDMs convey information about the quantity of tissue present at each brain location in a common template space.

TDMs corresponding to the i^{th} subject were vectorized, and the vectors of all 3 tissue types were concatenated into the vector $\mathbf{x}_i \in R^{1\times 3q}$. The vectors of various samples were then stacked together to form the matrix $\mathbf{X}$. Permutation testing was performed using 1000 permutations of labels. The null distribution, obtained using permutation tests, was compared to the model that was trained with the original labels, to obtain an experimental p-map. Similarly, the analytical null distributions predicted using the theory presented in Section 2 were used to generate an analytical p-map. The two p-maps are compared in Fig. 3a.

It is easy to note by visual inspection that the significant map obtained with the analytical approximate methods agrees with the one obtained through permutation testing. The main difference is that the proposed method required



(a) Representative slices of analytic and experimental p-maps for grey matter TDMs (left) and scatter plot of corresponding analytic and experimental p-values (for all three tissue types) for human brain data.



(b) Representative slices of analytic and experimental p-maps (left) and scatter plot of corresponding analytic and experimental p-values for mouse brain data.

significantly less computational time than permutation testing to produce a result of equivalent quality. To gain a quantitative view of the level of agreement of the two solutions, Fig. 3a also shows the scatter plot between the experimental and analytic p-maps.

Developing mouse brain data: In this experiment, we applied the proposed method to the problem of white matter maturation in mouse brains. We used ex vivo acquired Diffusion Tensor images of a population of 79 inbred mice of C57BL/6J strain. The imaged mouse correspond to different postnatal stages, ranging from day 2 to day 80 [9]. Early developmental stages were sampled more densely because development is more emphasized during that period.

The images were deformably registered to a template image chosen from the age group of day 10 using DROID [10]. DTI-Studio [11] was used to estimate tensors from which, the Fractional Anisotropy was calculated resulting in images with dimension $300 \times 300 \times 200$.

Similarly to the previous experiment, we compare the experimental p-map with the analytic one. By visually comparing correspond slices from the two mpas, we note that the predicted values closely follow the actual ones Fig. 3b. We also observe distinctively low p-values in the cortex and the genu of corpus callossum. These areas have been previously reported exhibiting noteworthy maturation profiles [9]. The scatter plot suggests that analytic and experimental p-values agree.

IV. Discussion

In this paper, we have provided the theoretical framework for analytically approximating permutation tests using SVRs. We have also provided a limited validation of this framework using two real datasets.

References

[1] C. M. Stonnington, C. Chu, S. Klppel, C. R. J. Jr., J. Ashburner, and R. S. Frackowiak, "Predicting clinical scores from magnetic resonance scans in alzheimer's disease." *NeuroImage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[2] E. Formisano, F. De Martino, and G. Valente, "Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning," *Magnetic Resonance Imaging*, vol. 26, no. 7, pp. 921–934, sep 2008.

[3] J. Ashburner, "A Fast Diffeomorphic Image Registration Algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95–113, 2007.

[4] D. Zhang, D. Shen, and A. D. N. Initiative, "Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers," *PLoS ONE*, vol. 7, no. 3, p. e33182, 2012.

[5] Y. Wang, Y. Fan, P. Bhatt, and C. Davatzikos, "High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables." *NeuroImage*, vol. 50, no. 4, pp. 1519–1535, 2010.

[6] B. Gaonkar and C. Davatzikos, "Deriving statistical significance maps for svm based image classification and group comparisons," in *MICCAI*, 2012, pp. 723–730.

[7] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[8] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[9] R. Verma, S. Mori, D. Shen, P. Yarowsky, J. Zhang, and C. Davatzikos, "Spatiotemporal maturation patterns of murine brain quantified by diffusion tensor MRI and deformation-based morphometry." *PNAS*, vol. 102, no. 19, pp. 6978–83, 2005.

[10] M. Ingalhalikar, J. Yang, C. Davatzikos, and R. Verma, "Dti-droid: Diffusion tensor imaging-deformable registration using orientation and intensity descriptors," *International Journal of Imaging Systems and Technology*, vol. 20, no. 2, pp. 99–107, 2010.

[11] H. Jiang, P. C. van Zijl, J. Kim, G. D. Pearlson, and S. Mori, "Dtistudio: Resource program for diffusion tensor computation and fiber bundle tracking," *Computer Methods and Programs in Biomedicine*, vol. 81, no. 2, pp. 106–116, 2006.