NOTES ON QUADRATIC PROGRAMMING FEATURE SELECTION

NIVAS

1. Determining The Significance of Features

The computed features were motivated by discussions with the subject matter experts, with the view of casting a meaningful but wide net to capture attributes of an epileptic seizure. However, this doesn't strictly preclude the possibility that, with respect to the phenomenon of epileptic seizure, certain features may be redundant, or low in predictive importance. Therefore, we quantify the predictive significance of the features in a natural but effective way, and score the features to maximize predictive importance and minimize redundancy using the method in [Rod Luj], which we summarize here.

1.1. Measuring Redundancy.

Our notion of redundancy arises naturally from the interpretation of brain activity as a stochastic process, whence the usual notion of linear dependence is replaced with the notion of statistical correlation. Specifically, suppose the data matrix, D, spanning t epochs, is $t \times n$, with n features, $z_i, 1 \leq i \leq n, z_i \in \mathbb{R}^t$. We define, the correlation matrix, $Q \in \mathbb{R}^{n,n}$, elementwise, where $Q(i,j)$ is the pearson correlation coefficient between the feature vectors z_i and z_j:

$$Q(i,j) = \frac{z_i^T z_j}{||z_i||\,||z_j||}.$$

The quadratic form $x^T Q x$ thus has the natural interpretation of yielding the sample-covariance of a compound feature, with coefficients contained in x, which is the notion of redundancy we wish to minimize.

1.2. Measuring Predictive Importance.

We first recall that the activity of the brain at time t is completely captured by d^t. We define the predictive importance, f_i, of a feature, z_i, as the r.m.s. influence of z_i^t on $z_j^{t+1}, 1 \leq j \leq n$, measured by the coefficients in the forecast operator corresponding to z_i. Formally, let $\Psi \in \mathbb{R}^{n,n+1}$ be the forecast operator. Then our best prediction of d^{t+1} is $\tilde{d}^{t+1}$ where

$$\tilde{d}^{t+1} = d^t \Psi^T.$$

The influence, $p_i(j)$, of feature z_i on z_j, contained in $p_i \in \mathbb{R}^n$, may be determined by predicting via Ψ using its indicator vector, e_i:

$$p_i = e_i \Psi^T,$$

whence the r.m.s. influence, f_i, of z_i is simply

$$f_i = ||p_i|| = ||\Psi_i||\,,$$

the column-norm of the forecast operator corresponding to column i. We define $f \in \mathbb{R}^n$ such that $f_i = ||\Psi_i||$, as the predictive importance vector.

1.3. Optimizing Redundancy and Predictive Importance.

We note that we may indicate features that maximize predictive importance and minimize redundancy by solving

$$\bar{x} = \arg\min x^T Q x - f^T x; \quad x \geq 0, \quad \sum_i x_i = 1,$$

where the constraints arise from forcing the resulting vector to be a distribution, from which we may omit an appropriately sized tail, should we choose to do so. However, the magnitudes of entries in Q, f may be vastly different for very similar data - for example, a data matrix aD would have the same correlation matrix as D, but its predictive importance vector would be a times as large. To make the objective function scale invariant, we normalize f to obtain

$$\hat{f} = f / ||f||_\infty\,.$$

Finally, to effect a meaningful trade-off between minimizing redundancy and maximizing predictive importance, we take a convex combination of the objective functions in both optimization problems:

$$q(x) = (1-\alpha) x^T Q x - \alpha f^T x,$$

where α is chosen, as in [rodLuj] as

$$\alpha = \frac{\sum_{i,j} Q(i,j)/n^2}{\sum_{i,j} Q(i,j)/n^2 + \sum_k f_k/n}.$$

We solve the resulting optimization problem:

$$x^* = \arg\min q(x); \quad s.t. x \geq 0, \quad \sum_i x_i = 1$$

to obtain an importance-distribution over the features.