Supplementary Material

1. Directional Derivatives

Start with a response vector $\mathbf{Y} = [y_1, \ \ldots, \ y_n]'$ and a predictor matrix $\mathbf{X}$, and suppose that there is a function $y_i = g(\mathbf{x}_i)$, where $\mathbf{x} = [x_1, \ \ldots, \ x_p]$ is a row-vector of $\mathbf{X}$. The gradient field $\nabla g(\cdot)$ is the set of gradient vectors $\nabla g(\mathbf{x})' = [\partial g(\mathbf{x})/\partial(x_1), \ \ldots, \ \partial g(\mathbf{x})/\partial(x_p)]$ at all points. At point $\mathbf{x}_i$, a vector of directional derivatives $D_v g(\mathbf{x})$ of function $g(\mathbf{x})$ in the direction of d unit vectors $\mathbf{V}$ is given by:

$$D_v g(\mathbf{x}_i) = \nabla g(\mathbf{x}_i)'\mathbf{V} \tag{S1}$$

where $\mathbf{V}$ is a $p \times d$ matrix. For example, the directional derivatives $\nabla g(\mathbf{x}_i)'\mathbf{I}_p$ are equal to the partial derivatives of $g(\mathbf{x}_i)$. Further, if $g(\mathbf{x}_i) = f(\mathbf{x}_i\mathbf{V})$ then $\nabla g(\mathbf{x}_i) = \mathbf{V}\nabla f(\mathbf{x}_i\mathbf{V})$ and

$$D_v g(\mathbf{x}_i) = \nabla g(\mathbf{x}_i)'\mathbf{V} = \nabla f(\mathbf{x}_i\mathbf{V})'\mathbf{V}'\mathbf{V} = \nabla f(\mathbf{x}_i\mathbf{V})' \tag{S2}$$

where $\nabla f(\mathbf{r})' = [\partial f(\mathbf{r})/\partial(r_1), \ \ldots, \ \partial f(\mathbf{r})/\partial(r_d)]$, and $\mathbf{r}$ is a row vector of $\mathbf{R} = \mathbf{X}\mathbf{V}$. It follows that the directions $\mathbf{V}$ in which $D_v g(\cdot)$ is maximized can be found by the eigenvalue decomposition of the average of the outer product of the gradient vectors at each point:

$$n^{-1} \sum_{i=1}^{n} (\nabla g(\mathbf{x}_i) \nabla g(\mathbf{x}_i)') = \mathbf{V}\mathbf{L}\mathbf{V}'. \tag{S3}$$

The diagonal elements of $\mathbf{L}$ are equal to the squared directional derivatives in the direction of each unit vector of $\mathbf{V}$. Note that as shown in Eq. S2 maximizing $D_v g(\mathbf{x}_i)$ is the same as maximizing $\nabla f(\mathbf{x}_i\mathbf{V})$.

Note that if $\mathbf{Y}$ is a matrix, then the only changes to the above are that $\nabla g(\mathbf{x})$, $\nabla f(\mathbf{r})$ and $D_v g(\mathbf{x})$ are also matrices.

2. Example of a Feature Space

In this example, let $\mathbf{X}$ be a predictor matrix with two variables $\mathbf{x}_i = [x_1\ x_2]_i$, where $\mathbf{x}$ is a row-vector of $\mathbf{X}$, and $i = (1,\ \ldots,\ n)$. The mapping functions $\Phi(\cdot)$ map from X-space to some feature space e.g.

$$\mathbf{x} \mapsto \Phi(\mathbf{x}) \tag{S4a}$$

$$[x_1\ x_2] \mapsto \begin{bmatrix} x_1 & x_2 & x_1x_2 & x_1^2 & x_2^2 & c \end{bmatrix} \tag{S4b}$$

so the feature space in this example is a 6-dimensional polynomial feature space. Now let y be a response variable that is a linear combination of the feature space $\Phi(\mathbf{x})$, then we can write $g(\mathbf{x}) = \Phi(\mathbf{x})\mathbf{b}$ or

$$y = b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2 + b_6c. \tag{S5}$$

In this example it is possible to create a simple 6-dimensional feature space from a 2-dimensional X-space, because $\mathbf{X}$ is a low-dimensional. But if $\mathbf{X}$ contains a lot of predictors, then it becomes difficult to explicitly map the predictors to a feature space that e.g. contains all the polynomial terms, including all the interaction terms.

3. Dual forms for the linear model and the nonlinear model

Start with the equation:

$$(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_p)\mathbf{X}' = \mathbf{X}'\mathbf{X}\mathbf{X}' + \lambda\mathbf{X}' = \mathbf{X}'(\mathbf{X}\mathbf{X}'+\lambda\mathbf{I}_n). \tag{S6}$$

Now left multiply each part of the above equation by $(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_p)^{-1}$, and then right multiply each part by $(\mathbf{X}\mathbf{X}'+\lambda\mathbf{I}_n)^{-1}\mathbf{Y}$, resulting in the identity:

$$\mathbf{X}'(\mathbf{X}\mathbf{X}'+\lambda\mathbf{I}_n)^{-1}\mathbf{Y} = (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{Y}. \tag{S7}$$

For the linear regression model $\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{B}$, the RHS of Eq. S7 is the conventional form of the regression matrix $\mathbf{B}$, and the LHS is the kernel form of $\mathbf{B}$. So in kernel form:

$$\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{X}'(\mathbf{X}\mathbf{X}'+\lambda\mathbf{I}_n)^{-1}\mathbf{Y}, \text{ or} \tag{S8a}$$

$$\widehat{\mathbf{Y}} = \mathbf{G}_X(\mathbf{G}_X+\lambda\mathbf{I}_n)^{-1}\mathbf{Y} \tag{S8b}$$

Now consider regression with respect to the feature space $\Phi(\mathbf{X})$:

$$\widehat{\mathbf{Y}} = \Phi(\mathbf{X})\mathbf{B} \tag{S9}$$

where $\mathbf{B} = [\Phi(\mathbf{X})'\Phi(\mathbf{X})+\lambda\mathbf{I}_p]^{-1}\Phi(\mathbf{X})'\mathbf{Y}$, and p is now the dimension of the feature space. In Eq. S6-S8 replace $\mathbf{X}$ by $\Phi(\mathbf{X})$, it should now be apparent that:

$$\widehat{\mathbf{Y}} = \Phi(\mathbf{X})\mathbf{B} = \mathbf{G}_X(\mathbf{G}_X+\lambda\mathbf{I}_n)^{-1}\mathbf{Y} \tag{S10}$$

where $\mathbf{G}_X = \Phi(\mathbf{X})\Phi(\mathbf{X})'$.

Lastly, if there exists a factorizable kernel $K(\cdot,\cdot)$ such that $\mathbf{G}_{ij} = K(\mathbf{x}_i,\mathbf{x}_j) = \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)'$, where $\mathbf{G}_{ij}$ are the matrix elements of $\mathbf{G}_X$, then:

$$\widehat{\mathbf{y}}_i = K(\mathbf{x}_i,\cdot)(\mathbf{G}_X+\lambda\mathbf{I}_n)^{-1}\mathbf{Y} = K(\mathbf{x}_i,\cdot)\mathbf{H} \tag{S11}$$

where $\widehat{\mathbf{y}}_i$ is a row-vector of $\widehat{\mathbf{Y}}$, $\mathbf{H} = (\mathbf{G}_X+\lambda\mathbf{I}_n)^{-1}\mathbf{Y}$, and $K(\mathbf{x}_i,\cdot) = [K(\mathbf{x}_i,\mathbf{x}_1),\ \ldots,\ K(\mathbf{x}_i,\mathbf{x}_n)]$ i.e $K(\mathbf{x}_i,\cdot)$ is a row-vector of $\mathbf{G}_X$.

Note for $\mathbf{Y} = \mathbf{XB}$, $K(\mathbf{x}_i, \cdot) = [\mathbf{x}_i\mathbf{x}_1', \ \ldots, \ \mathbf{x}_i\mathbf{x}_n']$. Also note in Eq. S9–S11 that $\widehat{\mathbf{Y}}$ has been expressed as a linear combination of $\Phi(\mathbf{X})$ and $K(\cdot,\cdot)$.

4. gKCCA and symmetry

Linear CCA is based on a eigen-decomposition which allows $\mathbf{X}$ and $\mathbf{Y}$ to be swapped i.e. the canonical correlations are the same for CCA($\mathbf{X},\mathbf{Y}$), and CCA($\mathbf{Y},\mathbf{X}$). Note that gKCCA makes use of the same eigendecomposition as linear CCA (Section 2d), but between $\mathbf{Y}$ and a nonlinear augmentation of $\mathbf{XW}$ i.e. CCA($\phi(\mathbf{XW}),\mathbf{Y}$). So in the CCA step of gKCCA, the canonical correlations are the same for CCA($\phi(\mathbf{XW}),\mathbf{Y}$) and CCA($\mathbf{Y},\phi(\mathbf{XW})$).

However, there is a more general question here about symmetry with respect to $\mathbf{X}$ and $\mathbf{Y}$ in gKCCA. For example, are there functions $f_j(\cdot)$ which satisfy?:

$$\mathbf{yA}_j = f_j(\mathbf{xW}_j) \tag{S12a}$$

$$\mathbf{xW}_j = f_j^{-1}(\mathbf{yA}_j) \tag{S12b}$$

where e.g. $\mathbf{y}$ is a row-vector of $\mathbf{Y}$, and $\mathbf{A}_j$ is the j^{th} column-vector of the matrix $\mathbf{A}$. One issue here is that the derivatives of $f_j(\cdot)$ and $f_j^{-1}(\cdot)$ are generally not equal because:

$$\frac{d(f^{-1}(y))}{dy} = \left(\frac{df(x)}{dx}\right)^{-1} \tag{S13}$$

where $y = f(x)$. So if the functions $f_j(\cdot)$ were strictly invertible, and $\mathbf{X}$ and $\mathbf{Y}$ were swapped, we may find the same subspaces $\mathbf{XW}$ and $\mathbf{YA}$ but with the components in a different order. Of course, the components could be reordered by further calculation.

Lastly, since there are not many functions $f_j(\cdot)$ which make Eq. S12 true, there will typically be an asymmetry between the roles of $\mathbf{X}$ and $\mathbf{Y}$ in gKCCA.