MTTTS16 Learning from Multiple Sources

5 ECTS credits

Autumn 2019, University of Tampere
Lecturer: Jaakko Peltonen

Lecture 1: Canonical correlation analysis

On this lecture:

- Canonical correlation analysis

Part 1: Canonical Correlation Analysis

Motivation for CCA

- Recap: **correlation** is one of the most basic statistical measures of dependency between two real-valued scalar variables.

- Covariance between two variables: cov(x,y) = E[(x-E[x])(y-E[y])]
  Depends on scales of the variables.

- **Correlation coefficient** (Pearson's product-moment correlation coefficient) between two variables:

$$corr(x,y) = \frac{cov(x,y)}{stdev(x)\,stdev(y)} = \frac{E[(x-E[x])(y-E[y])]}{(E[(x-E[x])^2])^{1/2}(E[(y-E[y])^2])^{1/2}}$$

- For zero-mean variables:

$$corr(x,y) = \frac{E[xy]}{(E[x^2]E[y^2])^{1/2}}$$

Terms can be estimated from samples $\{(x^i,y^i)\}_{i=1,\ldots,N}$

$$\hat{E}_{ML}[xy] = \frac{1}{N}\sum_{i=1}^{N} x^i y^i$$

Motivation for CCA

- If two variables have zero correlation (uncorrelated), they are **linearly independent**, otherwise they are **linearly dependent**.

- If two variables are linearly dependent, a linear predictor $y = ax+b$ can predict one variable from the other with some accuracy.
- Example: People's height and weight are positively correlated; poverty rate and education level might be negatively correlated

- If two variables have zero correlation, a linear predictor cannot predict one variable from the other.

- Two statistically independent variables have zero correlation.

Motivation for CCA

Caveats:

- Two variables with zero correlation can be dependent in nonlinear ways. If two variables have nonzero correlation, they may also have nonlinear dependency that correlation is unable to measure. Nonlinear dependencies can be exploited by a nonlinear predictor.
- Correlation does not imply causation: if x is correlated with y, then x might cause y, y might cause x, or x and y might be caused by other variables, some of which affect both x and y.

Motivation for CCA

- **More complicated situation 1**: What about when we have several scalar variables x_1, x_2,...,x_K, and we are not sure how they are related to y?

- It can happen that **individual** variables x_i are not strongly correlated with y, but there is a linear **combination** of variables that is strongly correlated to y.

- Example: total time of multi-hop bus journeys depends on the sum of journey lengths. The length of the first journey is not enough by itself to predict the total time.

- Idea: **which** linear combination of x_1, x_2,...,x_K is most strongly correlated with y? Can we find the best linear combination $a_1x_1 + a_2x_2 + ..., + a_Kx_K$, that is, the best weights a_1, ...,a_K ?

- Yes: this turns out to be the same as **ordinary least-squares linear regression**!

Canonical Correlation Analysis

- **More complicated situation 2:** What about when we have several scalar variables x_1, x_2,...,x_K, and also several scalar variables y_1, y_2,...,y_L, ? How can we now find what relationships exist between the x variables and the y variables?
- **Canonical correlation analysis (CCA)** is a method of correlating two multidimensional variables **x** = [x_1, x_2,...,x_K] and **y** = [y_1, y_2,...,y_L]
- Proposed by H. Hotelling. (*Hotelling, H. Relations between two sets of variates. Biometrika, vol. 28, pages 312–377, 1936.*)
- Idea: not all of **x** needs to be correlated with all of **y** or vice versa, as long as there is something correlated between **x** and **y**.
- Problem definition: given two sets of variables, find **basis vectors** (linear transformations), one for each set of variables, so that the correlations between the linear projections of the variables onto the basis vectors are mutually maximized (the projected coordinates are maximally correlated).

Canonical Correlation Analysis

- For **x**, find a projection $w_{x,1}x_1 + w_{x,2}x_2 + \ldots + w_{x,K}x_K$ where $\mathbf{w_x} = [\, w_{x,1}, w_{x,2}, \ldots, w_{x,K}\,]$ is the projection basis.
- For **y**, find a projection $w_{y,1}y_1 + w_{y,2}y_2 + \ldots + w_{y,L}y_L$ where $\mathbf{w_y} = [\, w_{y,1}, w_{y,2}, \ldots, w_{y,L}\,]$ is the projection basis.
- Find the projection bases by maximizing the correlation between the projections: maximize

$$corr\left(\mathbf{w}_x^T\mathbf{x}, \mathbf{w}_y^T\mathbf{y}\right) = \frac{E\left[\mathbf{w}_x^T\mathbf{x}\,\mathbf{w}_y^T\mathbf{y}\right]}{\left(E\left[\left(\mathbf{w}_x^T\mathbf{x}\right)^2\right]E\left[\left(\mathbf{w}_y^T\mathbf{y}\right)^2\right]\right)^{1/2}}$$

with respect to $\mathbf{w_x}$ and $\mathbf{w_y}$.

This definition assumes x and y are zero-mean, otherwise substract the means as in the original correlation definition.

- For a finite data set: maximize the sample correlation

$$\hat{corr}\left(\mathbf{w}_x^T\mathbf{x}, \mathbf{w}_y^T\mathbf{y}\right) = \frac{\hat{E}_{ML}\left[\mathbf{w}_x^T\mathbf{x}\,\mathbf{w}_y^T\mathbf{y}\right]}{\left(\hat{E}_{ML}\left[\left(\mathbf{w}_x^T\mathbf{x}\right)^2\right]\hat{E}_{ML}\left[\left(\mathbf{w}_y^T\mathbf{y}\right)^2\right]\right)^{1/2}}$$

Same definition as before

$$\hat{E}_{ML}[x\,y] = \frac{1}{N}\sum_{i=1}^{N} x^i y^i$$

Canonical Correlation Analysis

- How can we find the maximal correlation with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$?
- It turns out this can be rewritten as a **generalized matrix eigenvalue problem. Let's show how that is done.**
- First, rewrite the correlation to show it depends on covariance matrices:

$$\hat{corr}\left(\mathbf{w}_x^T\mathbf{x}, \mathbf{w}_y^T\mathbf{y}\right) = \frac{\hat{E}_{ML}\left[\mathbf{w}_x^T\mathbf{x}\,\mathbf{w}_y^T\mathbf{y}\right]}{\left(\hat{E}_{ML}\left[\left(\mathbf{w}_x^T\mathbf{x}\right)^2\right]\hat{E}_{ML}\left[\left(\mathbf{w}_y^T\mathbf{y}\right)^2\right]\right)^{1/2}}$$

$$= \frac{\hat{E}_{ML}\left[\mathbf{w}_x^T\mathbf{x}\,\mathbf{y}^T\mathbf{w}_y\right]}{\left(\hat{E}_{ML}\left[\mathbf{w}_x^T\mathbf{x}\,\mathbf{x}^T\mathbf{w}_x\right]\hat{E}_{ML}\left[\mathbf{w}_y^T\mathbf{y}\,\mathbf{y}^T\mathbf{w}_y\right]\right)^{1/2}}$$

$$= \frac{\mathbf{w}_x^T\hat{E}_{ML}\left[\mathbf{x}\,\mathbf{y}^T\right]\mathbf{w}_y}{\left(\mathbf{w}_x^T\hat{E}_{ML}\left[\mathbf{x}\,\mathbf{x}^T\right]\mathbf{w}_x\,\mathbf{w}_y^T\hat{E}_{ML}\left[\mathbf{y}\,\mathbf{y}^T\right]\mathbf{w}_y\right)^{1/2}}$$

$$= \frac{\mathbf{w}_x^T\hat{C}_{x,y}\mathbf{w}_y}{\left(\mathbf{w}_x^T\hat{C}_x\mathbf{w}_x\,\mathbf{w}_y^T\hat{C}_y\mathbf{w}_y\right)^{1/2}}$$

Sample estimate of the covariance matrix

$$\hat{C}_{x,y} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^i\mathbf{y}^{iT}$$

$$\hat{C}_x = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^i\mathbf{x}^{iT}$$

$$\hat{C} = \begin{bmatrix} \hat{C}_x & \hat{C}_{x,y} \\ \hat{C}_{y,x} & \hat{C}_y \end{bmatrix}$$

Canonical Correlation Analysis

Proof continues...

- Notice that the scales of $\mathbf{w}_x$ and $\mathbf{w}_y$ do not affect the correlation, so we can set any requirement that can be satisfied by changing the scales alone. For example, we can require $\boldsymbol{w}_x^T \hat{C}_x \boldsymbol{w}_x = 1$, $\boldsymbol{w}_y^T \hat{C}_y \boldsymbol{w}_y = 1$

- Then the optimization becomes a constrained optimization problem:

$$max_{\boldsymbol{w}_x, \boldsymbol{w}_y} [\boldsymbol{w}_x^T \hat{C}_{x,y} \boldsymbol{w}_y] \quad \text{such that} \quad \boldsymbol{w}_x^T \hat{C}_x \boldsymbol{w}_x = 1 , \ \boldsymbol{w}_y^T \hat{C}_y \boldsymbol{w}_y = 1$$

- Constrained optimization problems can be solved by the method of **Lagrange multipliers**, that is we maximize the Lagrangian

$$max_{\boldsymbol{w}_x, \boldsymbol{w}_y} L(\boldsymbol{w}_x, \boldsymbol{w}_y, \lambda_x, \lambda_y)$$

$$L(\boldsymbol{w}_x, \boldsymbol{w}_y, \lambda_x, \lambda_y) = \left(\boldsymbol{w}_x^T \hat{C}_{x,y} \boldsymbol{w}_y - \frac{\lambda_x}{2} (\boldsymbol{w}_x^T \hat{C}_x \boldsymbol{w}_x - 1) - \frac{\lambda_y}{2} (\boldsymbol{w}_y^T \hat{C}_y \boldsymbol{w}_y - 1) \right)$$

Canonical Correlation Analysis

Proof continues...

• At the optimum, the derivative of the Lagrangian must be zero with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$. This gives

$$\frac{dL}{d\mathbf{w}_x} = \hat{C}_{x,y}\mathbf{w}_y - \lambda_x \hat{C}_x \mathbf{w}_x = 0$$

$$\frac{dL}{d\mathbf{w}_y} = \hat{C}_{y,x}\mathbf{w}_x - \lambda_y \hat{C}_y \mathbf{w}_y = 0$$

$$\mathbf{w}_x^T \hat{C}_x \mathbf{w}_x = 1 \quad \mathbf{w}_y^T \hat{C}_y \mathbf{w}_y = 1$$

substract the $\mathbf{w}_x$ times the second equation from $\mathbf{w}_y$ times the first:

$$\begin{aligned}
&\mathbf{w}_x^T \hat{C}_{x,y}\mathbf{w}_y - \lambda_x \mathbf{w}_x^T \hat{C}_x \mathbf{w}_x - \mathbf{w}_y^T \hat{C}_{y,x}\mathbf{w}_x + \lambda_y \mathbf{w}_y^T \hat{C}_y \mathbf{w}_y \\
&= -\lambda_x \mathbf{w}_x^T \hat{C}_x \mathbf{w}_x + \lambda_y \mathbf{w}_y^T \hat{C}_y \mathbf{w}_y \\
&= -\lambda_x + \lambda_y \\
&= 0
\end{aligned}$$

because we required

$$\mathbf{w}_x^T \hat{C}_x \mathbf{w}_x = 1$$

$$\mathbf{w}_y^T \hat{C}_y \mathbf{w}_y = 1$$

Canonical Correlation Analysis

Proof continues...

- Therefore the Lagrange multipliers are equal, and

$$\frac{d\,L}{d\,\boldsymbol{w}_y} = \hat{C}_{y,x}\boldsymbol{w}_x - \lambda \hat{C}_y \boldsymbol{w}_y = 0 \rightarrow \boldsymbol{w}_y = (1/\lambda)\hat{C}_y^{-1}\hat{C}_{y,x}\boldsymbol{w}_x$$

- Inserting that into the other derivative equation gives

$$\hat{C}_{x,y}\hat{C}_y^{-1}\hat{C}_{y,x}\boldsymbol{w}_x = \lambda^2 \hat{C}_x \boldsymbol{w}_x$$

- This is a generalized eigenvalue equation which we can solve to get $\mathbf{w}_x$, and the previous equation then gives $\mathbf{w}_y$ from $\mathbf{w}_x$.
- If the covariance of **x** is invertible, this can also be written as a standard eigenproblem with some additional steps

Canonical Correlation Analysis

Generalizations of CCA:

• Several possible generalizations of CCA to more than two sets of variables have been proposed. One is (Hardoon, Szedmak, Taylor):

$$min_{\mathbf{W}^{(1)},\ldots,\mathbf{W}^{(M)}} \sum_{m,n=1,m\neq n}^{M} \left\| \mathbf{W}^{(m)T}\mathbf{x}^{(m)} - \mathbf{W}^{(n)T}\mathbf{x}^{(n)} \right\|_F$$

$$\mathbf{W}^{(m)T}\hat{C}_{mm}\mathbf{W}^{(m)} = I \qquad \mathbf{W}^{(m)T}\hat{C}_{mn}\mathbf{W}^{(n)} = 0$$

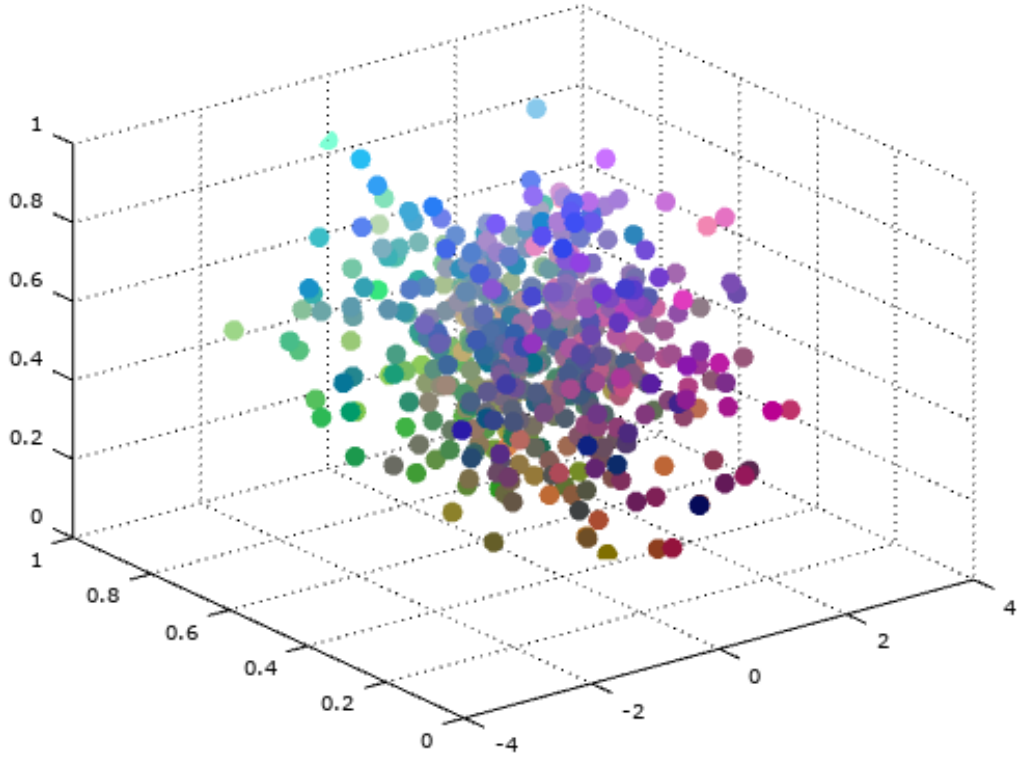
• CCA can be shown to be a generative model of data (next on this lectures).

• CCA can be generalized to analyze nonlinear correlations through a kernel mapping (later in the lectures)

• CCA can be incorporated with sparsity and other prior desires for the projection bases

Canonical Correlation Analysis, example

- In this example **x** and **y** are both 3-dimensional.
- **x** = [x_1 x_2 x_3] is randomly distributed according to a three-dimensional isotropic (spherical) Gaussian
- **y** = [y_1 y_2 y_3] = [red green blue], where "red" is proportional to $x_1+x_2+x_3$ and green and blue are randomly normally distributed.



In this picture **x** is shown as 3D spatial coordinates, and **y** is shown as red-green-blue (RGB) color components

Matlab/octave code:

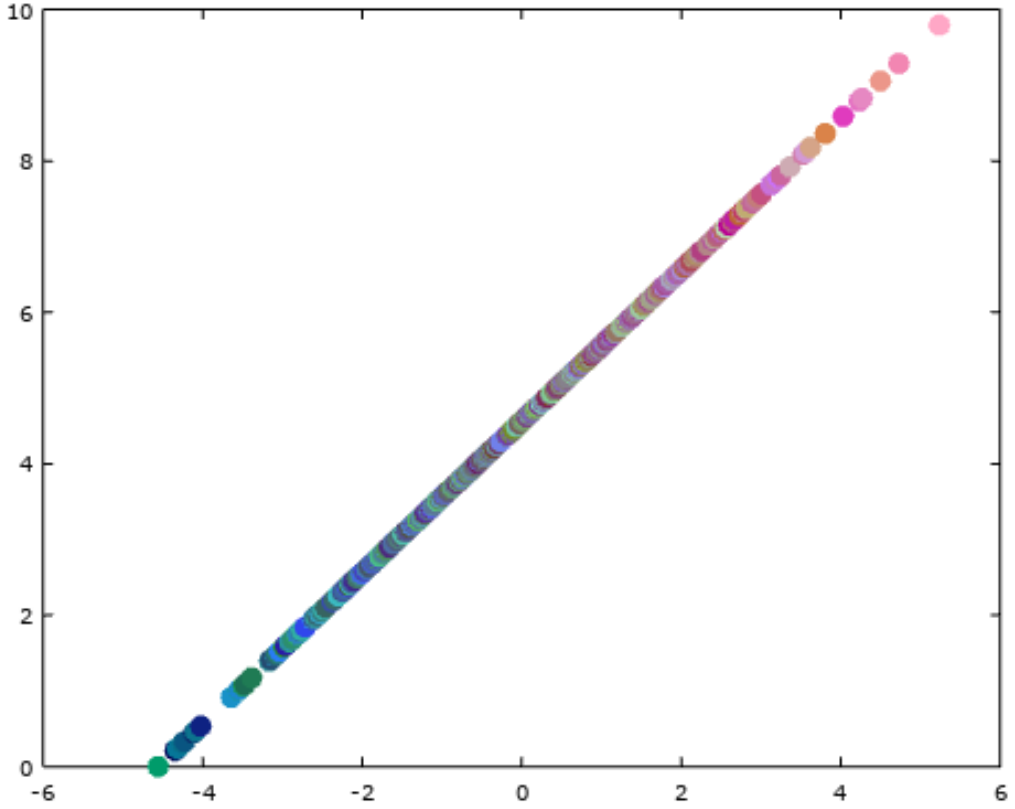
```
x=randn(500,3);
y=randn(500,3);
y(:,1)=sum(x,2);
for k=1:3,
y(:,k)=(y(:,k)-min(y(:,k)))/(max(y(:,k))-
min(y(:,k))); end;

scatter3(x(:,1),x(:,2),x(:,3), ...
   10*ones(size(x,1),1),y,'filled');
```

Canonical Correlation Analysis, example

- Let's compute and solve the eigenvalue equation.

```
C = cov([x y]);
Cx=C(1:3,1:3);
Cy=C(4:6,4:6);
Cxy=C(1:3,4:6);
A=Cxy*inv(Cy)*Cxy';
B=Cx;
[Wx,D]=eig(A,B);  % A*Wx=B*Wx*D
Wy=inv(Cy)*Cxy'*Wx*inv(D);
```




```
Wx =
   1.00000   0.82944  -0.70922
   1.00000   0.30068   1.00000
   1.00000  -1.00000  -0.40489

D =
   1.0000e+00         0         0
            0   1.6523e-04         0
            0         0   5.8796e-03

Wy =
   9.7948e+00   2.2388e+01  -1.5861e+00
  -3.0531e-16  -6.2843e+02  -4.2152e+01
   5.2736e-16  -2.2199e+02   8.7614e+01
```

The 1st column of Wx projects **x** to x1+x2+x3 (corresponds to largest eigenvalue in D)

1st column of Wy projects **y** to y1

Resulting projections match perfectly (correlation 1)

```
xp=x*Wx(:,1);
yp=y*Wy(:,1);

figure;
scatter(xp,yp,10*ones(size(x,1),1),y,'filled');
```

Canonical Correlation Analysis, example

Connection to least-squares regression:

- If y is one-dimensional (scalar):
  let $\mathbf{X}_{all} = [\mathbf{x}^1\ \mathbf{x}^2\ ...\ \mathbf{x}^N]$ and $\mathbf{y}_{all} = [y^1\ y^2\ ...\ y^N]^T$, then we have

$$\hat{C}_{x,y} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}^i y^i = \frac{1}{N}\boldsymbol{X}_{all}\boldsymbol{y}_{all}$$

$$\hat{C}_x = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}^i \boldsymbol{x}^{iT} = \frac{1}{N}\boldsymbol{X}_{all}\boldsymbol{X}_{all}^T \qquad \hat{C}_y = \frac{1}{N}\sum_{i=1}^{N} y^i y^i = \frac{1}{N}\boldsymbol{y}_{all}^T\boldsymbol{y}_{all}$$

- The ordinary (least-squared) linear regression solution is
  $\boldsymbol{w}_x = \left(\boldsymbol{X}_{all}\boldsymbol{X}_{all}^T\right)^{-1}\boldsymbol{X}_{all}\boldsymbol{y}_{all}$, we show it satisfies the eigenvalue equation.

- We must show that the solution satisfies for some scalar λ^2
  $\hat{C}_{x,y}\hat{C}_y^{-1}\hat{C}_{y,x}\boldsymbol{w}_x = \lambda^2\hat{C}_x\boldsymbol{w}_x$

Canonical Correlation Analysis, example

Denote the singular value decomposition $\boldsymbol{X}_{all} = \boldsymbol{UDV}$

- Right-hand side of the eigenvalue equation:

$$\lambda^2 \hat{C}_x \boldsymbol{w}_x = \lambda^2 \left(\frac{1}{N} \boldsymbol{X}_{all} \boldsymbol{X}_{all}^T\right)\left(\boldsymbol{X}_{all} \boldsymbol{X}_{all}^T\right)^{-1} \boldsymbol{X}_{all} \boldsymbol{y}_{all} = \frac{1}{N} \lambda^2 \boldsymbol{X}_{all} \boldsymbol{y}_{all}$$

- Left-hand side of the eigenvalue equation: $\hat{C}_{x,y} \hat{C}_y^{-1} \hat{C}_{y,x} \boldsymbol{w}_x =$

$$\left(\frac{1}{N} \boldsymbol{X}_{all} \boldsymbol{y}_{all}\right)\left(\frac{1}{N} \boldsymbol{y}_{all}^T \boldsymbol{y}_{all}\right)^{-1} \left(\frac{1}{N} \boldsymbol{X}_{all} \boldsymbol{y}_{all}\right)^T \left(\boldsymbol{X}_{all} \boldsymbol{X}_{all}^T\right)^{-1} \boldsymbol{X}_{all} \boldsymbol{y}_{all} =$$

$$\left(\left(N \boldsymbol{y}_{all}^T \boldsymbol{y}_{all}\right)^{-1}\right) \boldsymbol{X}_{all} \boldsymbol{y}_{all} \boldsymbol{y}_{all}^T \boldsymbol{X}_{all}^T \left(\boldsymbol{X}_{all} \boldsymbol{X}_{all}^T\right)^{-1} \boldsymbol{X}_{all} \boldsymbol{y}_{all} =$$

$$\left(\left(N \boldsymbol{y}_{all}^T \boldsymbol{y}_{all}\right)^{-1}\right) \boldsymbol{UDV}\, \boldsymbol{y}_{all} \boldsymbol{y}_{all}^T \boldsymbol{V}^T \boldsymbol{D} \boldsymbol{U}^T \left(\boldsymbol{UDV}\, \boldsymbol{V}^T \boldsymbol{D} \boldsymbol{U}^T\right)^{-1} \boldsymbol{UDV}\, \boldsymbol{y}_{all} =$$

$$\left(\left(N \boldsymbol{y}_{all}^T \boldsymbol{y}_{all}\right)^{-1}\right) \boldsymbol{UDV}\, \boldsymbol{y}_{all} \boldsymbol{y}_{all}^T \boldsymbol{V}^T \boldsymbol{D} \boldsymbol{U}^T \left(\boldsymbol{U}\, \boldsymbol{D}^2 \boldsymbol{U}^T\right)^{-1} \boldsymbol{UDV}\, \boldsymbol{y}_{all} =$$

$$\left(\left(N \boldsymbol{y}_{all}^T \boldsymbol{y}_{all}\right)^{-1}\right) \boldsymbol{UDV}\, \boldsymbol{y}_{all} \boldsymbol{y}_{all}^T \boldsymbol{V}^T \boldsymbol{D} \boldsymbol{U}^T \boldsymbol{U}\, \boldsymbol{D}^{-2} \boldsymbol{U}^T \boldsymbol{UDV}\, \boldsymbol{y}_{all} =$$

$$\left(\left(N \boldsymbol{y}_{all}^T \boldsymbol{y}_{all}\right)^{-1}\right) \boldsymbol{UDV}\, \boldsymbol{y}_{all} \boldsymbol{y}_{all}^T \boldsymbol{V}^T \boldsymbol{D}\, \boldsymbol{D}^{-2} \boldsymbol{DV}\, \boldsymbol{y}_{all} =$$

$$\left(\left(N \boldsymbol{y}_{all}^T \boldsymbol{y}_{all}\right)^{-1}\right) \boldsymbol{UDV}\, \boldsymbol{y}_{all} \boldsymbol{y}_{all}^T \boldsymbol{V}^T \boldsymbol{V}\, \boldsymbol{y}_{all} = \left(\left(N \boldsymbol{y}_{all}^T \boldsymbol{y}_{all}\right)^{-1}\right) \boldsymbol{UDV}\, \boldsymbol{y}_{all} \boldsymbol{y}_{all}^T \boldsymbol{y}_{all} =$$

$$\left(N^{-1}\right) \boldsymbol{UDV}\, \boldsymbol{y}_{all} = \left(N^{-1}\right) \boldsymbol{X}_{all} \boldsymbol{y}_{all}$$

Equality satisfied with lambda=1