

Cross-Validation

A Fast Algorithm by

Ole-Christian Galbo Engstrøm and Martin Holm Jensen

K

$\mathbf{X}_{11}$	$\mathbf{X}_{12}$	$\mathbf{X}_{13}$
$\mathbf{X}_{21}$	$\mathbf{X}_{22}$	$\mathbf{X}_{23}$
$\mathbf{X}_{31}$	$\mathbf{X}_{32}$	$\mathbf{X}_{33}$
$\mathbf{X}_{41}$	$\mathbf{X}_{42}$	$\mathbf{X}_{43}$

N

M

$\mathbf{Y}_{11}$	$\mathbf{Y}_{12}$
$\mathbf{Y}_{21}$	$\mathbf{Y}_{22}$
$\mathbf{Y}_{31}$	$\mathbf{Y}_{32}$
$\mathbf{Y}_{41}$	$\mathbf{Y}_{42}$

N

N

K

$\mathbf{X}_{11}$	$\mathbf{X}_{12}$	$\mathbf{X}_{13}$
$\mathbf{X}_{21}$	$\mathbf{X}_{22}$	$\mathbf{X}_{23}$
$\mathbf{X}_{31}$	$\mathbf{X}_{32}$	$\mathbf{X}_{33}$
$\mathbf{X}_{41}$	$\mathbf{X}_{42}$	$\mathbf{X}_{43}$

N

M

$\mathbf{Y}_{11}$	$\mathbf{Y}_{12}$
$\mathbf{Y}_{21}$	$\mathbf{Y}_{22}$
$\mathbf{Y}_{31}$	$\mathbf{Y}_{32}$
$\mathbf{Y}_{41}$	$\mathbf{Y}_{42}$

K

K

$\mathbf{X}^T\mathbf{X}_{11}$	$\mathbf{X}^T\mathbf{X}_{12}$	$\mathbf{X}^T\mathbf{X}_{13}$
$\mathbf{X}^T\mathbf{X}_{21}$	$\mathbf{X}^T\mathbf{X}_{22}$	$\mathbf{X}^T\mathbf{X}_{23}$
$\mathbf{X}^T\mathbf{X}_{31}$	$\mathbf{X}^T\mathbf{X}_{32}$	$\mathbf{X}^T\mathbf{X}_{33}$

K

M

$\mathbf{X}^T\mathbf{Y}_{11}$	$\mathbf{X}^T\mathbf{Y}_{12}$
$\mathbf{X}^T\mathbf{Y}_{21}$	$\mathbf{X}^T\mathbf{Y}_{22}$
$\mathbf{X}^T\mathbf{Y}_{31}$	$\mathbf{X}^T\mathbf{Y}_{32}$

	K		
N	X_{11}	X_{12}	X_{13}
	X_{21}	X_{22}	X_{23}
	X_{31}	X_{32}	X_{33}
	X_{41}	X_{42}	X_{43}

	M	
N	Y_{11}	Y_{12}
	Y_{21}	Y_{22}
	Y_{31}	Y_{32}
	Y_{41}	Y_{42}

	K				M	
K	$X^T X_{11}$	$X^T X_{12}$	$X^T X_{13}$	K	$X^T Y_{11}$	$X^T Y_{12}$
	$X^T X_{21}$	$X^T X_{22}$	$X^T X_{23}$		$X^T Y_{21}$	$X^T Y_{22}$
	$X^T X_{31}$	$X^T X_{32}$	$X^T X_{33}$		$X^T Y_{31}$	$X^T Y_{32}$

Ordinary/Partial Least Squares
Ridge Regression
Principal Component Analysis/Regression
Etc.

Figure 10 displays five 3D pie charts illustrating the distribution of Training and Validation data for different models. The legend indicates that orange represents Training data and blue represents Validation data. The charts show varying proportions of Training and Validation data, with Training generally being the larger portion.

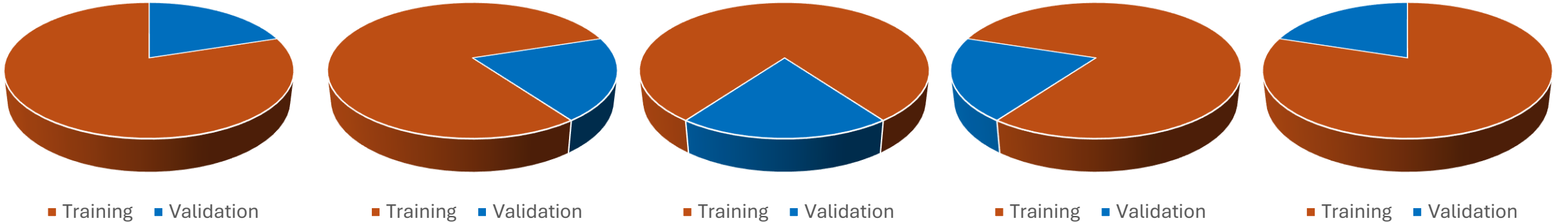
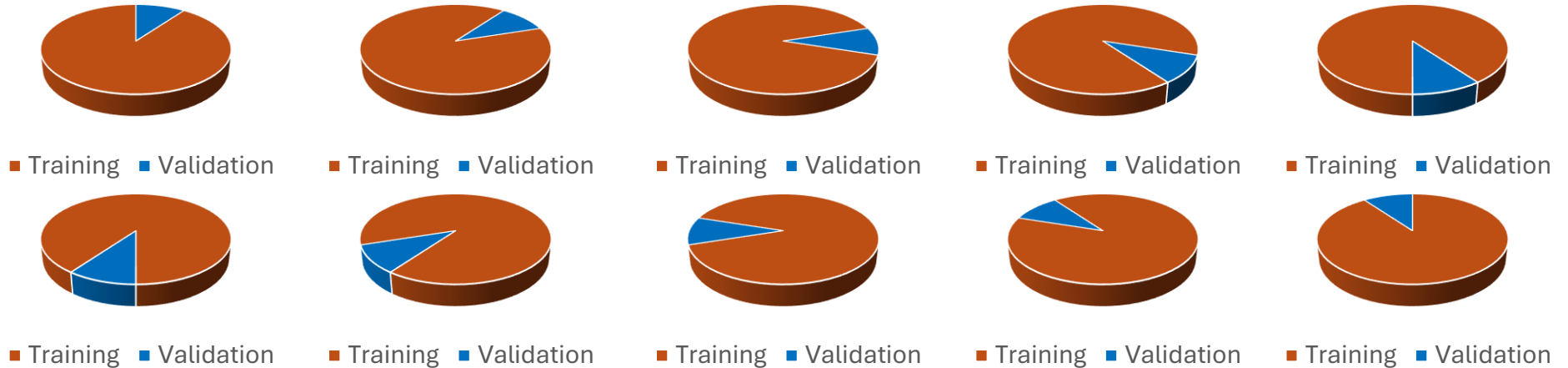


Figure 1 displays 10 3D pie charts arranged in two rows of five, illustrating the distribution of Training (orange) and Validation (blue) data for different model sizes and data splits. The top row shows models with 100,000 parameters, and the bottom row shows models with 1,000,000 parameters. The columns represent different data splits: 10/90, 20/80, 30/70, 40/60, and 50/50. In all cases, the Training data (orange) is the majority, and the Validation data (blue) is the minority. The Validation data is represented by a single blue slice in each pie chart.



Total number of samples: NP

Total number of validation samples: N

Total number of training samples: $NP - N = N(P-1) = \Theta(NP)$

X

$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$
$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	$\mathbf{x}_{33}$
$\mathbf{x}_{41}$	$\mathbf{x}_{42}$	$\mathbf{x}_{43}$

Y

$\mathbf{y}_{11}$	$\mathbf{y}_{12}$
$\mathbf{y}_{21}$	$\mathbf{y}_{22}$
$\mathbf{y}_{31}$	$\mathbf{y}_{32}$
$\mathbf{y}_{41}$	$\mathbf{y}_{42}$

$\mathbf{X}^T \mathbf{Y}$

$\mathbf{x}_{11} \mathbf{y}_{11} + \mathbf{x}_{21} \mathbf{y}_{21} +$ $\mathbf{x}_{31} \mathbf{y}_{31} + \mathbf{x}_{41} \mathbf{y}_{41}$	$\mathbf{x}_{11} \mathbf{y}_{12} + \mathbf{x}_{21} \mathbf{y}_{22} +$ $\mathbf{x}_{31} \mathbf{y}_{32} + \mathbf{x}_{41} \mathbf{y}_{42}$
$\mathbf{x}_{12} \mathbf{y}_{11} + \mathbf{x}_{22} \mathbf{y}_{21} +$ $\mathbf{x}_{32} \mathbf{y}_{31} + \mathbf{x}_{42} \mathbf{y}_{41}$	$\mathbf{x}_{12} \mathbf{y}_{12} + \mathbf{x}_{22} \mathbf{y}_{22} +$ $\mathbf{x}_{32} \mathbf{y}_{32} + \mathbf{x}_{42} \mathbf{y}_{42}$
$\mathbf{x}_{13} \mathbf{y}_{11} + \mathbf{x}_{23} \mathbf{y}_{21} +$ $\mathbf{x}_{33} \mathbf{y}_{31} + \mathbf{x}_{43} \mathbf{y}_{41}$	$\mathbf{x}_{13} \mathbf{y}_{12} + \mathbf{x}_{23} \mathbf{y}_{22} +$ $\mathbf{x}_{33} \mathbf{y}_{32} + \mathbf{x}_{43} \mathbf{y}_{42}$

$\mathbf{X}^T$

$\mathbf{x}_{11}$	$\mathbf{x}_{21}$	$\mathbf{x}_{31}$	$\mathbf{x}_{41}$
$\mathbf{x}_{12}$	$\mathbf{x}_{22}$	$\mathbf{x}_{32}$	$\mathbf{x}_{42}$
$\mathbf{x}_{13}$	$\mathbf{x}_{23}$	$\mathbf{x}_{33}$	$\mathbf{x}_{43}$

X

$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$
$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	$\mathbf{x}_{33}$
$\mathbf{x}_{41}$	$\mathbf{x}_{42}$	$\mathbf{x}_{43}$

Y

$\mathbf{y}_{11}$	$\mathbf{y}_{12}$
$\mathbf{y}_{21}$	$\mathbf{y}_{22}$
$\mathbf{y}_{31}$	$\mathbf{y}_{32}$
$\mathbf{y}_{41}$	$\mathbf{y}_{42}$



Training



Validation

Fold 1

$\mathbf{X}^T$

$\mathbf{x}_{11}$	$\mathbf{x}_{21}$	$\mathbf{x}_{31}$	$\mathbf{x}_{41}$
$\mathbf{x}_{12}$	$\mathbf{x}_{22}$	$\mathbf{x}_{32}$	$\mathbf{x}_{42}$
$\mathbf{x}_{13}$	$\mathbf{x}_{23}$	$\mathbf{x}_{33}$	$\mathbf{x}_{43}$

$\mathbf{X}^T \mathbf{Y}_{\text{Train}}$

$\mathbf{x}_{21} \mathbf{y}_{21} + \mathbf{x}_{31} \mathbf{y}_{31} +$ $\mathbf{x}_{41} \mathbf{y}_{41}$	$\mathbf{x}_{21} \mathbf{y}_{22} + \mathbf{x}_{31} \mathbf{y}_{32} +$ $\mathbf{x}_{41} \mathbf{y}_{42}$
$\mathbf{x}_{22} \mathbf{y}_{21} + \mathbf{x}_{32} \mathbf{y}_{31} +$ $\mathbf{x}_{42} \mathbf{y}_{41}$	$\mathbf{x}_{22} \mathbf{y}_{22} + \mathbf{x}_{32} \mathbf{y}_{32} +$ $\mathbf{x}_{42} \mathbf{y}_{42}$
$\mathbf{x}_{23} \mathbf{y}_{21} + \mathbf{x}_{33} \mathbf{y}_{31} +$ $\mathbf{x}_{43} \mathbf{y}_{41}$	$\mathbf{x}_{23} \mathbf{y}_{22} + \mathbf{x}_{33} \mathbf{y}_{32} +$ $\mathbf{x}_{43} \mathbf{y}_{42}$

X

$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$
$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	$\mathbf{x}_{33}$
$\mathbf{x}_{41}$	$\mathbf{x}_{42}$	$\mathbf{x}_{43}$

Y

$\mathbf{y}_{11}$	$\mathbf{y}_{12}$
$\mathbf{y}_{21}$	$\mathbf{y}_{22}$
$\mathbf{y}_{31}$	$\mathbf{y}_{32}$
$\mathbf{y}_{41}$	$\mathbf{y}_{42}$



Training



Validation

Fold 2

$\mathbf{X}^T$

$\mathbf{x}_{11}$	$\mathbf{x}_{21}$	$\mathbf{x}_{31}$	$\mathbf{x}_{41}$
$\mathbf{x}_{12}$	$\mathbf{x}_{22}$	$\mathbf{x}_{32}$	$\mathbf{x}_{42}$
$\mathbf{x}_{13}$	$\mathbf{x}_{23}$	$\mathbf{x}_{33}$	$\mathbf{x}_{43}$

$\mathbf{X}^T \mathbf{Y}_{\text{Train}}$

$\mathbf{x}_{11} \mathbf{y}_{11} + \mathbf{x}_{31} \mathbf{y}_{31} +$ $\mathbf{x}_{41} \mathbf{y}_{41}$	$\mathbf{x}_{11} \mathbf{y}_{12} + \mathbf{x}_{31} \mathbf{y}_{32} +$ $\mathbf{x}_{41} \mathbf{y}_{42}$
$\mathbf{x}_{12} \mathbf{y}_{11} + \mathbf{x}_{32} \mathbf{y}_{31} +$ $\mathbf{x}_{42} \mathbf{y}_{41}$	$\mathbf{x}_{12} \mathbf{y}_{12} + \mathbf{x}_{32} \mathbf{y}_{32} +$ $\mathbf{x}_{42} \mathbf{y}_{42}$
$\mathbf{x}_{13} \mathbf{y}_{11} + \mathbf{x}_{33} \mathbf{y}_{31} +$ $\mathbf{x}_{43} \mathbf{y}_{41}$	$\mathbf{x}_{13} \mathbf{y}_{12} + \mathbf{x}_{33} \mathbf{y}_{32} +$ $\mathbf{x}_{43} \mathbf{y}_{42}$

X

$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$
$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	$\mathbf{x}_{33}$
$\mathbf{x}_{41}$	$\mathbf{x}_{42}$	$\mathbf{x}_{43}$

Y

$\mathbf{y}_{11}$	$\mathbf{y}_{12}$
$\mathbf{y}_{21}$	$\mathbf{y}_{22}$
$\mathbf{y}_{31}$	$\mathbf{y}_{32}$
$\mathbf{y}_{41}$	$\mathbf{y}_{42}$



Training



Validation

Fold 3

$\mathbf{X}^T$

$\mathbf{x}_{11}$	$\mathbf{x}_{21}$	$\mathbf{x}_{31}$	$\mathbf{x}_{41}$
$\mathbf{x}_{12}$	$\mathbf{x}_{22}$	$\mathbf{x}_{32}$	$\mathbf{x}_{42}$
$\mathbf{x}_{13}$	$\mathbf{x}_{23}$	$\mathbf{x}_{33}$	$\mathbf{x}_{43}$

$\mathbf{X}^T \mathbf{Y}_{\text{Train}}$

$\mathbf{x}_{11} \mathbf{y}_{11} + \mathbf{x}_{21} \mathbf{y}_{21} + \mathbf{x}_{41} \mathbf{y}_{41}$	$\mathbf{x}_{11} \mathbf{y}_{12} + \mathbf{x}_{21} \mathbf{y}_{22} + \mathbf{x}_{41} \mathbf{y}_{42}$
$\mathbf{x}_{12} \mathbf{y}_{11} + \mathbf{x}_{22} \mathbf{y}_{21} + \mathbf{x}_{42} \mathbf{y}_{41}$	$\mathbf{x}_{12} \mathbf{y}_{12} + \mathbf{x}_{22} \mathbf{y}_{22} + \mathbf{x}_{42} \mathbf{y}_{42}$
$\mathbf{x}_{13} \mathbf{y}_{11} + \mathbf{x}_{23} \mathbf{y}_{21} + \mathbf{x}_{43} \mathbf{y}_{41}$	$\mathbf{x}_{13} \mathbf{y}_{12} + \mathbf{x}_{23} \mathbf{y}_{22} + \mathbf{x}_{43} \mathbf{y}_{42}$

X

$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$
$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$
$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	$\mathbf{x}_{33}$
$\mathbf{x}_{41}$	$\mathbf{x}_{42}$	$\mathbf{x}_{43}$

Y

$\mathbf{y}_{11}$	$\mathbf{y}_{12}$
$\mathbf{y}_{21}$	$\mathbf{y}_{22}$
$\mathbf{y}_{31}$	$\mathbf{y}_{32}$
$\mathbf{y}_{41}$	$\mathbf{y}_{42}$



Training



Validation

Fold 4

$\mathbf{X}^T$

$\mathbf{x}_{11}$	$\mathbf{x}_{21}$	$\mathbf{x}_{31}$	$\mathbf{x}_{41}$
$\mathbf{x}_{12}$	$\mathbf{x}_{22}$	$\mathbf{x}_{32}$	$\mathbf{x}_{42}$
$\mathbf{x}_{13}$	$\mathbf{x}_{23}$	$\mathbf{x}_{33}$	$\mathbf{x}_{43}$

$\mathbf{X}^T \mathbf{Y}_{\text{Train}}$

$\mathbf{x}_{11} \mathbf{y}_{11} + \mathbf{x}_{21} \mathbf{y}_{21} + \mathbf{x}_{31} \mathbf{y}_{31}$	$\mathbf{x}_{11} \mathbf{y}_{12} + \mathbf{x}_{21} \mathbf{y}_{22} + \mathbf{x}_{31} \mathbf{y}_{32}$
$\mathbf{x}_{12} \mathbf{y}_{11} + \mathbf{x}_{22} \mathbf{y}_{21} + \mathbf{x}_{32} \mathbf{y}_{31}$	$\mathbf{x}_{12} \mathbf{y}_{12} + \mathbf{x}_{22} \mathbf{y}_{22} + \mathbf{x}_{32} \mathbf{y}_{32}$
$\mathbf{x}_{13} \mathbf{y}_{11} + \mathbf{x}_{23} \mathbf{y}_{21} + \mathbf{x}_{33} \mathbf{y}_{31}$	$\mathbf{x}_{13} \mathbf{y}_{12} + \mathbf{x}_{23} \mathbf{y}_{22} + \mathbf{x}_{33} \mathbf{y}_{32}$

N rows in **X** and **Y**

K columns in **X**

M columns in **Y**

Each $\mathbf{X}^T \mathbf{Y}_{\text{Train}}$
Time: $\Theta(NKM)$

P folds:
Time: $\Theta(PNKM)$

N rows in $\mathbf{X}$ and $\mathbf{Y}$
K columns in $\mathbf{X}$
M columns in $\mathbf{Y}$

... Can we do better?

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

X

x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
x_{31}	x_{32}	x_{33}
x_{41}	x_{42}	x_{43}

Y

y_{11}	y_{12}
y_{21}	y_{22}
y_{31}	y_{32}
y_{41}	y_{42}



Training



Validation

Fold 1

 $X^T Y$ _{Train}

$x_{11}y_{11} + x_{21}y_{21} + x_{31}y_{31}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32}$

=

 $X^T Y$

$x_{11}y_{11} + x_{21}y_{21} + x_{31}y_{31} + x_{41}y_{41}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32} + x_{41}y_{42}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31} + x_{42}y_{41}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32} + x_{42}y_{42}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31} + x_{43}y_{41}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32} + x_{43}y_{42}$

-

 $X^T Y$ _{Val}

$x_{11}y_{11}$	$x_{11}y_{12}$
$x_{12}y_{11}$	$x_{12}y_{12}$
$x_{13}y_{11}$	$x_{13}y_{12}$

X

x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
x_{31}	x_{32}	x_{33}
x_{41}	x_{42}	x_{43}

Y

y_{11}	y_{12}
y_{21}	y_{22}
y_{31}	y_{32}
y_{41}	y_{42}



Training



Validation

Fold 2

 $X^T Y_{\text{Train}}$

$x_{11}y_{11} + x_{31}y_{31} + x_{41}y_{41}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32}$

=

 $X^T Y$

$x_{11}y_{11} + x_{21}y_{21} + x_{31}y_{31} + x_{41}y_{41}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32} + x_{41}y_{42}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31} + x_{42}y_{41}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32} + x_{42}y_{42}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31} + x_{43}y_{41}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32} + x_{43}y_{42}$

-

 $X^T Y_{\text{Val}}$

$x_{21}y_{21}$	$x_{21}y_{22}$
$x_{22}y_{21}$	$x_{22}y_{22}$
$x_{23}y_{21}$	$x_{23}y_{22}$

X

x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
x_{31}	x_{32}	x_{33}
x_{41}	x_{42}	x_{43}

Y

y_{11}	y_{12}
y_{21}	y_{22}
y_{31}	y_{32}
y_{41}	y_{42}



Training



Validation

Fold 3

 $X^T Y$ _{Train}

$x_{11}y_{11} + x_{21}y_{21} + x_{41}y_{41}$	$x_{11}y_{12} + x_{21}y_{22} + x_{41}y_{42}$
$x_{12}y_{11} + x_{22}y_{21} + x_{42}y_{41}$	$x_{12}y_{12} + x_{22}y_{22} + x_{42}y_{42}$
$x_{13}y_{11} + x_{23}y_{21} + x_{43}y_{41}$	$x_{13}y_{12} + x_{23}y_{22} + x_{43}y_{42}$

 $X^T Y$

=

$x_{11}y_{11} + x_{21}y_{21} + x_{31}y_{31} + x_{41}y_{41}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32} + x_{41}y_{42}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31} + x_{42}y_{41}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32} + x_{42}y_{42}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31} + x_{43}y_{41}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32} + x_{43}y_{42}$

-

 $X^T Y$ _{Val}

$x_{31}y_{31}$	$x_{31}y_{32}$
$x_{32}y_{31}$	$x_{32}y_{32}$
$x_{33}y_{31}$	$x_{33}y_{32}$

X

x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
x_{31}	x_{32}	x_{33}
x_{41}	x_{42}	x_{43}

Y

y_{11}	y_{12}
y_{21}	y_{22}
y_{31}	y_{32}
y_{41}	y_{42}



Training



Validation

Fold 4

 $X^T Y_{\text{Train}}$

$x_{11}y_{11} + x_{21}y_{21} + x_{31}y_{31}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32}$

 $X^T Y$

=

$x_{11}y_{11} + x_{21}y_{21} + x_{31}y_{31} + x_{41}y_{41}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32} + x_{41}y_{42}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31} + x_{42}y_{41}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32} + x_{42}y_{42}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31} + x_{43}y_{41}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32} + x_{43}y_{42}$

-

 $X^T Y_{\text{Val}}$

$x_{41}y_{41}$	$x_{41}y_{42}$
$x_{42}y_{41}$	$x_{42}y_{42}$
$x_{43}y_{41}$	$x_{43}y_{42}$

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

Once:

$$X^T Y$$

P times:

$$X^T Y_{\text{Val}}$$

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

N rows in **X** and **Y**
 K columns in **X**
 M columns in **Y**

Once:
 $X^T Y$

$\Theta(NKM)$

P times:

$X^T Y_{\text{Val}}$

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

$\Theta(NKM)$

$\Theta(PKM) = O(NKM)$

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

N rows in **X** and **Y**
 K columns in **X**
 M columns in **Y**

Once:
 $X^T Y$

P times:

$$X^T Y_{\text{Val}}$$

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

$$\Theta(NKM)$$

$$\Theta(NKM)$$

$$\Theta(PKM) = O(NKM)$$

$$\Theta(NKM)$$

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

N rows in **X** and **Y**
 K columns in **X**
 M columns in **Y**

Asymptotically independent on P! ←

Once:
 $X^T Y$

P times:

$$X^T Y_{\text{Val}} \\ X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

$\Theta(NKM)$

$\Theta(NKM)$

$\Theta(PKM) = O(NKM)$

$\Theta(NKM)$

Cost of computing all $P X^T Y_{\text{Train}}$
Improved from $\Theta(PNKM)$ to $\Theta(NKM)$

Cost of computing all $P X^T Y_{\text{Train}}$
Improved from $\Theta(PNKM)$ to $\Theta(NKM)$

Cost of computing all $P X^T X_{\text{Train}}$
Improved from $\Theta(PNKK)$ to $\Theta(NKK)$

Cost of computing all $P X^T Y_{\text{Train}}$ and $X^T X_{\text{Train}}$
Improved from $\Theta(PNK(K+M))$ to $\Theta(NK(K+M))$

... Any row-wise preprocessing works out of the box

SNV, i.e., (row-wise) centering and scaling
Convolution (e.g. with Savitzky-Golay filters)



$$\begin{array}{|c|c|} \hline \textcolor{brown}{X}_{11}\textcolor{brown}{Y}_{11} + \textcolor{brown}{X}_{21}\textcolor{brown}{Y}_{21} + & \textcolor{brown}{X}_{11}\textcolor{brown}{Y}_{12} + \textcolor{brown}{X}_{21}\textcolor{brown}{Y}_{22} + \\ \textcolor{brown}{X}_{31}\textcolor{brown}{Y}_{31} & \textcolor{brown}{X}_{31}\textcolor{brown}{Y}_{32} \\ \hline \textcolor{brown}{X}_{12}\textcolor{brown}{Y}_{11} + \textcolor{brown}{X}_{22}\textcolor{brown}{Y}_{21} + & \textcolor{brown}{X}_{12}\textcolor{brown}{Y}_{12} + \textcolor{brown}{X}_{22}\textcolor{brown}{Y}_{22} + \\ \textcolor{brown}{X}_{32}\textcolor{brown}{Y}_{31} & \textcolor{brown}{X}_{32}\textcolor{brown}{Y}_{32} \\ \hline \textcolor{brown}{X}_{13}\textcolor{brown}{Y}_{11} + \textcolor{brown}{X}_{23}\textcolor{brown}{Y}_{21} + & \textcolor{brown}{X}_{13}\textcolor{brown}{Y}_{12} + \textcolor{brown}{X}_{23}\textcolor{brown}{Y}_{22} + \\ \textcolor{brown}{X}_{33}\textcolor{brown}{Y}_{31} & \textcolor{brown}{X}_{33}\textcolor{brown}{Y}_{32} \\ \hline \end{array}
 =
 \begin{array}{|c|c|} \hline \textcolor{brown}{X}_{11}\textcolor{brown}{Y}_{11} + \textcolor{brown}{X}_{21}\textcolor{brown}{Y}_{21} + & \textcolor{brown}{X}_{11}\textcolor{brown}{Y}_{12} + \textcolor{brown}{X}_{21}\textcolor{brown}{Y}_{22} + \\ \textcolor{brown}{X}_{31}\textcolor{brown}{Y}_{31} + \textcolor{blue}{X}_{41}\textcolor{blue}{Y}_{41} & \textcolor{brown}{X}_{31}\textcolor{brown}{Y}_{32} + \textcolor{blue}{X}_{41}\textcolor{blue}{Y}_{42} \\ \hline \textcolor{brown}{X}_{12}\textcolor{brown}{Y}_{11} + \textcolor{brown}{X}_{22}\textcolor{brown}{Y}_{21} + & \textcolor{brown}{X}_{12}\textcolor{brown}{Y}_{12} + \textcolor{brown}{X}_{22}\textcolor{brown}{Y}_{22} + \\ \textcolor{brown}{X}_{32}\textcolor{brown}{Y}_{31} + \textcolor{blue}{X}_{42}\textcolor{blue}{Y}_{41} & \textcolor{brown}{X}_{32}\textcolor{brown}{Y}_{32} + \textcolor{blue}{X}_{42}\textcolor{blue}{Y}_{42} \\ \hline \textcolor{brown}{X}_{13}\textcolor{brown}{Y}_{11} + \textcolor{brown}{X}_{23}\textcolor{brown}{Y}_{21} + & \textcolor{brown}{X}_{13}\textcolor{brown}{Y}_{12} + \textcolor{brown}{X}_{23}\textcolor{brown}{Y}_{22} + \\ \textcolor{brown}{X}_{33}\textcolor{brown}{Y}_{31} + \textcolor{blue}{X}_{43}\textcolor{blue}{Y}_{41} & \textcolor{brown}{X}_{33}\textcolor{brown}{Y}_{32} + \textcolor{blue}{X}_{43}\textcolor{blue}{Y}_{42} \\ \hline \end{array}
 -
 \begin{array}{|c|c|} \hline \textcolor{blue}{X}_{41}\textcolor{blue}{Y}_{41} & \textcolor{blue}{X}_{41}\textcolor{blue}{Y}_{42} \\ \hline \textcolor{blue}{X}_{42}\textcolor{blue}{Y}_{41} & \textcolor{blue}{X}_{42}\textcolor{blue}{Y}_{42} \\ \hline \textcolor{blue}{X}_{43}\textcolor{blue}{Y}_{41} & \textcolor{blue}{X}_{43}\textcolor{blue}{Y}_{42} \\ \hline \end{array}$$

... What about (column-wise) centering and scaling (autoscaling)?

X		
x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
x_{31}	x_{32}	x_{33}
x_{41}	x_{42}	x_{43}

Y	
y_{11}	y_{12}
y_{21}	y_{22}
y_{31}	y_{32}
y_{41}	y_{42}

Fold 4

 Training
 Validation

$X^T Y_{\text{Train}}$	
$x_{11}y_{11} + x_{21}y_{21} + x_{31}y_{31}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32}$

=

$X^T Y$	
$x_{11}y_{11} + x_{21}y_{21} + x_{31}y_{31} + x_{41}y_{41}$	$x_{11}y_{12} + x_{21}y_{22} + x_{31}y_{32} + x_{41}y_{42}$
$x_{12}y_{11} + x_{22}y_{21} + x_{32}y_{31} + x_{42}y_{41}$	$x_{12}y_{12} + x_{22}y_{22} + x_{32}y_{32} + x_{42}y_{42}$
$x_{13}y_{11} + x_{23}y_{21} + x_{33}y_{31} + x_{43}y_{41}$	$x_{13}y_{12} + x_{23}y_{22} + x_{33}y_{32} + x_{43}y_{42}$

-

$X^T Y_{\text{Val}}$	
$x_{41}y_{41}$	$x_{41}y_{42}$
$x_{42}y_{41}$	$x_{42}y_{42}$
$x_{43}y_{41}$	$x_{43}y_{42}$

$$X^T Y_{\text{Train}} = X^T Y - X^T Y_{\text{Val}}$$

$X^T Y_{\text{Train}}$

$X_{11} Y_{11} + X_{21} Y_{21} +$ $X_{31} Y_{31}$	$X_{11} Y_{12} + X_{21} Y_{22} +$ $X_{31} Y_{32}$
$X_{12} Y_{11} + X_{22} Y_{21} +$ $X_{32} Y_{31}$	$X_{12} Y_{12} + X_{22} Y_{22} +$ $X_{32} Y_{32}$
$X_{13} Y_{11} + X_{23} Y_{21} +$ $X_{33} Y_{31}$	$X_{13} Y_{12} + X_{23} Y_{22} +$ $X_{33} Y_{32}$

$X^T Y$

$X_{11} Y_{11} + X_{21} Y_{21} +$ $X_{31} Y_{31} + X_{41} Y_{41}$	$X_{11} Y_{12} + X_{21} Y_{22} +$ $X_{31} Y_{32} + X_{41} Y_{42}$
$X_{12} Y_{11} + X_{22} Y_{21} +$ $X_{32} Y_{31} + X_{42} Y_{41}$	$X_{12} Y_{12} + X_{22} Y_{22} +$ $X_{32} Y_{32} + X_{42} Y_{42}$
$X_{13} Y_{11} + X_{23} Y_{21} +$ $X_{33} Y_{31} + X_{43} Y_{41}$	$X_{13} Y_{12} + X_{23} Y_{22} +$ $X_{33} Y_{32} + X_{43} Y_{42}$

$X^T Y_{\text{Val}}$

$X_{41} Y_{41}$	$X_{41} Y_{42}$
$X_{42} Y_{41}$	$X_{42} Y_{42}$
$X_{43} Y_{41}$	$X_{43} Y_{42}$

Break up statistics into parts

Once:

- Compute parts for entire dataset

P times:

- Compute parts for validation set

- Derive training set parts

16 combinations of centering and scaling for **X** and **Y**

16 combinations of centering and scaling for **X** and **Y**

4 unique configurations for **$X^T X$**

16 combinations of centering and scaling for **X** and **Y**

4 unique configurations for **X^TX**

Only 8 unique configurations of **X^TY** (out of 16)

16 combinations of centering and scaling for $\mathbf{X}$ and $\mathbf{Y}$

4 unique configurations for $\mathbf{X}^T\mathbf{X}$

Only 8 unique configurations of $\mathbf{X}^T\mathbf{Y}$ (out of 16)

All improved from $\Theta(PNK(K+M))$ to $\Theta(NK(K+M))$

	K				M			K				M	
N	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$	N	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	K	$\mathbf{x}^T\mathbf{x}_{11}$	$\mathbf{x}^T\mathbf{x}_{12}$	$\mathbf{x}^T\mathbf{x}_{13}$	K	$\mathbf{x}^T\mathbf{y}_{11}$	$\mathbf{x}^T\mathbf{y}_{12}$
	$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$		$\mathbf{y}_{21}$	$\mathbf{y}_{22}$		$\mathbf{x}^T\mathbf{x}_{21}$	$\mathbf{x}^T\mathbf{x}_{22}$	$\mathbf{x}^T\mathbf{x}_{23}$		$\mathbf{x}^T\mathbf{y}_{21}$	$\mathbf{x}^T\mathbf{y}_{22}$
	$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	$\mathbf{x}_{33}$		$\mathbf{y}_{31}$	$\mathbf{y}_{32}$		$\mathbf{x}^T\mathbf{x}_{31}$	$\mathbf{x}^T\mathbf{x}_{32}$	$\mathbf{x}^T\mathbf{x}_{33}$		$\mathbf{x}^T\mathbf{y}_{31}$	$\mathbf{x}^T\mathbf{y}_{32}$
	$\mathbf{x}_{41}$	$\mathbf{x}_{42}$	$\mathbf{x}_{43}$		$\mathbf{y}_{41}$	$\mathbf{y}_{42}$							

Algorithm 7 Fast Cross-Validation Algorithm with Centering and Scaling

Input: $\mathbf{X} \in \mathbb{R}^{N \times K}$, $\mathbf{Y} \in \mathbb{R}^{N \times M}$, $\mathcal{P}$ is a scalable partitioning of $R = \{1, \dots, N\}$

- 1: $R \leftarrow \{1, \dots, N\}$
 - 2: Let $\mathcal{V}$ be the result of executing Algorithm 1 with $\mathcal{P}$ and R as input
 - 3: $\mathbf{X}\mathbf{T}\mathbf{X} \leftarrow \mathbf{X}^T \mathbf{X}$, $\mathbf{X}\mathbf{T}\mathbf{Y} \leftarrow \mathbf{X}^T \mathbf{Y}$, $\mu_{\mathbf{X}} \leftarrow \bar{\mathbf{x}}$, $\mu_{\mathbf{Y}} \leftarrow \bar{\mathbf{y}}$
 - 4: $\Sigma \mathbf{X} \leftarrow \sum_{n \in R} \mathbf{X}_n$, $\Sigma \mathbf{Y} \leftarrow \sum_{n \in R} \mathbf{Y}_n$, $\Sigma sq \mathbf{X} \leftarrow \sum_{n \in R} (\mathbf{X}_n^{\circ 2})$, $\Sigma sq \mathbf{Y} \leftarrow \sum_{n \in R} (\mathbf{Y}_n^{\circ 2})$
 - 5: **for** each partition $p \in \{1, \dots, P\}$ **do**
 - 6: $V \leftarrow \mathcal{V}[p]$
 - 7: $\mathbf{X}_V \leftarrow \mathbf{X}_V$, $\mathbf{Y}_V \leftarrow \mathbf{Y}_V$
 - 8: $\mathbf{X}\mathbf{T}\mathbf{X}_V \leftarrow \mathbf{X}_V^T \mathbf{X}_V$, $\mathbf{X}\mathbf{T}\mathbf{Y}_V \leftarrow \mathbf{X}_V^T \mathbf{Y}_V$
 - 9: $\mathbf{X}\mathbf{T}\mathbf{X}_T \leftarrow \mathbf{X}\mathbf{T}\mathbf{X} - \mathbf{X}\mathbf{T}\mathbf{X}_V$, $\mathbf{X}\mathbf{T}\mathbf{Y}_T \leftarrow \mathbf{X}\mathbf{T}\mathbf{Y} - \mathbf{X}\mathbf{T}\mathbf{Y}_V$ $\triangleright$ Obtain $\mathbf{X}_T^T \mathbf{X}_T$ and $\mathbf{X}_T^T \mathbf{Y}_T$
 - 10: $\mu_{\mathbf{X}_V} \leftarrow \bar{\mathbf{x}}_V$, $\mu_{\mathbf{Y}_V} \leftarrow \bar{\mathbf{y}}_V$
 - 11: $N_V \leftarrow |V|$, $N_T \leftarrow N - N_V$
 - 12: $\mu_{\mathbf{X}_T} \leftarrow \frac{N}{N_T} \mu_{\mathbf{X}} - \frac{N_V}{N_T} \mu_{\mathbf{X}_V}$ $\triangleright$ Obtain $\bar{\mathbf{x}}_T$
 - 13: $\mu_{\mathbf{Y}_T} \leftarrow \frac{N}{N_T} \mu_{\mathbf{Y}} - \frac{N_V}{N_T} \mu_{\mathbf{Y}_V}$ $\triangleright$ Obtain $\bar{\mathbf{y}}_T$
 - 14: $\mu_{\mathbf{X}_T} \mathbf{T} \mathbf{x}_T \leftarrow N_T (\mu_{\mathbf{X}_T}^T \mu_{\mathbf{X}_T})$, $\mu_{\mathbf{X}_T} \mathbf{T} \mathbf{y}_T \leftarrow N_T (\mu_{\mathbf{X}_T}^T \mu_{\mathbf{Y}_T})$ $\triangleright$ Obtain $|T| \bar{\mathbf{x}}_T^T \bar{\mathbf{x}}_T$ and $|T| \bar{\mathbf{x}}_T^T \bar{\mathbf{y}}_T$
 - 15: $\mathbf{c}\mathbf{X}\mathbf{T}\mathbf{X}_T \leftarrow \mathbf{X}\mathbf{T}\mathbf{X}_T - \mu_{\mathbf{X}_T} \mathbf{T} \mathbf{x}_T$ $\triangleright$ Obtain $\mathbf{X}_T^{\mathbf{c}T} \mathbf{X}_T^{\mathbf{c}}$
 - 16: $\mathbf{c}\mathbf{X}\mathbf{T}\mathbf{Y}_T \leftarrow \mathbf{X}\mathbf{T}\mathbf{Y}_T - \mu_{\mathbf{X}_T} \mathbf{T} \mathbf{y}_T$ $\triangleright$ Obtain $\mathbf{X}_T^{\mathbf{c}T} \mathbf{Y}_T^{\mathbf{c}}$
 - 17: $\Sigma \mathbf{X}_V \leftarrow \sum_{n \in V} \mathbf{X}_n$, $\Sigma \mathbf{Y}_V \leftarrow \sum_{n \in V} \mathbf{Y}_n$
 - 18: $\Sigma \mathbf{X}_T \leftarrow \Sigma \mathbf{X} - \Sigma \mathbf{X}_V$, $\Sigma \mathbf{Y}_T \leftarrow \Sigma \mathbf{Y} - \Sigma \mathbf{Y}_V$ $\triangleright$ Obtain $\Sigma_{n \in T} \mathbf{X}_n$ and $\Sigma_{n \in T} \mathbf{Y}_n$
 - 19: $\Sigma sq \mathbf{X}_V \leftarrow \sum_{n \in V} (\mathbf{X}_n^{\circ 2})$, $\Sigma sq \mathbf{Y}_V \leftarrow \sum_{n \in V} (\mathbf{Y}_n^{\circ 2})$
 - 20: $\Sigma sq \mathbf{X}_T \leftarrow \Sigma sq \mathbf{X} - \Sigma sq \mathbf{X}_V$ $\triangleright$ Obtain $\Sigma_{n \in T} (\mathbf{X}_n^{\circ 2})$
 - 21: $\Sigma sq \mathbf{Y}_T \leftarrow \Sigma sq \mathbf{Y} - \Sigma sq \mathbf{Y}_V$ $\triangleright$ Obtain $\Sigma_{n \in T} (\mathbf{Y}_n^{\circ 2})$
 - 22: $\sigma_{\mathbf{X}_T} \leftarrow \left(\frac{1}{N_T - 1} (-2\mu_{\mathbf{X}_T} \odot \Sigma \mathbf{X}_T + N_T \mu_{\mathbf{X}_T}^{\circ 2} + \Sigma sq \mathbf{X}_T) \right)^{\circ \frac{1}{2}}$ $\triangleright$ Obtain $\widehat{\mathbf{x}}_T$
 - 23: $\sigma_{\mathbf{Y}_T} \leftarrow \left(\frac{1}{N_T - 1} (-2\mu_{\mathbf{Y}_T} \odot \Sigma \mathbf{Y}_T + N_T \mu_{\mathbf{Y}_T}^{\circ 2} + \Sigma sq \mathbf{Y}_T) \right)^{\circ \frac{1}{2}}$ $\triangleright$ Obtain $\widehat{\mathbf{y}}_T$
 - 24: Replace with 1 any entry in $\sigma_{\mathbf{X}_T}$ and $\sigma_{\mathbf{Y}_T}$ that is 0.
 - 25: $\sigma_{\mathbf{X}_T} \mathbf{T} \mathbf{x}_T \leftarrow \sigma_{\mathbf{X}_T}^T \sigma_{\mathbf{X}_T}$, $\sigma_{\mathbf{X}_T} \mathbf{T} \mathbf{y}_T \leftarrow \sigma_{\mathbf{X}_T}^T \sigma_{\mathbf{Y}_T}$ $\triangleright$ Obtain $\widehat{\mathbf{x}}_T^T \widehat{\mathbf{x}}_T$ and $\widehat{\mathbf{x}}_T^T \widehat{\mathbf{y}}_T$
 - 26: $\mathbf{cs}\mathbf{X}\mathbf{T}\mathbf{X}_T \leftarrow \mathbf{c}\mathbf{X}\mathbf{T}\mathbf{X}_T \oslash \sigma_{\mathbf{X}_T} \mathbf{T} \mathbf{x}_T$ $\triangleright$ Obtain $\mathbf{X}_T^{\mathbf{cs}T} \mathbf{X}_T^{\mathbf{cs}}$
 - 27: $\mathbf{cs}\mathbf{X}\mathbf{T}\mathbf{Y}_T \leftarrow \mathbf{c}\mathbf{X}\mathbf{T}\mathbf{Y}_T \oslash \sigma_{\mathbf{X}_T} \mathbf{T} \mathbf{y}_T$ $\triangleright$ Obtain $\mathbf{X}_T^{\mathbf{cs}T} \mathbf{Y}_T^{\mathbf{cs}}$
 - 28: **end for**
-

Does it really matter?

Does it really matter?

Demo time!

Contributions

Faster cross-validation for $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Y}$

Solved for any combination of centering and scaling

Software implementing the algorithm

`pip install cvmatrix`

Software combining the algorithm with PLS

`pip install ikpls`

Article

