Exact Incremental and Distributed Regression

Ying-zong Huang

May 6, 2016

1 Canonical regression

Given sequences \( \{x_i\}, \{y_i\}, \{n_i\} \in \mathbb{R}^n \), a matrix \( A \in \mathbb{R}^{n \times k} \), and a noisy linear system

\[
\begin{bmatrix}
  y_i \\
  \vdots \\
  n_i
\end{bmatrix} = \begin{bmatrix}
  \vdots \\
  A \\
  \vdots \\
\end{bmatrix} \begin{bmatrix}
  x_i \\
  \vdots \\
  n_i
\end{bmatrix}
\]

(1)

the canonical regression problem is the inversion

\[
A^* = \arg \min_A \sum_i \|Ax_i - y_i\|_2^2
\]

(2)

this being the equivalent to the least-squares estimator

\[
A^* = \arg \min_A \mathbb{E} \|Ax - y\|_2^2
\]

(3)

for the same linear system \( y = Ax + n \) when \( x, y, n \) are zero-mean random variables and \( n \) is uncorrelated with \( x \), under the objective of minimizing estimator noise variance.

The objective of Eq. (2) is quadratic in \( A \), and we formally differentiate and apply critical point methods, i.e.

\[
\nabla_A \sum_i \|Ax_i - y_i\|_2^2 = \sum_i 2(Ax_i - y_i)x_i^T
\]

(4)

\[
= 2 \sum_i Ax_i x_i^T - y_i x_i^T
\]

(5)

where

\[
\nabla_A f = \begin{bmatrix}
  \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1k}} \\
  \vdots & \ddots & \vdots \\
  \frac{\partial f}{\partial A_{n1}} & \cdots & \frac{\partial f}{\partial A_{nk}}
\end{bmatrix}^T = \begin{bmatrix}
  \nabla_a^{T(0)} f \\
  \vdots \\
  \nabla_a^{T(n-1)} f
\end{bmatrix}^T
\]

(6)
More pedantically, the objective of Eq. (2) is separable in each row of

$$A = \begin{bmatrix} a_{(0)}^T \\ \vdots \\ a_{(n-1)}^T \end{bmatrix}$$

because we can separately minimize

$$a_{(u)}^* = \arg \min_{a_{(u)}} \sum_i \| a_{(u)}^T x_i - y_i^{(u)} \|_2^2$$

Without loss of generality, take $a \triangleq a_{(u)}$, $x \triangleq x_i$, $y \triangleq y_i^{(u)}$. Expanding each term,

$$\| a^T x - y \|_2^2 = (x^T a - y^T) (a^T x - y) = x^T a a^T x + y^2 - 2y a^T x$$

and differentiating,

$$\nabla_a \| a^T x - y \|_2^2 = (x^T a)^T \nabla_a (a^T x) + \nabla_a (x^T a) (a^T x) - 2y \nabla_a (a^T x)$$

$$= a^T x x^T + (a^T x)^T \nabla_a (x^T a) - 2y x^T$$

$$= a^T x x^T + x^T a x T - 2y x^T$$

$$= 2a^T x x^T - 2y x^T$$

Summing across all samples and applying the critical condition,

$$\sum_i a_i^* x_i x_i^T - y_i x_i^T = 0$$

$$a^* T \sum_i x_i x_i^T = \sum_i y_i x_i^T$$

Letting $M$ be the number of samples in the sequences $\{x_i\}$, $\{y_i\}$, put

$$X = \begin{bmatrix} x_0 & \cdots & x_{M-1} \end{bmatrix}$$

and

$$Y = \begin{bmatrix} y_0 & \cdots & y_{M-1} \end{bmatrix}$$

Eq. 15 can be re-written as

$$a^* T X X^T = Y X^T$$

Essentially we need to compute the sample covariance matrix $U = X X^T$ and the sample cross-covariance matrix $V = Y X^T$. With these two matrices, the solution of

$$a^T U = V$$

is

$$a^* = V U^{-1}$$
provided \( U \) is invertible, i.e. there are at least as many samples as \( k \), the dimension of each \( x_i \).

It is straightforward to show that the same applies when each \( y_i \) is the original in Eq. 2 and \( a^T \) is \( A \), that is,

\[
A^* = VU^{-1} 
\]

Verification of the linear prediction given \( x_i \):

\[
A^* x_i = YX^T(XX^T)^{-1} x_i
\]

\[
= YX^i x_i
\]

The second form is in the form of the Moore-Penrose pseudo-inverse \( X^\dagger \).

The residual error over all samples:

\[
e = \sum \| A^* x_i - y_i \|_2^2 = \text{tr} \left[ (VU^{-1}X - Y)(VU^{-1}X - Y)^T \right]
\]

\[
= \text{tr} \left[ VU^{-1}XX^T U^{-1}V^T + W - 2YX^TU^{-1}V^T \right]
\]

\[
= \text{tr} [W] - \text{tr} [VU^{-1}V^T]
\]

\[
= \text{tr} [W] - \text{tr} [A^*V^T]
\]

where \( W = YY^T \).

2 Incremental regression

The operative quantities are the sample covariance and sample cross-covariance matrices:

The covariances

\[
U \triangleq \sum_i x_i x_i^T = XX^T
\]

\[
W \triangleq \sum_i y_i y_i^T = YY^T
\]

The cross-covariance

\[
V \triangleq \sum_i y_i x_i^T = YX^T
\]

These terms can be computed incrementally. The only non-linear operation is the inversion of \( U \).

The correction to \( A^* = VU^{-1} \) is

\[
\Delta A^* = V\Delta U^{-1} + \Delta VU^{-1} + \Delta V \Delta U^{-1}
\]

\[
\Delta e = y_i^T y_i - \text{tr} [A^* \Delta V^T + \Delta A^* V^T + \Delta A^* \Delta V^T]
\]
Here, $\Delta V = y_i x_i^T$ is an incremental cross-covariance, easily computed. $\Delta U^{-1}$ is more complicated. We appeal to the Sherman-Morrison formula, which states

$$\begin{align*}
(U + x_i x_i^T)^{-1} &= U^{-1} + \Delta U^{-1} \\
&= U^{-1} - \frac{U^{-1} x_i x_i^T U^{-1}}{1 + x_i^T U^{-1} x_i} 
\end{align*}$$

or more generally the Woodbury matrix identity, which states

$$\begin{align*}
(U + XX^T)^{-1} &= U^{-1} + \Delta U^{-1} \\
&= U^{-1} - U^{-1} X (I + X U^{-1} X)^{-1} X^T U^{-1} 
\end{align*}$$

if the increment comprises multiple (e.g. $\Delta M$) samples in $X$ and $Y$.

This method is actually not computationally efficient, but contains the memory usage to the storage of each increment, if that should be a concern.

A way to balance the complexity of the update is the following rank condition heuristic: if $\Delta M \geq k$, then compute $(U + \Delta U)^{-1}$ via direct inversion; if $\Delta M < k$, then compute $U^{-1} + \Delta U^{-1}$ via low-rank update.

## 3 Distributed regression

The incremental version of regression does not lend itself naturally to distribution, since there is serial dependency from one increment to the next.\(^1\)

**Scheme A** — One possible map-reducible scheme is to compute only local $V_s$ and $U_s$, so the final reduction is

$$A^* = \left( \sum s' V_{s'} \right) \left( \sum t' U_{t'} \right)^{-1}$$

If a central node is responsible for dispatching data as well as reduction, we can use the rank condition heuristic to decide whether to perform a central low-rank update, or a map-reduced (i.e. sharded) direct inversion.

**Scheme B** — If shards are allowed direct communication with each other, and synchronization issues are resolved, another possible scheme is to have each shard be responsible for computing all cross terms involving $V_s$, e.g. compute $V_s, U_s$ locally, pull $U_t$ for all $t \neq s$ from other shards, then compute

$$A^*_s = V_s U^{-1} = V_s \left(U_s + \sum_{t \neq s} U_t \right)^{-1}$$

and the final reduction becomes only addition,

$$A^* = \sum_s A^*_s$$

\(^1\)See also [http://www.csee.umbc.edu/~hillol/PUBS/Papers/sdm08_bhaduri.pdf](http://www.csee.umbc.edu/~hillol/PUBS/Papers/sdm08_bhaduri.pdf)
Scheme C — We can combine Scheme B with incrementalism to create a two-phase protocol. In the communication phase, suppose the shards each have $V_s$ and the sum $U = U_s + \sum_{t \neq s} U_t$. They exchange local increments $\Delta U_s$ so that each shard updates to $U := U + \Delta U_s + \sum_{t \neq s} \Delta U_t$. Further each shard updates $A^*_s = V_s U^{-1}$ as before. In the collection phase, each shard receives local increments $X_s, Y_s$ and computes $\Delta V_s, \Delta U_s$. Finally the reduction computes $A^* = \sum_s A^*_s$. If we do not demand consistency, these phases can happen asynchronously. In particular, receiving new data at any shard triggers a collection phase locally, followed by a communication phase at all shards, followed by a reduction.

4 Ridge regression

The prior methods apply verbatim to Ridge regression. To show this, modify canonical regression with a regularization term to obtain the Ridge solution

$$A^* = \arg \min_A \sum \|Ax_i - y_i\|^2 + \lambda \|A\|^2_F$$  (41)

The additional term in the derivative is $2\lambda \sum A_{uv}$, giving the modified critical condition (c.f. Eq. 15)

$$a^*^T \sum_i x_i x_i^T = \sum_i y_i x_i^T - \lambda a^*^T$$  (42)

The single dimension case reduces to

$$a^*^T \left( \lambda + \sum_i x_i x_i^T \right) = \sum_i y_i x_i^T$$  (43)

The multidimensional case is similar

$$A^* \left( \lambda I_k + \sum_i x_i x_i^T \right) = \sum_i y_i x_i^T$$  (44)

Thus the solution (Eq. 21) is only slightly perturbed, by replacing $U = XX^T$ with

$$U = \lambda I_k + XX^T$$  (45)

which is to say, that the covariance of $\{x_i\}_i$ is augmented by a self-variance. In the limit of large $\lambda$, we have

$$A^* = VU^{-1} \approx V\lambda^{-1} \rightarrow 0$$  (46)

Consider $\lambda I_k$ as the prior covariance assumption, worth exactly $\lambda$ additional white samples of $x$ with unit per-dimension variance and uncorrelated with corresponding $y$ samples (hence nothing on RHS of Eq. 44). Adding uncorrelated
samples to the data has the effect of mixing a null model ($A = 0$) into the solution in the absence of data — also ensuring $U$ is invertible.

The residual error of Eq. (28) is also slightly changed

$$e = \sum_i \|A^* x_i - y_i\|_2^2$$

(47)

$$e = \text{tr} [VU^{-1}XX^TU^{-1}V^T + W - 2XY^TU^{-1}V^T]$$

(48)

$$e = \text{tr} [VU^{-1}(U - \lambda I_k)U^{-1}V^T + W - 2VU^{-1}V^T]$$

(49)

$$e = \text{tr} [W] - \text{tr} [A^*V^T] - \lambda \|A^*\|_F^2$$

(50)

i.e., there is an additional term due to regularization. If we retain $\text{tr} [W] - \text{tr} [A^*V^T] = e + \lambda \|A^*\|_F^2$ as the error expression (the Ridge error), then it includes the model penalization term automatically, useful against overfitting.

Since the only modification from canonical regression is the computation of $U$, the incremental and distributed versions are the same except for initialization.

5 Miscellany

5.1 Cholesky decomposition

The matrix equation $A^* = VU^{-1}$ can be solved less expensively without matrix inversion by using a (pre-computed) Cholesky decomposition of $U = LL^T$, from which a dedicated Cholesky solver can solve

$$LL^T A^*T = V^T$$

(51)

for $A^*$. There are also low-rank updates to the Cholesky expression directly.\(^2\)

5.2 Low-rank downdate

The Sherman-Morrison and Woodbury identities can be modified to allow downdates, by switching two signs:

$$\left(U - xx_i^T\right)^{-1} = U^{-1} + \Delta U^{-1} = U^{-1} + \frac{U^{-1}xx_i^TU^{-1}}{1 - x_i^TU^{-1}x_i}$$

(52)

and

$$\left(U - XX^T\right)^{-1} = U^{-1} + \Delta U^{-1} = U^{-1} + U^{-1}X(I_{\Delta M} - X^TU^{-1}X)^{-1}X^TU^{-1}$$

(53)

5.3 Intercept

If \( \{x_i\}_i, \{y_i\}_i \) have non-zero means \( \bar{x}, \bar{y} \), then the means can be removed first. Or, we can compute the intercept \( b^* \) from the relation

\[
y_i - \bar{y} = A^*(x_i - \bar{x}) + n_i
\]

namely

\[
b^* = \bar{y} - A^*\bar{x}
\]

The estimates for \( A^* \) and \( e \) will also be computed from the mean-removed moments instead, i.e.

\[
U = (X - \bar{X})(X - \bar{X})^T = XX^T - M\bar{x}\bar{x}^T
\]

\[
V = (Y - \bar{Y})(X - \bar{X})^T = YX^T - M\bar{y}\bar{x}^T
\]

\[
W = (Y - \bar{Y})(Y - \bar{Y})^T = YY^T - M\bar{y}\bar{y}^T
\]

where \( \bar{X} \triangleq [\bar{x} \cdots \bar{x}] \) and \( \bar{Y} \triangleq [\bar{y} \cdots \bar{y}] \) are of dimensions \( k \times M \) and \( n \times M \), respectively.