Lecture 10

Partial least squares

Let $x \in \mathbb{R}^m$ and $y \in \mathbb{R}$ be two zero-mean random variables. Denote $\Sigma = \text{cov}(x)$ and $\delta = \text{cov}(x, y)$. Assume $y = x\beta + e$, $\text{E}(e) = 0$ and $\text{cov}(e) = \sigma^2$. This is called a population model.

Now, let $Y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$ be a vector of n independent copies of y and $X = (x_1^t, \ldots, x_n^t)^t \in \mathbb{R}^{n \times m}$ be a matrix of n independent copies of x. In particular, $Y = X\beta + \epsilon$, where $\text{E}(\epsilon) = 0_n$ and $\text{cov}(\epsilon) = \sigma^2 I_n$.

Denote $A = X^t X$ and $b = X^t Y$; note that A and b are proportional to the sample estimators of Σ and δ, respectively.

The ordinary least squares estimator (OLS) of β is defined as

$$\hat{\beta}_{LS} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2.$$

Solution of this problem is equivalent to the solution of the normal equations $(X^t X)\beta = X^t Y$, or in our notation $A\beta = b$.

If A is ill-conditioned (e.g., if some columns in X are (nearly) linear dependent), then one can define an OLS estimator via the pseudoinverse of A. Let $X = V\Lambda S^t$, for $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and orthogonal matrices V and S, be the singular value decomposition of X, so that $X^t X = A = S\Lambda^2 S^t$. Then,

$$\hat{\beta}_{LS} = A^- b = \sum_{i=1}^{rk(A)} V_i^t Y S_i / \lambda_i =: \sum_{i=1}^{rk(A)} p_i.$$

There are many methods that provide a regularised solution to $A\beta = b$ in case of an ill-conditioned matrix A. For example, principal component regression derives an estimator for β as

$$\hat{\beta}_{PC}^d = \sum_{i=1}^{d} p_i,$$

where $d \leq rk(A)$ should be chosen data-driven. In ridge regression regularisation depends on a parameter γ, which also has to be chosen data-driven:

$$\hat{\beta}_{RR} = \sum_{i=1}^{rk(A)} \frac{\lambda_i^2}{\lambda_i^2 + \gamma} p_i.$$

Another approach offers the so-called partial least squares algorithm which is closely related to the conjugate gradient algorithm for the solution of $A\beta = b$. Recall that according to Cayley-Hamilton theorem every square matrix satisfies its characteristic equation. That is, if $P(x) = |xI_n - A|$, then $P(A) = 0$. From this one finds that

$$A^{-1} = (-1)^{m-1} \sum_{i=1}^{m} c_i A^{i-1}$$

for suitable $c_i \in \mathbb{R}$ and hence, $\hat{\beta}_{LS} = \sum_{i=1}^{m} (-1)^{m-1} c_i A^{i-1} b$.
Denote now $K_d(b, A) = \text{span}(b, Ab, \ldots, A^{d-1}b)$ the d-dimensional Krylov space. Then, the partial least squares estimator can be defined as

$$\hat{\beta}^d_{PLS} = \arg\min_{\beta \in K_d} \|Y - X\beta\|^2,$$

which is a regularised version of $\hat{\beta}_{LS} = \sum_{i=1}^{m} \tilde{c}_i A^{i-1} b$ for $d < m$ for ill-conditioned A.

Another view on principal component regression and partial least squares allows to highlight the differences in the performance of both algorithms.

Assume we would like to find such a linear combinations of entries of the vector x, that has a maximal covariance. That is, we are looking for

$$\alpha_1 = \arg\max_{\|\alpha\|=1} \text{cov}(x^t \alpha).$$

This problem is equivalent to maximisation of $\alpha \text{cov}(x) \alpha^t - \lambda(\alpha^t \alpha - 1)$ or to the solution to $\{\text{cov}(x) - \lambda I_m\}\alpha = 0$. Hence, α_1 is the eigenvector corresponding to the maximal eigenvalue of $\text{cov}(x) = \Sigma$. Next α_2 should also maximise $\text{cov}(x^t \alpha)$ subject to $\alpha_2 \perp \alpha_1$, which is the eigenvector, corresponding to the second largest eigenvalue. Replacing x by xH, where $H = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^{m \times d}$ we obtain

$$\beta^d_{PC} = H \arg\min_{\xi \in \mathbb{R}^d} \mathrm{E}\|y - x^t H \xi\|^2 = H(H^t \Sigma H)^{-1} H^t \delta.$$

That is, we regress on those linear combinations of x that have the largest covariance.

Since in practice instead of x we have a matrix X of n independent realisations of x and $A = X^t X = S \Lambda^2 S^t$ is proportional to the estimator of Σ, to get the sample version

$$\hat{\beta}^d_{PC} = S_d (S_d^t X^t X S_d)^{-1} S_d^t X^t Y sm_{i=1}^d p_i,$$

where S_d denotes the first d columns of the matrix of eigenvalues S.

Partial least squares follow another route. In the population model we are looking for such linear combination of entries of the vector x, which have the largest covariance between x and y. That is,

$$\alpha_1 = \arg\max_{\|\alpha\|=1} \operatorname{cov}(x^t\alpha, y)^2 \propto x^t y$$

Next, we get $y - x^t\beta^1_{PLS}$, where $\beta^1_{PLS} = \alpha_1(\alpha_1^t\Sigma\alpha_1)^{-1}\alpha_1\delta$ and look for

$$\alpha_2 = \arg\max_{\|\alpha\|=1} \operatorname{cov}(x^t\alpha, y - x^t\beta^1_{PLS})^2.$$

By definition, $\alpha_2 \perp \alpha_1$. Further orthogonal components are found in the same way. It is easy to show that $\alpha_i \in K_i(\delta, \Sigma)$.

Again, to obtain a sample version of PLS, the population covariances Σ and δ are replaced by their sample versions A and b, respectively, yielding $\hat{\beta}^d_{PLS}$. Note that PLS estimators are invariant to scaling of A and b.

Obviously, the calculation of the PLS estimator using its definition via Krylov spaces is highly numerically unstable. In practice, there are several PLS algorithms developed, which are all shown to be equivalent. In fact, PLS was developed by scientists working in chemometrics as an algorithm first and links of PLS to conjugate gradient and Krylov spaces were established later.

It is assumed that

$$\begin{aligned} X &= TP^t + E \\ Y &= UQ^t + F, \end{aligned}$$

where $X \in \mathbb{R}^{n\times m}$, $Y \in \mathbb{R}^{n\times 1}$ (can be $n \times k$), $T, U \in \mathbb{R}^{n\times d}$ are scores and $P \in \mathbb{R}^{m\times d}$, $Q \in \mathbb{R}^{1\times d}$ are loadings. Matrices E and F have the same dimensions as X, Y, respectively and are residuals.
NIPALS algorithm is given as

1. Initialise $X_0 = X$, $Y_0 = Y$
2. Repeat until convergence of q_i, for $i = 1, 2, \ldots$
   (a) $w_i = X^t_{i-1}Y_{i-1}/(Y^t_{i-1}Y_{i-1})$
   (b) Normalise w_i to 1

(c) $t_i = X_{i-1} w_i$

(d) $q_i = X_{i-1}^t t_i / (t_i^t t_i)$

(e) $X_i = X_{i-1} - t_i q_i^t$

(f) $c_i = Y_{i-1}^t t_i / (t_i^t t_i)$

(g) Normalise c_i to 1

(i) $Y_i = Y_{i-1} - c_i t_i^t$

Another well-known algorithm is SIMPLS by de Jong.

The choice of d is important in practice. Too small d might lead to a high bias in the model (some features are not captured), while too large d might increase the variance. A typical choice is the cross-validation.

Let $\kappa : \{1, \dots, n\} \mapsto \{1, \dots, K\}$ be an indexing function that indicates the partition to which observation i is allocated by randomisation. Denote $\hat{f}^{-k}(d)$ the model fit, calculated without the kth part of the data. For example, for PLS regression $\hat{f}(d) = X\hat{\beta}^d_{PLS}$ and $\hat{f}(x_i; d) = X_i\hat{\beta}^d_{PLS}$. The K-fold cross-validation is defined as

$$CV(d) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{f}^{-\kappa(i)}(x_i; d)\}^2.$$

If $K = N$, the this criterion is known as leave-one-out cross-validation. In this case $\kappa(i) = i$ and for the ith observation the fit is computed using all the data but the ith pair (X_i, Y_i). Another typical choice for K is 5 or 10. While leave-one-out CV with $K = N$ is (asymptotically) unbiased, it is more computationally intensive and has a large variance. CV with $K = 5$ or 10 are known to be biased, but have considerably smaller variance and are faster to calculate.