EM for LS

Given a set of data samples $\{X_i, y_i\}, \forall i \in \{1, ..., n\}$, where $X_i \in R^m$, we desire to fit the relation $y_i = X_i w + \epsilon_i$, which is trained by minimizing the least-squared error.

$$\min \sum_i (y_i - X_i w)^2$$

The quality of each data sample are not consistent, and there may exist anomaly in the data samples. We would like to assign different importance to different data samples, which can be implemented by assigning weights in regression loss.

$$\min \sum_i \alpha_i (y_i - X_i w)^2$$

How to compute α_i? We assume $\epsilon_i \sim N(0, \lambda_i^2 \sigma^2)$, such that the error of each data sample yields a Gaussian distribution with distinct error. Then the data sample $\{X_i, y_i\}$ follows the distribution $y_i \sim N(X_i w, \lambda^2 \sigma^2)$. We desire to optimize the likelihood

$$l(w) = \prod_i p(y_i|w) = \prod_i \frac{1}{\sqrt{2\pi}\lambda_i \sigma} \exp\left[-\frac{(y_i - X_i w)^2}{2\lambda_i^2 \sigma^2}\right]$$

Maximize the logarithm of $l(w)$ is equivalent to minimizing $\sum_i \frac{1}{\lambda_i^2}(y_i - X_i w)^2$, thus, $\alpha_i = \frac{1}{\lambda_i^2}$

As the parameters of distribution $\epsilon_i \sim N(0, \lambda_i^2 \sigma^2)$ cannot be known in advance, we need to optimize the likelihood in a EM fashion while considering $\lambda_i^2 \sigma^2$ as a hidden variable.

Given an initial group of model coefficients w^0 (can be computed using OLS), we can estimate the variances of each data sample $\lambda_i^2 \sigma^2$ and update the weight for next iteration.

How to compute $\lambda_i^2 \sigma^2$? Since it cannot be estimated using a single data sample, we can run a clustering method to group the data samples (based on the residual of each data sample) and those in the same group are assigned the same variance. Or we can start from a unified variance and the variance of each data sample is adapted in each iteration.

EM for TLS

In the setup of TLS, we desire to fit the relation $y_i + r_i = (X_i + E_i)w$, and we let $\widehat{X}_i = X_i + E_i$, is the estimated value of X_i. We consider both the error for estimating y_i as r_i and the error for measuring X_i as E_i. We assume r_i and E_{ij}, the j-th element of E_i, follow the Gaussian distribution $N(0, \sigma^2)$, the probability of data sample i conditioned on the model w is $p(X_i, y_i|w) = p(y_i|w) \prod_j P(X_{ij}|w)$, where $p(y_i|w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \widehat{X}_i w)^2}{2\sigma^2}\right]$ and $p(X_{ij}|w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(X_{ij} - \widehat{X}_{ij})^2}{2\sigma^2}\right]$.

Thus, the likelihood is :

$$l(w) = \prod_i p(y_i|w) \times \prod_i \prod_j p(X_{ij}|w)$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n(m+1)} \exp\left[-\frac{\sum_i (y_i - \widehat{X}_i w)^2 + \sum_i \sum_j (X_{ij} - \widehat{X}_{ij})^2}{2\sigma^2}\right]$$

Similarly, we can assume that r_i and E_{ij} follow $N(0, \lambda_i^2\sigma^2)$. Thus, the likelihood is updated to

$$l(w) = \prod_i p(y_i|w) \times \prod_i \prod_j p(X_{ij}|w)$$
$$= \prod_i \left[\frac{1}{\sqrt{2\pi}\lambda_i\sigma} \exp\left[-\frac{(y_i - \widehat{X}_i w)^2}{2\lambda_i^2\sigma^2}\right] \times \prod_j \frac{1}{\sqrt{2\pi}\lambda_i\sigma} \exp\left[-\frac{(X_{ij} - \widehat{X}_{ij})^2}{2\lambda_i^2\sigma^2}\right]\right]$$
$$= \prod_{\mathrm{i}} \left[\frac{1}{(\sqrt{2\pi}\lambda_i\sigma)^{m+1}} \exp\left[-\frac{(y_i - \widehat{X}_i w)^2}{2\lambda_i^2\sigma^2}\right] \times \exp\left[-\frac{\sum_j (X_{ij} - \widehat{X}_{ij})^2}{2\lambda_i^2\sigma^2}\right]\right]$$
$$= \prod_{\mathrm{i}} \left\{\frac{1}{(\sqrt{2\pi}\lambda_i\sigma)^{m+1}} \exp\left[-\frac{(y_i - \widehat{X}_i w)^2 + \sum_j (X_{ij} - \widehat{X}_{ij})^2}{2\lambda_i^2\sigma^2}\right]\right\}$$

Maximize the log-likelihood is equivalent to minimize

$$\sum_{\mathrm{i}} \frac{1}{\lambda_i^2}\left[r_i^2 + \sum_j E_{ij}^2\right]$$

Given a w, use KKT to compute $r = \frac{Xw-y}{\|w\|_2^2+1}$, $E = -\frac{(Xw-y)w^T}{\|w\|_2^2+1}$. We can thus estimate $\lambda_i^2\sigma^2$ for each data sample and use it to optimize w in next iteration. Given λ_i^2, how to compute w? Just scale each data sample by multiplying both X_i and y_i to $\frac{1}{\lambda_i}$

Different Noise Level on Each Column

We assume r_i, follow the Gaussian distribution $N(0, \sigma_y^2)$, and E_{ij} follows the Gaussian distribution $N\left(0, \sigma_{X_j}^2\right)$. the probability of data sample i conditioned on the model w is $p(X_i, y_i|w) = p(y_i|w)\prod_{\mathrm{j}} P(X_{ij}|w)$, where $p(y_i|w) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left[-\frac{(y_i - \widehat{X}_i w)^2}{2\sigma_y^2}\right]$ and $p(X_{ij}|w) = \frac{1}{\sqrt{2\pi}\sigma_{X_j}} \exp\left[-\frac{(X_{ij} - \widehat{X}_{ij})^2}{2\sigma_{X_j}^2}\right]$.

Thus, the likelihood is :

$$l(w) = \prod_i p(y_i|w) \times \prod_i \prod_j p(X_{ij}|w)$$

$$= \prod_{\mathrm{i}} \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left[-\frac{\left(y_i - \widehat{X}_i w\right)^2}{2\sigma_y^2}\right] \times \prod_{\mathrm{i}} \prod_j \frac{1}{\sqrt{2\pi}\sigma_{X_j}} \exp\left[-\frac{\left(X_{ij} - \widehat{X}_{ij}\right)^2}{2\sigma_{X_j}^2}\right]$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_y^n} \exp\left[-\frac{\sum_i \left(y_i - \widehat{X}_i w\right)^2}{2\sigma_y^2}\right] \times \prod_{\mathrm{j}} \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_{X_j}^n} \exp\left[-\frac{\sum_i \left(X_{ij} - \widehat{X}_{ij}\right)^2}{2\sigma_{X_j}^2}\right]$$

Maximize this likelihood is equivalent to minimize

$$\sum_i \frac{\sum_i \left(y_i - \widehat{X}_i w\right)^2}{\sigma_y^2} + \sum_j \frac{\sum_i \left(X_{ij} - \widehat{X}_{ij}\right)^2}{\sigma_{X_j}^2}$$

The corresponding optimization problem is

$\min_{\mathrm{w,E,r}} \frac{1}{\sigma_y^2}\|r\|_2^2 + \sum_{\mathrm{j}} \frac{1}{\sigma_{X_j}^2}\left\|E_j\right\|_2^2$, s.t. $(X+E)w = y + r$

We consider two scenarios:

When the variances of all columns in $\boldsymbol{X}$ are identical such that $\boldsymbol{\sigma_{X_j}^2 = \sigma_X^2}$

Let $\eta = \frac{\sigma_{X_j}^2}{\sigma_y^2}$, $\min_{\mathrm{w,E,r}} \|E\|_F^2 + \eta\|r\|_2^2$, s.t. $(X+E)w = y + r$

$L(E, r, \lambda) = \| E\|_F^2 + \eta \| r\|_2^2 + \lambda[(X+E)w - y - r],\ \lambda \in R^m$

KKT Condition: $2E + \lambda w^T = 0$ (1), $2\eta r - \lambda = 0$ (2), $(X+E)w = y + r$ (3)

From (1) and (2): $2\eta r w^T - \lambda w^T = 0$, $2\eta r w^T + 2E = 0$, $E = -\eta r w^T$ (4)

From (3) and (4): $(X - \eta r w^T)w = y + r$, $Xw - y = \eta r w^T w + r$, $r = \frac{Xw - y}{\eta\|w\|_2^2 + 1}$ (5)

From (4) and (5): $E = -\frac{\eta(Xw-y)w^T}{\eta\|w\|_2^2+1}$ (6)

Put (5) and (6) into the objective function:

$$\| E\|_F^2 + \eta \| r\|_2^2 = \frac{\eta^2\|Xw - y\|_2^2\|w\|_2^2 + \eta\|Xw - y\|_2^2}{\left(\eta\|w\|_2^2 + 1\right)\left(\eta\|w\|_2^2 + 1\right)} = \eta\frac{\eta\|Xw - y\|_2^2\|w\|_2^2 + \|Xw - y\|_2^2}{\left(\eta\|w\|_2^2 + 1\right)\left(\eta\|w\|_2^2 + 1\right)}$$

$$= \eta\frac{\left(\eta\|w\|_2^2 + 1\right)\|Xw - y\|_2^2}{\left(\eta\|w\|_2^2 + 1\right)\left(\eta\|w\|_2^2 + 1\right)} = \frac{\eta\|Xw - y\|_2^2}{\eta\|w\|_2^2 + 1} = \frac{\|Xw - y\|_2^2}{\|w\|_2^2 + \frac{1}{\eta}}$$

$\min_{\mathrm{x,E,r}} \|E\|_F^2 + \eta\|r\|_2^2,\ s.t.\ (X+E)w = y + r$ -> $\min_{\mathrm{w}} \frac{\|Xw-y\|_2^2}{\|w\|_2^2 + \frac{1}{\eta}}$ -> $\min_{\mathrm{w}} \frac{\left\|X(\sqrt{\eta}w) - \sqrt{\eta}y\right\|_2^2}{\left\|\sqrt{\eta}w\right\|_2^2 + 1}$

Compute $\sqrt{\eta}w$ by solving the standard TLS problem, and then compute w

Initialize w and

E: Estimate $\boldsymbol{\sigma_X^2}$ and σ_y^2, $E = -\frac{\eta(Xw-y)w^T}{\eta\|w\|_2^2+1}$, $r = \frac{Xw-y}{\eta\|w\|_2^2+1}$, compute η

M: Compute w based on given η

When the variances of all columns in $\boldsymbol{X}$ are not assumed to be identical

$\min_{\mathrm{w,E,r}} \frac{1}{\sigma_y^2}\|r\|_2^2 + \sum_{\mathrm{j}} \frac{1}{\sigma_{X_j}^2}\left\|E_j\right\|_2^2$, s.t. $(X+E)w = y + r$

$\min_{\mathrm{w,E,r}} \|r\|_2^2 + \sum_{\mathrm{j}} \frac{\sigma_y^2}{\sigma_{X_j}^2}\left\|E_j\right\|_2^2$, s.t. $(X+E)w = y + r$

$\min_{\mathrm{w,E,r}} \|r\|_2^2 + \|E\Sigma_x\|_F^2$, s.t. $(X+E)w = y + r,\ \Sigma_x = diag\left[\frac{\sigma_y}{\sigma_{X_1}}, \ldots, \frac{\sigma_y}{\sigma_{X_m}}\right]$

$L(E, r, \lambda) = \|r\|_2^2 + \|E\Sigma_x\|_F^2 + \lambda[(X+E)w - y - r],\ \lambda \in R^m$

KKT:

$2E\Sigma_x^2 + \lambda w^T = 0$ (1)

$2r - \lambda = 0$ (2)

$(X+E)w = y + r$ (3)

From (1) and (2): $2rw^T - \lambda w^T = 0$, $2E\Sigma_x^2 + 2rw^T = 0$, $E = -rw^T\Sigma_x^{-2}$ (4)

From (3) and (4): $(X - rw^T\Sigma_x^{-2})w = y + r$, $Xw - rw^T\Sigma_x^{-2}w = y + r$, $Xw - y = r + w^T\Sigma_x^{-2}wr = (1 + w^T\Sigma_x^{-2}w)r$, $r = \frac{Xw-y}{1+w^T\Sigma_x^{-2}w}$ (5)

From (4) and (5): $E = -\frac{(Xw-y)w^T\Sigma_x^{-2}}{1+w^T\Sigma_x^{-2}w}$ (6)

Put (5) and (6) into the objective function:

$$\|r\|_2^2 + \|E\Sigma_x\|_F^2 = \frac{\|Xw-y\|_2^2 + \left\|(Xw-y)w^T\Sigma_x^{-1}\right\|_F^2}{(1+w^T\Sigma_x^{-2}w)(1+w^T\Sigma_x^{-2}w)} = \frac{\|Xw-y\|_2^2(1+w^T\Sigma_x^{-2}w)}{(1+w^T\Sigma_x^{-2}w)(1+w^T\Sigma_x^{-2}w)}$$

$$= \frac{\|Xw-y\|_2^2}{(1+w^T\Sigma_x^{-2}w)} = \frac{\left\|X\Sigma_x\Sigma_x^{-1}w - y\right\|_2^2}{(1+w^T\Sigma_x^{-2}w)}$$

Let $\Sigma_x^{-1}w = w'$, the problem is equivalent to $\min \frac{\|X\Sigma_x w' - y\|_2^2}{\|w'\|_2^2 + 1}$

Initialize w and Σ_x

E: Estimate variances $r = \frac{Xw-y}{1+w^T\Sigma_x^{-2}w}$, $E = -\frac{(Xw-y)w^T\Sigma_x^{-2}}{1+w^T\Sigma_x^{-2}w}$, update Σ_x

M: Compute w based on current Σ_x