확률론적 다변량 회귀분석

김동순, 이인범
포항공과대학교 환경공학부

Probabilistic Latent Score Regression

Dongsoon Kim, In-Beum Lee
School of Environmental Science and Engineering, POSTECH, South Korea

Introduction

There can be a bunch of measurements in a process, for instance of activated sludge process, BOD, COD, temperature, SVI, DO, MLSS, turbidity, color, etc. Among the measurements some are easily measurable while the others are not, e.g. BOD needs 5 days while DO for every minute. Multivariate regression method is favorable candidate to overcome the time mismatch. If there is a relation between the readily and hardly measurements, combination of the handies can be used to predict the nuisances. This research suggests a probabilistic method for the regression in which holds two critical concepts: the *latent variable* called hidden, caused, principal component or factor to represents condition of the process; and the *probabilistic reasoning* to interpret the regression results. Combining them enable engineers to analyze the process by substitution the headaches for handies.

Theory

Let's consider the standard regression formula as Eq. (1).

$$y = \boldsymbol{c}^{\mathrm{T}}\cdot\boldsymbol{z} + v \qquad (1)$$

where regressor variable $\boldsymbol{z} \in \Re^{L} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_z)$ and response variable $y \in \Re^{1} \sim \mathcal{N}(0, \lambda_y)$ are assumed. The best linear unbiased estimator (BLUE) of $\boldsymbol{c}$ is the least-square estimator (LSE).

$$\boldsymbol{c}^{\mathrm{T}}_{\mathrm{LS}} = \boldsymbol{y}\cdot\mathbf{Z}^{\mathrm{T}}\cdot(\mathbf{Z}\cdot\mathbf{Z}^{\mathrm{T}})^{-1} = \boldsymbol{y}\cdot\mathbf{Z}^{+} \qquad (2)$$

where $\boldsymbol{y} = \{y^{(n)}\}$ and $\mathbf{Z} = \{\boldsymbol{z}^{(n)}\}$ for sample number $n \in \{1,\ldots,N\}$, and superscript '+' represents the Moore-Penrose generalized matrix inverse. Note that it is the result of an optimization problem, i.e. $\boldsymbol{c}_{\mathrm{LS}} = \arg_{\boldsymbol{c}}$ min: $\lambda_v = \langle(y - \boldsymbol{c}^{\mathrm{T}}\cdot\boldsymbol{z})^2\rangle$. When the LSE was used, regression error is to be $v \sim \mathcal{N}(0, \lambda_v)$ since Gaussianity is closed for linear operation, and regressed $y = \boldsymbol{c}_{\mathrm{LS}}{}^{\mathrm{T}}\cdot\boldsymbol{z}$. Furthermore, if $\lambda_v = \langle(y - \boldsymbol{c}^{\mathrm{T}}{}_{\mathrm{LS}}\cdot\boldsymbol{z})^2\rangle \leq \delta\cdot\lambda_y$ for $\delta \in (0, 1)$ then y is regressible by $\boldsymbol{c}^{\mathrm{T}}{}_{\mathrm{LS}}\cdot\boldsymbol{z}$ with $r^2 = (1-\delta)$ regressibility. Hence the absorption ratio of λ_y by $\boldsymbol{c}_{\mathrm{LS}}{}^{\mathrm{T}}\cdot\boldsymbol{z}$ is expressed by Eq.(3).

$$r^2 = \boldsymbol{y}\cdot\mathbf{Z}^{+}\cdot\mathbf{Z}\cdot\boldsymbol{y}^{+} \qquad (3)$$

where $0 \leq r^2 \leq 1$. Note that $r^2 = 1$ indicates $\lambda_v = 0$, and hence no estimation errors. H-principal emphasizes that $\boldsymbol{c}_{\mathrm{LS}}$ should be balanced between minimizing λ_v and is robust. The robustness of $\boldsymbol{c}_{\mathrm{LS}}$ is checked by the condition number of $\mathbf{Z}$, denoted by $\eta_{\mathbf{Z}}$, because Euclidian norm of it indicates $\|\boldsymbol{c}_{\mathrm{LS}}\|_{\mathrm{E}}^2 = y\cdot\mathbf{Z}^{\mathrm{T}}\cdot(\mathbf{Z}\cdot\mathbf{Z}^{\mathrm{T}})^{-2}\cdot\mathbf{Z}\cdot y^{\mathrm{T}}$. Thus it is reasonable to say that "y is regressible by $\boldsymbol{c}^{\mathrm{T}}{}_{\mathrm{LS}}\cdot\boldsymbol{z}$ with r^2 regressibility, and if $\eta_{\mathbf{Z}} \le \Delta$ for a large Δ, then $\boldsymbol{c}_{\mathrm{LS}}$ is robust".

Various multivariate calibration methods

All measurements $\boldsymbol{x} \in \Re^P \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_x)$ can be used for the regressor variable $\boldsymbol{z}$. It is the well-known multiple linear regression (MLR) method. Let's denote the regression coefficient vector of $\boldsymbol{x}$ as $\boldsymbol{b}$. Then $\boldsymbol{b} = \boldsymbol{c}_{\mathrm{LS}}$, and hence $\underline{y} = \boldsymbol{b}^{\mathrm{T}}\cdot\boldsymbol{x}$ in MLR. Additionally, suppose a unitary matrix $\mathbf{P}$ that rotates $\boldsymbol{z}$, and the rotation result is $\boldsymbol{x}$, e.g. $\boldsymbol{x} = \mathbf{P}\cdot\boldsymbol{z}$ where $\mathbf{P}^{\mathrm{T}}\cdot\mathbf{P} = \mathbf{I}_P$. Then $\boldsymbol{z}$ can be recovered by latent score filter $\mathbf{Q} = \mathbf{P}^{-1} = \mathbf{P}^{\mathrm{T}}$ such as $\underline{\boldsymbol{z}} = \mathbf{Q}\cdot\boldsymbol{x}$, and hence Eq. (4) represents.

$$\boldsymbol{b}^{\mathrm{T}} = \boldsymbol{c}^{\mathrm{T}}{}_{\mathrm{LS}}\cdot\mathbf{Q} \tag{4}$$

Suppose $\boldsymbol{x} \in \Re^P = \mathbf{P}\cdot\boldsymbol{z} + \boldsymbol{e}$, where $\boldsymbol{z} \in \Re^L \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_z)$, $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$, $\mathbf{P}^{\mathrm{T}}\cdot\mathbf{P} = \mathbf{I}_L$, and $L \le P$. It implies a high dimensional measurement vector is the results of a transform of a low dimensional latent vector. If $(P\text{–}L)$ elements of which small variances are eliminated from $\boldsymbol{x}$, the robustness of $\boldsymbol{c}_{\mathrm{LS}}$ is guaranteed, i.e. $\eta_{\mathbf{Z}} \le \Delta$. In this case, the hidden signals are recovered by Eq. (5).

$$\underline{\boldsymbol{z}} = \mathbf{Q}\cdot\boldsymbol{x} \text{ and } \underline{\boldsymbol{e}} = \mathbf{W}\cdot\boldsymbol{x} \tag{5}$$

where $\mathbf{Q} = \mathbf{P}^+ = (\mathbf{P}^{\mathrm{T}}\cdot\mathbf{P})^{-1}\cdot\mathbf{P}^{\mathrm{T}} = \mathbf{P}^{\mathrm{T}}$ and $\mathbf{W} = (\mathbf{I} - \mathbf{P}\cdot\mathbf{P}^{\mathrm{T}})$. Note that $\boldsymbol{\Sigma}_z = \mathbf{Q}\cdot\boldsymbol{\Sigma}_x\cdot\mathbf{Q}^{\mathrm{T}}$ and $\boldsymbol{\Sigma}_e = \mathbf{W}\cdot\boldsymbol{\Sigma}_x\cdot\mathbf{W}^{\mathrm{T}}$. Therefore y is regressible by $\boldsymbol{b}^{\mathrm{T}}\cdot\boldsymbol{x}$ with $r^2(L) = \boldsymbol{y}\cdot(\mathbf{P}^{\mathrm{T}}\cdot\mathbf{X})^+\cdot(\mathbf{P}^{\mathrm{T}}\cdot\mathbf{X})\cdot\boldsymbol{y}^+$ regressibility. Note that $r^2(i) \le r^2(j)$ for $i < j$, $r^2(P) = \boldsymbol{y}\cdot\mathbf{X}^+\cdot\mathbf{X}\cdot\boldsymbol{y}^+$, and $L = \arg_l \min: |\, r^2{}_{\mathrm{desire}} - r^2(l)\,|$. (See also Figure 1).

If an orthogonal basis set $\mathbf{P}$ were set, then $\mathbf{Q}$, $\mathbf{W}$, $\boldsymbol{c}^{\mathrm{T}}$ and $\boldsymbol{b}^{\mathrm{T}}$ are uniquely determined, and $\underline{\boldsymbol{z}}$ and $\underline{\boldsymbol{e}}$ are found from $\mathbf{Q}$ and $\mathbf{W}$, respectively. There is an abundance methods to find $\mathbf{P}$, e.g. $\mathbf{P}$ = any unitary matrix is MLR, $\mathbf{P} = \{\boldsymbol{u}^{(l)}\}$ is PCR, $\mathbf{P} = \{\boldsymbol{g}^{(l)}\}$ is PLS1, where $\boldsymbol{u}^{(l)}$ and $\boldsymbol{g}^{(l)}$ are the l^{th} left singular vectors of $\mathbf{X}$, and PLS basis vector of $\mathbf{X}$, respectively. CPR finds $\mathbf{P}$ by input modifying $\mathbf{X}_\alpha = \mathbf{U}\cdot\mathbf{S}^\alpha\cdot\mathbf{V}^{\mathrm{T}}$ to PLS algorithm, and it results MLR if $\alpha = 0$, PLS1 if $\alpha = 1$, PCR if $\alpha \approx \infty$. CSR obtains $\mathbf{P}$ by running PLS algorithm with approximated $\mathbf{X}_L{}^J$, it represents MLR if $L = J = P$, PLS1 if $L = P$, PCR if $L = J$. Refer to [1].

PPCR calibration method

Probabilistic principal component regression (PPCR) has its foundation on probabilistic PCA (PPCA) proposed by [2]. It has a model that $\boldsymbol{x} = \mathbf{P}\cdot\boldsymbol{z} + \boldsymbol{e}$, where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})$ and $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \lambda\cdot\mathbf{I})$. PPCA seeks to find the most probable parameter set $\theta = \{\mathbf{P}, \lambda\}$ in the model under given experience $\mathbf{X}$ by the expectation and maximization (EM) algorithm [3]. In brief, EM is an iterative algorithm that maximizes the complete data log likelihood function. Let's denote log likelihood of the i^{th} θ as $\mathcal{L}(\theta_i) =$

$\log\{\mathcal{P}(\mathbf{X}\,|\theta_i)\}$, and its difference for a new estimate as $\Delta\mathcal{L} = \mathcal{L}(\theta) - \mathcal{L}(\theta_i)$. Then $\Delta\mathcal{L}(\theta) = \log\int \mathcal{P}(z\,|\,\mathbf{X},\theta_i)\cdot\mathcal{P}(z, \mathbf{X}\,|\theta)\cdot\mathcal{P}(z, \mathbf{X}\,|\theta_i)^{-1}\,dz$ in which contains the probability density information of latent variable. EM optimize the lower bound of $\Delta\mathcal{L}(\theta)$, that is $Q(\theta) = \int \mathcal{P}(z\,|\,\mathbf{X},\theta_i)\cdot\log\{\mathcal{P}(z, \mathbf{X}\,|\theta)\cdot\mathcal{P}(z, \mathbf{X}\,|\theta_i)^{-1}\}\,dz$, instead of $\Delta\mathcal{L}(\theta)$ itself since $0 = Q(\theta_i\,|\theta_i) \le Q(\theta_{i+1}|\theta_i) \le \mathcal{L}(\theta_{i+1}) - \mathcal{L}(\theta_i) = \Delta\mathcal{L}$. It is the reason that EM can never decrease the log likelihood as iteration proceeds. The optimum is calculated by both solving $(\partial/\partial\mathbf{P})\cdot Q(\mathbf{P}, \lambda) = 0$ that results Eq.(6.1), and $(\partial/\partial\lambda)\cdot Q(\mathbf{P}, \lambda) = 0$ which produces Eq.(6.2) iteratively.

$$\mathbf{P} = \mathbf{X}\cdot\mathbf{Z}^{\mathrm{T}}\cdot(N\cdot\lambda\cdot\mathbf{M} + \mathbf{Z}\cdot\mathbf{Z}^{\mathrm{T}})^{-1} \qquad (6\text{-}1)$$

$$\lambda = (P\cdot N)^{-1}\cdot\mathrm{Tr}(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{E}) \qquad (6\text{-}2)$$

where $\mathbf{M} = (\mathbf{P}^{\mathrm{T}}\cdot\mathbf{P}+\lambda\cdot\mathbf{I})^{-1}$, $\mathbf{Z} = \mathbf{M}\cdot\mathbf{P}^{\mathrm{T}}\cdot\mathbf{X}$ and $\mathbf{E} = (\mathbf{I}-\mathbf{P}\cdot\mathbf{M}\cdot\mathbf{P}^{\mathrm{T}})\cdot\mathbf{X}$. EM also results two posteriors, i.e. $\boldsymbol{z}|\boldsymbol{x} \sim \mathcal{N}(\mathbf{M}\cdot\mathbf{P}^{\mathrm{T}}\cdot\boldsymbol{x}, \lambda\cdot\mathbf{M})$ and $\boldsymbol{e}|\boldsymbol{x} \sim \mathcal{N}(\{\mathbf{I}-\mathbf{P}\cdot\mathbf{M}\cdot\mathbf{P}^{\mathrm{T}}\}\cdot\boldsymbol{x}, \lambda\cdot\mathbf{P}\cdot\mathbf{M}\cdot\mathbf{P}^{\mathrm{T}})$. So, Eq.(7) is obtained.

$$\underline{z} = \mathbf{M}\cdot\mathbf{P}^{\mathrm{T}}\cdot\boldsymbol{x} \text{ and } \underline{e} = \{\mathbf{I}-\mathbf{P}\cdot\mathbf{M}\cdot\mathbf{P}^{\mathrm{T}}\}\cdot\boldsymbol{x} \qquad (7)$$

Therefore $\mathbf{Q} = \mathbf{M}\cdot\mathbf{P}^{\mathrm{T}}$ and $\mathbf{W} = (\mathbf{I} - \mathbf{P}\cdot\mathbf{M}\cdot\mathbf{P}^{\mathrm{T}})$. In case of PPCR, y is regressible by $\boldsymbol{b}^{\mathrm{T}}\cdot\boldsymbol{x}$ with $r^2(L)$ regressibility, where $r^2(L) = \boldsymbol{y}\cdot(\mathbf{M}\cdot\mathbf{P}^{\mathrm{T}}\cdot\mathbf{X})^{+}\cdot(\mathbf{M}\cdot\mathbf{P}^{\mathrm{T}}\cdot\mathbf{X})\cdot\boldsymbol{y}^{+}$, and here $\eta_{\mathbf{Z}} = 1$.

Suppose a new measurement set $\{x, y\}$ is obtained from the process. Is y regressible by $\boldsymbol{b}^{\mathrm{T}}\cdot x$? If $\underline{e} = \mathbf{W}\cdot x \sim \mathcal{N}(\boldsymbol{0}, \lambda\cdot\mathbf{I})$ then x follows the PPCA model. Therefore y is expected to be regressible by $\boldsymbol{b}^{\mathrm{T}}\cdot x$ with α level of significance. Eq.(8) is the test statistics for the regressibility of y.

$$\|\underline{e}\|_{\mathrm{M}}^2 \in [\,0, \chi^{-2}{}_{(1-\alpha,\,P)}) \text{ or } \lambda^{-0.5}\cdot\underline{e}_p \in [\mathcal{N}_s^{-1}{}_{(0.5\cdot\alpha)}, \mathcal{N}_s^{-1}{}_{(1-0.5\cdot\alpha)})\ \forall p \qquad (8)$$

where $\|\underline{e}\|_{\mathrm{M}}^2 = \lambda^{-1}\cdot x^{\mathrm{T}}\cdot\mathbf{W}^{\mathrm{T}}\cdot\mathbf{W}\cdot x$, and $\underline{e}_p = (\mathbf{W}\cdot x)_p$ denotes the p^{th} element of $\underline{e}$. Additionally, in-control criterion can also be set as Eq. (9).

$$\|\underline{z}\|_{\mathrm{M}}^2 \in [\,0, \chi^{-2}{}_{(1-\alpha,\,L)}) \text{ or } \underline{z}_l \in [\mathcal{N}_s^{-1}{}_{(0.5\cdot\alpha)}, \mathcal{N}_s^{-1}{}_{(1-0.5\cdot\alpha)}) \qquad (9)$$

where $\|\underline{z}\|_{\mathrm{M}}^2 = x^{\mathrm{T}}\cdot\mathbf{Q}^{\mathrm{T}}\cdot\mathbf{Q}\cdot x$ and $\underline{z}_l$ denotes the l^{th} element of $\underline{z} = \mathbf{Q}\cdot x$.

Results and Discussion

Various types of multivariate regression methods can be unified by the block diagram shown in the left of Figure 1 not only the orthogonal basis methods but also the probabilistic method, i.e. PPCR. If the mixing matrix **P** were set, then all of the filters **Q**, **W**, $\boldsymbol{c}_{\mathrm{LS}}$ and $\boldsymbol{b}$, and the recovered scores $\underline{z}$ and $\underline{e}$ are uniquely determined by Eq. (5) for the orthogonal methods, and Eq.(7) for the probabilistic method. Figure 2 shows an illustrative example for the PPCR with respect to the test set $\{x, y\}$. As shown in the figure, the main advantage of PPCR over the other methods is that it can suggest the regressibility for a new comer whether z is still the common factor both of x and y or not. If z is the common factor then

y can be expected to regressible, else irregressible.

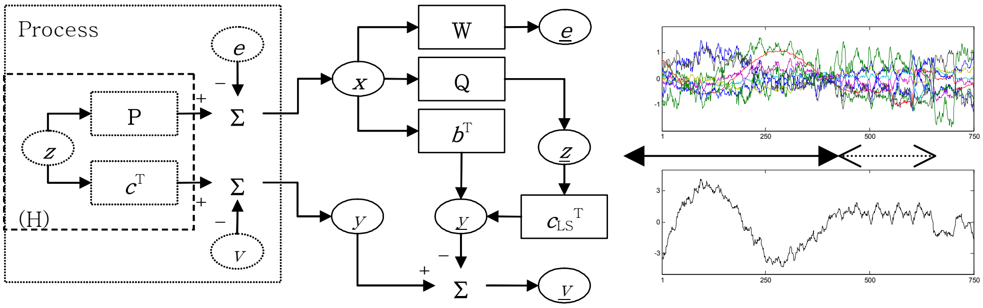

Figure 1: (Left) Block diagram for multivariate regression methods under the assumption that latent variable exists. If r^2 is sufficiently large then it implies (H) block in the figure is correct, else there is another latent sources which were not measured by $\boldsymbol{x}$. (Right) Data set for model calibration $\{\boldsymbol{x}^{(n)}, y^{(n)}\}$ for n={1,…,500}, and validation $\{x^{(k)}, y^{(k)}\}$ for k={1,…,250}.

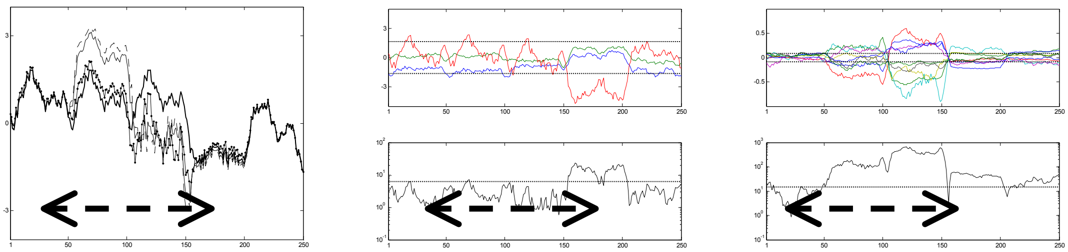

Figure 2: (Left) Regression results for the test set $\{x^{(k)}, y^{(k)}\}$ for k={1,…,250} by MLR, PCR, PLS1 and PPCR. Dotted arrow indicates the irregressible region. (Middle) Process monitoring result to check whether the process is under in-control or not, e.g. $\|\underline{z}\|_M^2$ for the top and $\underline{z}_l$ $\forall l$ for the bottom. (Right) Regressibility test plot to check whether x is still useful to estimate y or not, where $\|\underline{e}\|_M^2$ for the top and $\lambda^{-0.5}\cdot\underline{e}_p$ $\forall p$ for the bottom.

Acknowledgement: This work is supported by the BK21 program.

References

[1] J.H. Kalivas, "Basis sets for multivariate regression," *Anal. Chim. Acta*, **428**, 31(2001).

[2] M. Tipping & C. Bishop, "Probabilistic Principal Component Analysis," *J. Roy. Stat. Soc. B. Sta.*, **61**, 611(1997).

[3] [0]A.P. Dempster, N.M. Laird & D.B. Rubin, "Maximum likelihood form incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B. Sta.*, **39**, 1(1977).