Fault Variable Identification in Hotelling's T^2 procedure

Joungyoun Kim

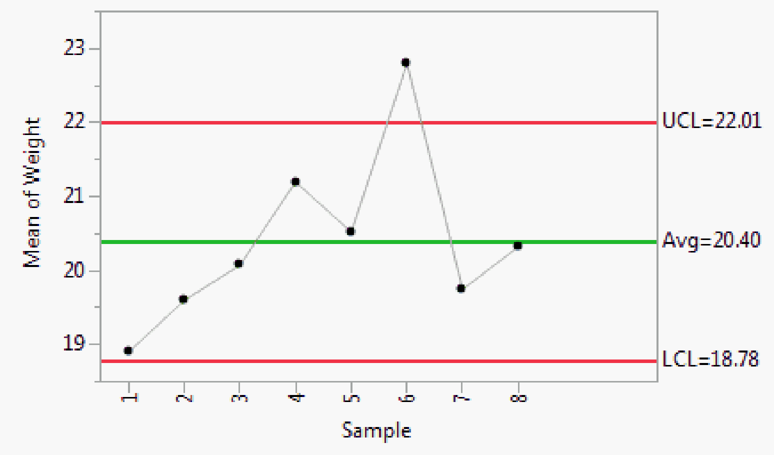
Yonsei University
College of Nursing

May 28, 2021

Outlines

Example

Introduction: Example



- Statistical process control (SPC)

Statistical process control (SPC)

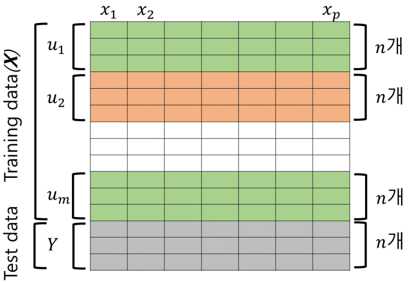
- Statistical process control (SPC)
  - A method of quality control
  - To monitor and control a process.

$$\text{Efficiency} = \begin{cases} \text{More products;} \\ \text{Less wastes.} \end{cases}$$

- Control chart: a tool of SPC

Introduction: Hotelling's T^2

- Data structure



- $\bar{\mathrm{u}}_i$, S_i: sample mean and covariance of u_i

$$\bar{\bar{\mathrm{x}}} = \frac{1}{m}\sum_{i=1}^{m} \bar{\mathrm{u}}_i \text{ and } \bar{\mathrm{S}}_{\mathrm{X}} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{S}_i,$$

- $\bar{\mathrm{y}}$: sample mean of Y
- Hotelling's T^2

$$\mathrm{T}^2 = n(\bar{\mathrm{y}} - \bar{\bar{\mathrm{x}}})^\top \bar{\mathrm{S}}_{\mathrm{X}}^{-1}(\bar{\mathrm{y}} - \bar{\bar{\mathrm{x}}}).$$

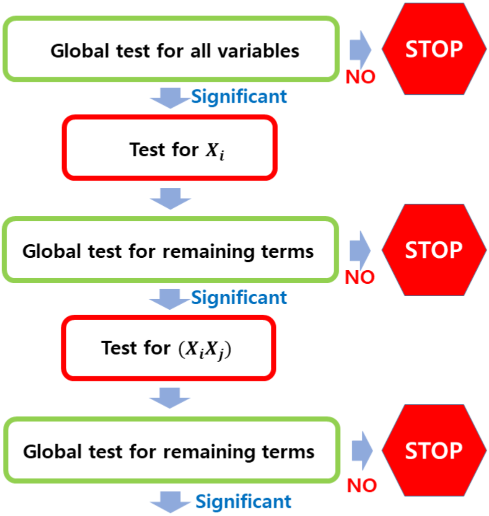
- Upper Control Limit (UCL)

$$\mathrm{UCL} = \frac{p(m+1)(n-1)}{mn - m - p + 1} F_{(\alpha, df_1, df_2)}$$

$$df_1 = p,\ df_2 = mn - m - p + 1$$

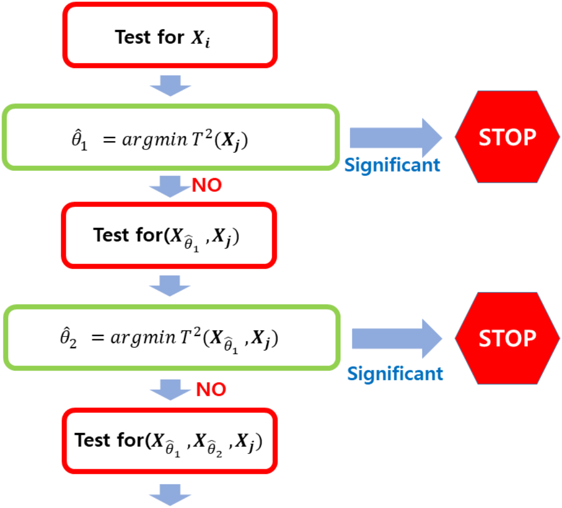
Post HT procedure: MTY

- Mason, R.L., Tracy, N.D., and Young, J.C. (1995)



Post HT procedure: Adaptive Step-down procedure (ASD)

- Kim, J., Jeong, M.K., Elsayed, E.A., Al-Khalifa, K.N., and Hamouda, A.M.S. (2016).



Model

- $\boldsymbol{\mu}_{\mathsf{X}} = \left(\mu_{\mathsf{X}1}, \mu_{\mathsf{X}2}, \ldots, \mu_{\mathsf{X}p}\right)^{\top}$
- $\boldsymbol{\mu}_{\mathsf{Y}} = \left(\mu_{\mathsf{Y}1}, \mu_{\mathsf{Y}2}, \ldots, \mu_{\mathsf{Y}p}\right)^{\top}$
- A latent variable

$$\boldsymbol{\gamma} = \left(\gamma_1, \gamma_2, \ldots, \gamma_p\right)^{\top}$$

$$\gamma_i = \begin{cases} 0 & \text{if } \mu_{\mathsf{Y}i} = \mu_{\mathsf{X}i}; \\ 1 & \text{if } \mu_{\mathsf{Y}i} \neq \mu_{\mathsf{X}i}. \end{cases}$$

Model

- $\boldsymbol{\mu}_\mathsf{X}(\boldsymbol{\gamma})$, $\boldsymbol{\mu}_\mathsf{Y}(\boldsymbol{\gamma})$, $\bar{\mathsf{y}}(\boldsymbol{\gamma})$, $\bar{\bar{\mathsf{x}}}(\boldsymbol{\gamma})$ and $\bar{\mathsf{S}}_{\mathsf{X}(\boldsymbol{\gamma})}$: the sub-vectors (matrix) of $\boldsymbol{\mu}_\mathsf{X}$, $\boldsymbol{\mu}_\mathsf{Y}$, $\bar{\mathsf{y}}$, $\bar{\bar{\mathsf{x}}}$ and $\bar{\mathsf{S}}$ corresponding to the non-zero elements of $\boldsymbol{\gamma}$
- Hotelling's T^2

$$\mathsf{T}^2(\boldsymbol{\gamma}) = n\big(\bar{\mathsf{y}}(\boldsymbol{\gamma}) - \bar{\bar{\mathsf{x}}}(\boldsymbol{\gamma})\big)^\top \bar{\mathsf{S}}^{-1}_{\mathsf{X}(\boldsymbol{\gamma})}\big(\bar{\mathsf{y}}(\boldsymbol{\gamma}) - \bar{\bar{\mathsf{x}}}(\boldsymbol{\gamma})\big), \tag{1}$$

- $C(\boldsymbol{\gamma})$: p-value of $\mathsf{T}^2(\boldsymbol{\gamma})$.
- Boltzman type distribution

$$P(\boldsymbol{\gamma}) = \frac{1}{\Psi(\beta)} \exp\big\{-\beta \cdot C(\boldsymbol{\gamma})\big\}, \beta > 0. \tag{2}$$

- Goal: to find $\boldsymbol{\gamma}$ with the maximum $P(\boldsymbol{\gamma})$

Method: Shotgun Stochastic Search

- Neighborhood $N(\gamma)$ when γ=(1, 1, 1, 0, 0), $p = 5$

	$N(\gamma)$	γ^*	$T(\gamma^*)$	df_1	df_2	$C(\gamma^*)$
Add	γ^+	1 1 1 1 0				
		1 1 1 0 1				
Delete	γ^-	0 1 1 0 0				
		1 0 1 0 0				
		1 1 0 0 0				
Swap	γ^0	0 1 1 1 0				
		0 1 1 0 1				
		1 0 1 1 0				
		1 0 1 0 1				
		1 1 0 1 0				
		1 1 0 0 1				

Method: Shotgun Stochastic Search

- Propose γ^* with probability

$$q(\gamma^* \mid \gamma) = \frac{P(\gamma^*)\mathrm{I}(\gamma^* \in N(\gamma))}{\sum_{\mathsf{s}\in N(\gamma)} P(\mathsf{s})},$$

- Accept γ^* with probability

$$\begin{aligned}\alpha &= \min\left\{1, \sum_{\mathsf{s}\in N(\gamma)} P(\mathsf{s}) \Big/ \sum_{\mathsf{s}\in N(\gamma^*)} P(\mathsf{s})\right\} \\ &= \min\left\{1, \sum_{\mathsf{s}\in N(\gamma)} \exp(-\beta C(\mathsf{s})) \Big/ \sum_{\mathsf{s}\in N(\gamma^*)} \exp(-\beta C(\mathsf{s}))\right\}\end{aligned}$$

Numerical study: Setting

- Setting
  - Control mean vector $\mathcal{H}_0 : \mu_\mathsf{Y} = \mu_\mathsf{X}$
    - ★ p=25
    - ★ $\mathcal{H}_{\rm 5th} : \boldsymbol{\mu}_\mathsf{Y} = \boldsymbol{\mu}_\mathsf{X} + a \times \sqrt{p/5} \sum_{j=1}^{5} (-1)^{j-1} \times \mathrm{e}_j$
    - ★ $\mathcal{H}_{\rm 10th} : \boldsymbol{\mu}_\mathsf{Y} = \boldsymbol{\mu}_\mathsf{X} + a \times \sqrt{p/10} \sum_{j=1}^{10} (-1)^{j-1} \times \mathrm{e}_j$
  - Control distribution: generate X from Multivariate Normal or t(5)
  - Control covariance matrix
    - ★ IND: $\Sigma_1 = \mathrm{diag}\big(\lambda_1, \lambda_2, \lambda_3, 1_{p-3}\big)$, where $\lambda_1 = 4$, $\lambda_2 = 3$, $\lambda_3 = 2$, and 1_{p-3} is the $(p-3)$-dimensional row vector of all ones.
    - ★ AR: $\Sigma_2 = \Sigma_1 + \Big(\mathrm{A}(\rho) - \mathrm{I}_p\Big)$, where $\mathrm{A}(\rho) = \big(a_{ij}\big)_{1 \le i,j \le p}$ with $a_{ij} = \rho^{|i-j|}$ and ρ is set as 0.5.
    - ★ PC: $\Sigma_3 = \mathrm{LL}^T + \mathrm{I}_p$, where $\mathrm{L}(p \times q, q < p)$ and $L_{ij} \sim N(0,1)$.

Numerical study: Setting

- Existing methods
  - MTY: Mason, Tracy and Young (1997)
  - ASD: Kim *et al.*(2016)
  - LASSO: Zou *et al.* (2009), Zou and Qiu (2009)

Numerical study: Result (IND case)

		Mean-sen.		Mean-spec.	
		$\mathcal{H}_{5\text{th}}$	$\mathcal{H}_{10\text{th}}$	$\mathcal{H}_{5\text{th}}$	$\mathcal{H}_{10\text{th}}$
N	S1	4.840	8.940	13.100	11.160
		(0.370)	(1.331)	(1.669)	(1.405)
	S3	4.833	8.293	12.633	9.893
		(0.263)	(0.616)	(1.031)	(1.000)
	MTY	5.000	9.260	19.000	13.880
		(0.000)	(0.899)	(0.881)	(2.135)
	ASD:T	4.880	7.660	19.900	14.900
		(0.385)	(1.533)	(0.303)	(0.303)
	ASD:S	4.860	7.900	19.660	14.740
		(0.351)	(1.329)	(0.557)	(0.487)
	LASSO	3.040	1.700	18.220	14.280
		(2.157)	(2.957)	(3.388)	(2.603)
t(5)	S1	4.660	7.920	12.200	10.020
		(0.557)	(1.368)	(1.863)	(1.868)
	S3	4.680	7.587	11.900	9.467
		(0.375)	(0.882)	(1.334)	(1.302)
	MTY	4.780	7.780	18.340	12.420
		(0.507)	(1.718)	(2.925)	(4.607)
	ASD:T	4.760	7.320	19.020	14.040
		(0.555)	(1.720)	(1.097)	(1.106)
	ASD:S	4.420	5.740	19.400	14.480
		(0.731)	(1.482)	(0.904)	(0.707)
	LASSO	3.220	2.160	17.300	14.560
		(2.053)	(2.780)	(3.460)	(1.387)

Numerical study: Result (AR case)

		Mean-sen.		Mean-spec.	
		$\mathcal{H}_{5\text{th}}$	$\mathcal{H}_{10\text{th}}$	$\mathcal{H}_{5\text{th}}$	$\mathcal{H}_{10\text{th}}$
N	S1	5.000 (0.000)	9.820 (0.482)	13.000 (1.143)	11.540 (1.265)
	S3	4.993 (0.047)	8.787 (0.355)	12.393 (0.779)	9.280 (1.040)
	MTY	4.980 (0.141)	9.340 (0.717)	18.780 (1.112)	13.660 (2.228)
	ASD:T	4.880 (0.328)	7.980 (1.286)	19.900 (0.364)	14.860 (0.405)
	ASD:S	4.920 (0.274)	8.020 (1.237)	19.760 (0.517)	14.780 (0.507)
	LASSO	4.520 (0.707)	8.580 (1.500)	18.300 (4.287)	14.280 (2.322)
t(5)	S1	4.960 (0.198)	9.520 (0.762)	12.680 (1.406)	10.460 (2.082)
	S3	4.980 (0.080)	8.667 (0.522)	12.047 (1.052)	9.067 (0.901)
	MTY	4.820 (0.388)	8.300 (1.359)	17.660 (4.680)	12.460 (4.546)
	ASD:T	4.780 (0.465)	7.620 (1.276)	19.060 (1.434)	14.100 (1.313)
	ASD:S	4.540 (0.579)	6.240 (1.001)	19.540 (0.762)	14.360 (0.942)
	LASSO	4.680 (0.513)	8.180 (2.116)	17.340 (4.525)	13.220 (3.164)

Numerical study: Result (PC case)

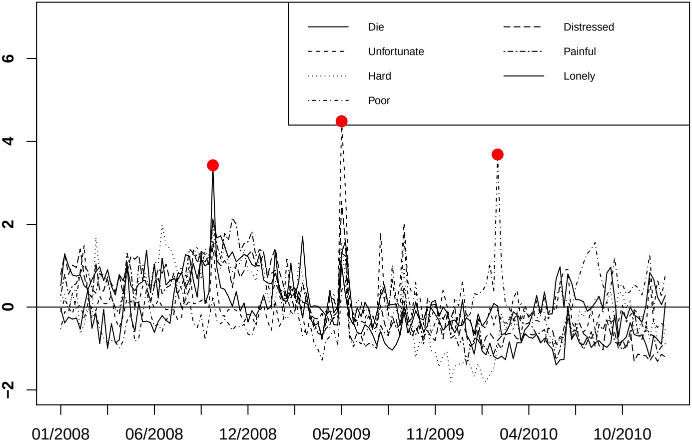
		Mean-sen.		Mean-spec.	
		$\mathcal{H}_{5\text{th}}$	$\mathcal{H}_{10\text{th}}$	$\mathcal{H}_{5\text{th}}$	$\mathcal{H}_{10\text{th}}$
N	S1	3.300	6.280	10.700	7.680
		(0.953)	(1.796)	(2.468)	(2.316)
	S3	3.260	5.960	10.447	7.467
		(0.766)	(1.217)	(1.692)	(1.534)
	MTY	2.340	5.600	11.780	7.320
		(2.115)	(4.076)	(8.918)	(6.310)
	ASD:T	0.760	1.160	19.320	14.240
		(0.771)	(0.889)	(0.768)	(0.716)
	ASD:S	0.860	1.240	19.320	14.000
		(0.857)	(0.822)	(0.891)	(0.969)
	LASSO	2.380	3.540	14.560	11.440
		(1.689)	(2.270)	(5.257)	(3.453)
t(5)	S1	3.120	5.660	10.900	7.420
		(1.206)	(1.479)	(2.243)	(2.071)
	S3	3.100	5.587	10.533	7.547
		(0.879)	(1.085)	(1.911)	(1.505)
	MTY	3.100	6.080	9.180	5.980
		(1.951)	(4.208)	(8.817)	(6.473)
	ASD:T	1.020	1.900	18.120	13.340
		(0.869)	(1.632)	(1.649)	(2.219)
	ASD:S	0.700	1.120	19.080	14.020
		(0.863)	(1.1)	(0.986)	(1.059)
	LASSO	2.080	3.560	15.060	10.820
		(1.805)	(2.815)	(5.247)	(4.183)

Blog data

- Moon and Lee (2013)
- DAUM blog data from Jan.1, 2008–Dec.31, 2010 (156 weeks)
- Daily number of blogs per 100K blogs that contains
  - Die: 죽고싶다
  - Unfortunate: 안타깝다
  - Hard: 힘들다
  - Poor (or Pitiful): 불쌍하다
  - Distressed: 괴롭다
  - Painful: 아프다
  - Lonely: 외롭다
- $p = 7$
- Use the latest 12 weeks as training data: m=12, n=7

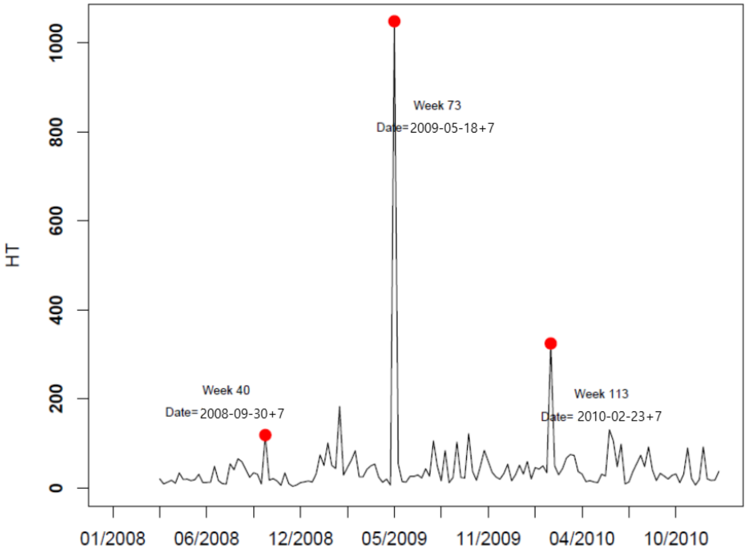
Blog data

- Weekly mean number of blogs per 100K blogs that contains the seven words



Blog data

- Trace plot of Hotelling's T^2 over weeks



Blog data

Week 40	Die	Unfort.	Hard	Poor	Distr.	Pain.	Lonely	#	$\log(C(\gamma))$
S3	1	1	0	1	0	0	1	4	-28.48
	1	1	0	0	0	0	1	3	-28.03
	1	1	1	1	0	0	1	5	-26.85
MTY	1	1	1	0	1	1	1	6	-23.32
ASD	1	1	0	0	0	1	1	4	-26.33
LASSO	1	0	0	1	0	0	0	2	-21.86
univariate t	-7.89	-5.48	-2.60	-1.41	-3.09	-4.01	-5.98		
Week 73	Die	Unfort.	Hard	Poor	Distr.	Pain.	Lonely	#	$\log(C(\gamma))$
S3	0	1	0	0	0	0	0	1	-90.99
	0	1	0	0	1	1	0	3	-90.22
	1	1	0	0	1	0	0	3	-90.10
MTY	1	1	1	1	1	1	1	7	-79.91
ASD	1	1	1	1	1	1	0	6	-82.43
LASSO	0	1	0	0	0	0	0	1	-90.99
univariate t	-3.54	-28.94	-7.84	-3.89	-14.65	-7.18	-5.79		
Week 113	Die	Unfort.	Hard	Poor	Distr.	Pain.	Lonely	#	$\log(C(\gamma))$
S3	0	0	0	1	0	0	1	2	-54.01
	0	0	0	1	1	0	1	3	-53.84
	0	1	0	1	0	0	1	3	-52.28
MTY	0	1	1	1	0	0	1	4	-50.24
ASD	0	0	1	1	0	0	1	3	-51.63
LASSO	0	0	1	1	1	1	1	5	-49.64
univariate t	-3.54	-28.94	-7.84	-3.89	-14.65	-7.18	-5.79		

Conclusion

- Our proposed method can be applied to any global testing statistic whose p-value or selection criterion is analytically available.
- We need to find a numerical study setting which can explain the blog data result.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov Chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, **90(431),** 909-920.

Mason, R.L., Tracy, N.D., and Young, J.C. (1995). Decomposition of T^2 for multivariate control chart interpretation. *Journal of Quality Technology*, **27(2)**, 99-108.

Mason, R.L., Tracy, N.D., and Young, J.C. (1995). A practical approach for interpreting multivariate T^2 control chart signals. *Journal of Quality Technology*, **29(4)**, 396-406.

Hans, C., Dobra, A., West, M. (2007). Shotgun Stochastic Search for "Large p" Regression. *Journal of the American Statistical Association*, **102**, 507–516.

Montgomery, D.C. (2009) *Introduction to Statistical Quality Control (6th edition)*, John Wiley & Sons, New York.

Zou, C. and Qiu, P. (2009). Multivariate statistical process control using LASSO. *Journal of the American Statistical Association*, **104**, 1586-1596.

Zou, C., Jiang, W., and Tsung, F. (2012). A lasso-based diagnostic framework for multivariate statistical process control. *Technometrics*, **53(3)**, 297-309.

Moon, J. and Lee, S. (2013). Detection of the Change in Blogger Sentiment using Multivariate Control Charts. *The Korean Journal of Applied Statistics*, **26(6)**, 903–913. (in Korean)

Kim, J., Jeong, M.K., Elsayed, E.A., Al-Khalifa, K.N., and Hamouda, A.M.S. (2016). An adaptive step-down procedure for fault variable identification. *International Journal of Production Research*, **54(11)**, 3187-3200.

Lee, S. and Lim, J. (2017). Phase 2 monitoring of changes in mean from high dimensional data. *Applied Stochastic Models in Business and Industry*, **33**, 626-639.