Journal of Data Science **1**(2003), 27-41

A Statistical Method of Detecting Bioremediation

Dechang Chen[1], Michael Fries[2], and John M. Lyon[1]
[1] *University of Wisconsin-Green Bay,* [2] *DePaul University*

Abstract: Hydrocarbon contaminated soils result from pipeline ruptures, petroleum manufacture spills, as well as storage and transportation accidents (Bossert and Bartha (1984)). The cost of removal of the contaminated solids followed by incineration or by disposal in a landfill is prohibitive. Bioremediation - the use of microorganism populations to eliminate hydrocarbon contaminations from the environment - is the most acceptable technology for hydrocarbon cleanup (Bossert and Bartha (1984)). It can be argued that a decrease of the oil concentration in soil is not due to biodegradation but due to sorption. If this were the case, since mass transfer of sorption is a gradual process, a slow decrease in the oil recovery rate may be observed after a spill. However, a rapid or sudden decrease in the oil concentration during the incubation should exclude sorption as the primary mechanism contributing to the observed hydrocarbon loss. A Bayesian procedure is given to detect a change of the linear relationship between the oil concentration (the dependent variable) and the time in days since the addition of the oil (the independent variable). The advantage of this procedure is that it does not need to assume that the variance of the error before the change is equal to that after the change. The implementation of this procedure is straightforward.

Key words: Bioremediation, change point, linear model, posterior distribution.

1. Introduction

In Ma (1999), an investigation of the potential enhanced rate, the potential enhanced extent, or both of bioremediation of 20W motor oil contaminated soil by inoculation of microorganisms was undertaken. The inoculant of the soil was with a commercially available product containing microbial isolates with bioremediation abilities. The product package also included a nutrient product for feeding the microbes, which was added to all treatments. Also examined was the effect of different soil types (sandy loam with low organic matter, denoted as sandy loam, sandy loam with high organic matter, denoted silt loam, and loam with high clay content, denoted loam) on bioremediation.

Ma noted that it could be argued that the observed decrease of the oil concentration in all soil types was not due to biodegradation but due to its sorption effects. Ma also noted that the mass transfer of sorption is a gradual process, and should result in an unchanged degradation rate. Rapid decrease of the oil concentration during the incubation period would exclude sorption as the primary mechanism contributing to the observed hydrocarbon loss, and therefore providing evidence of effective bioremediation.

When oil concentration is modeled via linear regression lines, such a rapid change would be indicated by a change in the regression line at some time during the study period. A statistical test that such a change exists would provide an objective method of eliminating sorption as the primary mechanism contributing to the observed hydrocarbon loss, and supporting the existence of bioremediation. In addition, estimating the time when such a change takes place is then meaningful and potentially useful. For example, the knowledge of the change point under a given set of conditions could determine the choice of inoculant or affect the strategy of dealing with a particular hydrocarbon spill.

Statistical methods have been developed to detect change points in regression. For example, one may use likelihood procedures (see, for example, Worsley (1983) and Srivastava and Worsley (1986)) and Bayesian methods (see, for example, Broemeling and Chin Choy (1981), Moen and Broemeling (1984), and Guttman and Srivastava (1987)). This paper focuses on Bayesian methods. In a Bayesian analysis one needs to give prior probability distributions to both change points and the parameters. The resulting posterior probabilities, based on the data, are then used to make neces-

sary inferences. In particular, the posterior distribution of the change point can be employed to locate the "actual time" at which a sequence of observations undergoes sudden changes, that is, the time where the posterior probability assumes its maximum will be used to estimate the actual time of the change. Such a methodology has found many interesting applications in practice. For example, Smith (1975) developed Bayesian tests for structural stability, a topic of interest to economists. Moen and Broemeling (1984) developed a procedure for testing whether or not a change has occurred in the regression matrix of a multivariate linear model. The resulting test is based on the marginal posterior distribution of the change point. A numerical example using a bivariate regression model was used to illustrate the test procedure. Guttman and Srivastava (1987) provided a Bayesian method of finding the change point for the general multivariate linear model in which it is suspected that a change occurs from one linear model to another where the different models have some common parameters. They also discussed a certain change-point problem that involves a switch at some time from one growth-curve model to another. They illustrated the general results through the example of locating the time at which the effects of labor inputs on gross domestic product may undergo.

In this article, we developed a Bayesian method for detecting the time where the regression lines differ. One advantage of the procedure is that it does not assume that the variance of the error before the change of regression is equal to that after the change. This is practical since in many biological situations, the variance of concentrations changes with the magnitude of the concentrations.

The rest of the paper is organized as follows. Section 2 describes the experiment and dataset. Section 3 introduces the statistical models. Section 4 presents the analysis of bioremediation. Section 5 concludes our study. The derivation of formulas is given in an Appendix.

2. Experiment

Three types of uncontaminated soils (sandy loam, silt loam and loam) were collected from different locations of the University of Wisconsin-Green Bay. This soil was air-dried and screened by using U.S. Standard Sieve Series

No. 10 (2.00 mm). 20W motor oil was added to simulate a hydrocarbon spill with a concentration of 3000 mg oil per kilogram of dry soil. Each type of soil was then divided into three plots. These plots received one of three treatments. For treatment one, the control, no additional material was added. For treatment two, only the nutrients were added. This treatment was to provide information on whether the nutrients could stimulate the degradation of the oil by indigenous microorganisms. And for treatment three, both the nutrients and the inoculum were added. This treatment was to provide information on whether the inoculum could enhance the rate of oil degradation. These treated plots of the soil were then divided in half to provide a duplication of each treatment on each type of soil. Soil moisture was adjusted twice a week to maintain between a 40 and 50 percent of the soil content by weight at saturation. The soil was also mixed thoroughly twice per week to homogenize the soil and oil contaminant distribution and to enhance aeration. The results of both duplications of each of the three treatments on each of the three soil types are given in Table 1.

3. Models

The simplest framework of detecting change point in regression may be stated as follows. Consider a sequence of n pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$, where y_i is the value of the response variable (dependent variable) in the ith trial and x_i is the known value of the independent variable in the ith trial. Suppose the following model:

$$\begin{aligned} y_i &= \beta_{11} + \beta_{12} x_i + \epsilon_{1i}, \quad i = 1, \cdots, t, \\ y_i &= \beta_{21} + \beta_{22} x_i + \epsilon_{2i}, \quad i = t+1, \cdots, n, \end{aligned} \tag{3.1}$$

where $3 \leq t \leq n-3$, ϵ_{1i} are independent $N(0, \sigma_1^2)$, ϵ_{2i} are independent $N(0, \sigma_2^2)$, and ϵ_{1i} and ϵ_{2i} are independent. The parameters β_{ij}, σ_i^2, and t are all unknown. The above model indicates a switch in regression of y on x at "time" t. The main task is to estimate t. The range $3 \leq t \leq n-3$ is necessary in order to allow for the estimation of the two regression lines as well as the estimates of the associated error terms.

For completeness of the theory, we shall consider the following general

Table 1: 20W oil concentration (mg/kg d.w.) in loam. Treatment 1 = Control, Treatment 2 = Nutrient, Treatment 3 = Nutrient and Inoculum.

Days	3	7	11	14	18	21	25	32	39	46
Sandy Loam										
Treatment 1										
Dup 1	2660	3050	2920	2590	2450	2910	2510	2480	2570	2410
Dup 2	3160	3110	3040	2690	2940	3210	2820	2850	2520	2690
Treatment 2										
Dup 1	3900	3200	3060	2850	2830	3000	2610	2740	2550	2690
Dup 2	2740	2820	2920	2930	2880	2810	2820	2710	2630	2410
Treatment 3										
Dup 1	3280	2960	2990	2700	2730	2850	2540	2870	2880	2270
Dup 2	2870	3090	2950	2990	2710	3190	2800	2700	2690	3040
Silt Loam										
Treatment 1										
Dup 1	3490	2530	3670	2880	2770	1910	2850	1740	1950	1500
Dup 2	2310	3080	3100	2810	2590	2660	2520	1900	1840	1950
Treatment 2										
Dup 1	2390	3360	3500	3540	2420	2500	1390	2680	2400	1890
Dup 2	2100	3210	3250	3310	2620	2520	3160	3130	2600	2770
Treatment 3										
Dup 1	2550	3180	3030	3120	2690	2450	2840	2170	2110	2100
Dup 2	2110	3110	3190	3230	2930	3030	2740	2420	2600	2480
Loam										
Treatment 1										
Dup 1	3100	1990	2570	1640	1280	1560	1240	1650	1690	1150
Dup 2	2400	3710	2940	1540	1270	1480	970	2010	1780	1340
Treatment 2										
Dup 1	2520	2620	2570	1370	1320	1620	1300	1680	1940	1380
Dup 2	2330	2380	2290	1220	1450	1680	1040	1480	1700	1430
Treatment 3										
Dup 1	2160	2180	2410	1080	1250	1410	990	1360	1300	870
Dup 2	2470	2320	3110	1600	1730	1460	1180	1490	1620	1190

Source: Ma (1999)

case:

$$\begin{aligned} y_i &= \boldsymbol{\beta}_1' \mathbf{x}_i + \epsilon_{1i}, \quad i = 1, \cdots, t, \\ y_i &= \boldsymbol{\beta}_2' \mathbf{x}_i + \epsilon_{2i}, \quad i = t+1, \cdots, n, \end{aligned} \tag{3.2}$$

where $p+1 \leq t \leq n-p-1$, $\mathbf{x}_i$ is a known p vector of independent variables, and both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are p vectors of unknown parameters. This model shows that the linear relationship between the dependent and independent variables changes at "time" t. The range $p+1 \leq t \leq n-p-1$ is necessary in order to allow for the estimation of the two regression lines as well as the estimates of the associated error terms.

In the following, we shall impose some prior information on $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1^2, σ_2^2, and t, and derive the posterior probability mass function of the change point t. This posterior distribution will be used to estimate t in (3.2).

Set the matrices X_1 and X_2 and vectors $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}$ as follows.

$$X_1' = (\mathbf{x}_1, \cdots, \mathbf{x}_t), \qquad X_2' = (\mathbf{x}_{t+1}, \cdots, \mathbf{x}_n),$$

$$\mathbf{y}_1' = (y_1, \cdots, y_t), \quad \mathbf{y}_2' = (y_{t+1}, \cdots, y_n), \quad \text{and} \quad \mathbf{y}' = (\mathbf{y}_1', \mathbf{y}_2').$$

Assume that X_i is of full rank, i.e., the rank of X_i is p, the number of columns of X_i, for $i = 1, 2$. Then the square matrix $X_i'X_i$ is of full rank so that the inverse of $X_i'X_i$ is defined in the usual way. Let $S_i = (\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}}_i)'(\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}}_i)$ with $\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\mathbf{y}_i$, $i = 1,\ 2$.

Given that no other information is available, what prior might we put on the parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1^2, σ_2^2, and t ? Intuitively, "noninformative" priors should be given in the absence of "enough" information. Let us consider first the time t. Clearly, we should assume that each of the time values $p+1$, $\cdots$, $n-p-1$ is equally likely. This fact is indicated by the uniform prior $p(t) \propto$ constant. For the vector $\boldsymbol{\beta}_1$, we may also assume the uniform prior, that is, uniform over the entire p-dimensional space. This is expressed by $p(\boldsymbol{\beta}_1) \propto$ constant. Similarly, we have $p(\boldsymbol{\beta}_2) \propto$ constant. To find a prior for σ_1^2, we consider $\ln \sigma_1^2$. Since $\ln \sigma_1^2$ can take any value between $-\infty$ and ∞, we may assume the uniform prior on $\ln \sigma_1^2$, that is, uniform over the entire one-dimensional space. Thus we have $p(\ln \sigma_1^2) \propto$ constant, so that $p(\sigma_1^2) \propto 1/\sigma_1^2$. Same reasoning yields $p(\sigma_2^2) \propto 1/\sigma_2^2$. Note that the noninformative priors for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\ln \sigma_1^2$, and $\ln \sigma_2^2$ should be understood in terms of Jeffreys' improper probability distribution functions. Now we assume that all the

above priors are independent. Then we have the following (uniform) prior in $\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2,\ \ln\sigma_1^2,\ \ln\sigma_2^2,$ and t:

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t) \propto \frac{1}{\sigma_1^2\sigma_2^2} \text{ for } p+1 \le t \le n-p-1. \tag{3.3}$$

It can be shown that given the prior in (3.3) and the data $\mathbf{y}$, the posterior probability that a change point occurs at time t is (see the Appendix)

$$p(t|\mathbf{y}) = K 2^{\frac{n}{2}-p}(|X_1'X_1||X_2'X_2|)^{-\frac{1}{2}}\Gamma\Big(\frac{t-p}{2}\Big)\Gamma\Big(\frac{n-t-p}{2}\Big)S_1^{-\frac{t-p}{2}}S_2^{-\frac{n-t-p}{2}}, \tag{3.4}$$

where the constant K is such that

$$K^{-1} = 2^{\frac{n}{2}-p}\sum_{t=p+1}^{n-p-1}(|X_1'X_1||X_2'X_2|)^{-\frac{1}{2}}\Gamma\Big(\frac{t-p}{2}\Big)\Gamma\Big(\frac{n-t-p}{2}\Big)S_1^{-\frac{t-p}{2}}S_2^{-\frac{n-t-p}{2}}. \tag{3.5}$$

Now one may estimate t by the mode of the distribution in (3.4), that is by that value of t at which $p(t|\mathbf{y})$ has its maximum.

In this paper we use only noninformative priors (i.e., where no prior information is explicitly imposed). Noninformative priors are used when information about parameters is completely unknown or when proper priors such as conjugate priors do not apply. When more information on the parameters is known so that a reliable proper prior can be employed, the resulting estimate of the change point can be expected to be more accurate. For a vigorous discussion on the choice of priors, see Box and Tiao (1992).

4. Analysis

For simplicity, in this paper we will consider in detail the case of loam with high clay content receiving treatment 3, where both nutrients and inoculum were added. For this purpose we isolate the results of both duplications of treatment 3 on the loam from Table 1 to form Table 2.

Upon examination of the data in Table 2, it can be noted that the oil concentration measurements in Dup 2 are consistently larger than the oil concentration measurements in Dup 1. We suspect that there is some block

Table 2: 20W oil concentration in loam with treatment 3

Days	3	7	11	14	18	21	25	32	39	46
Dup 1	2160	2180	2410	1080	1250	1410	990	1360	1300	870
Dup 2	2470	2320	3110	1600	1730	1460	1180	1490	1620	1190

effect associated with Dup 2. Specifically, the set of conditions such as raw material source, raw material purity, and room temperature that resulted in the oil concentrations observed in Dup 1 are not exactly the same as the set of conditions that resulted in the oil concentrations observed in Dup 2. Such differences may result in the fact that all responses in Dup 2 will be τ units lower (or higher) than in Dup 1, that is, $\tilde{y} = y + \tau$, where $\tilde{y}$ is a Dup 2 observation and y is an observation produced under the conditions for Dup 1. To get valid data for our change point analysis from Dup 2, we need to account for this effect τ. A straightforward way of doing this is as follows. The average from Dup 1 is $A_1 = (2160+2180+\cdots+870)/10 = 1501$, and the average from Dup 2 is $A_2 = (2470 + 2320 + \cdots + 1190)/10 = 1817$. The difference $A_2 - A_1 = 316$, denoted $\hat{\tau}$, will be used to estimate τ. Now subtract $\hat{\tau}$ from each observation in Dup 2. See Table 3 for the modified dataset. This new dataset will be used to replace the old one (Table 2) for our analysis. In the above we discussed a simple way to remove the block effect from Dup 2. For more information on block effects, see Montgomery (2001).

Table 3: 20W oil concentration in loam with treatment 3 (Modified)

Days	3	7	11	14	18	21	25	32	39	46
Dup 1	2160	2180	2410	1080	1250	1410	990	1360	1300	870
Dup 2	2154	2004	2794	1284	1414	1144	864	1174	1304	874

The data in Table 3 suggest that the relation between the oil concentration and the time changes after 11 days, since the observations during the first 11 days look larger than those for the following days. We now use the procedure described in Section 3 to estimate the time when such a change took place.

Let y denote the (modified) oil concentration, and let $\mathbf{x} = (1, x_2)$, where x_2 denotes the time in days. In order to apply the model (3.2) to the two duplications it is necessary to enumerate the observations. We do so alternating between duplicate 1 and duplicate 2, giving the sequence: $(2160, 2154, 2180, 2004, \cdots, 870, 874)$. The posterior probabilities that a particular observation is where a change point occurs are derived by (3.4) and (3.5). They are given in Table 4.

Table 4: Posterior probabilities that a change point occurs

Days	3	7	11	14	18	21	25	32	39	46
Dup 1	—	0.001	0.000	0.005	0.000	0.000	0.000	0.000	0.002	—
Dup 2	—	0.000	0.732	0.001	0.000	0.000	0.001	0.257	—	—

The maximum probability $p(t|\mathbf{y}) = 0.732$ is seen to be when $t = 6$, indicating a change in response after 11 days. A graph of the experimental points and the two resulting regression lines are given in Figure 1.

Note that in the above we used the sequence $(2160, 2154, 2180, 2004, \cdots, 870, 874)$ as our data $\mathbf{y}$, that is, we arranged the data as if we always observed Dup 1 first in a given day. The fact is that the order of displaying Dup 1 and Dup 2 observations in a given day will not affect our final estimate of the change point. More specifically, if we display the data in the order of time (days) but without caring about the order of arranging Dup 1 and Dup 2 observations in a given day, then we have $2^{10} = 1024$ ways to write down our $\mathbf{y}$, and each $\mathbf{y}$ will lead to the same conclusion that a change in response occurs after 11 days. Clearly, different $\mathbf{y}$'s may yield different posterior distributions $p(t|\mathbf{y})$. For example, let us use the sequence $(2154, 2160, 2004, 2180, \cdots, 874, 870)$ as our $\mathbf{y}$. Then using (3.4) and (3.5), we obtain Table 5 listing the posterior probabilities that a change point occurs. Note that Table 5 is not identical with Table 4. However, since the maximum posterior probability 0.731 occurs when $t = 6$, again we see that a change in response takes place after 11 days.

In the analysis of this data in the original bioremediation study Ma (1999)

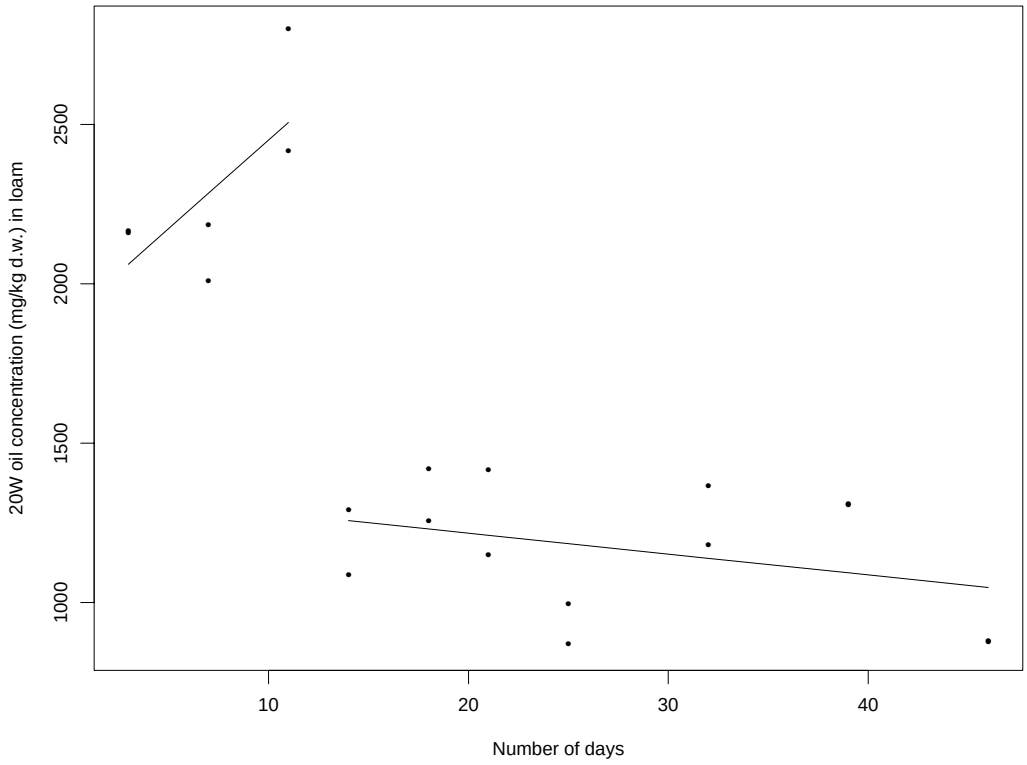

Figure 1: Oil concentration in loam as a function of days of bioremediation. The scatter represents the experimental data. The lines represent the two fitted linear models.

Table 5: Posterior probabilities based on data
$(2154, 2160, 2004, 2180, \cdots, 874, 870)$

Days	3	7	11	14	18	21	25	32	39	46
Dup 1	—	0.000	0.731	0.001	0.000	0.000	0.001	0.256	—	—
Dup 2	—	0.001	0.001	0.006	0.000	0.000	0.000	0.000	0.002	—

noted change points only in the loam with high clay content under all treatments by visual inspection. By applying our methodology to treatment 1 and treatment 2 in loam, we found the change point occurs during day 11. Therefore, in all three treatments, the change takes place during day 11. This further corroborates the conclusions in Ma (1999) that "... the rapid decrease after day 11 of the incubation with loam with high clay content should exclude sorption as the primary mechanism contributing to the observed hydrocarbon loss. The microbial activity can change exponentially which may indicate that the rapid decrease of hydrocarbon in loam soil was contributed by biodegradation."

5. Conclusion

We have derived a practical Bayesian method for the estimation of a change point in a linear regression that does not rely on the hypothesis that the same variance exists both before and after the change point. Through our analysis of bioremediation data, we have illustrated the methods use and have shown it to be an effective means in estimating the time of a change point in this situation. As the literature contains many examples where the application of this method of finding the time of a change point would be useful for other fields, including economics and biology, we feel that this method can become a useful tool in a variety of undertakings.

References

Bossert, I., and Bartha, R. (1984). The Fate of Petroleum in Soil Ecosystems. In *Petroleum Microbiology*, R. M. Atlas (ed.), Macmillan, New York, 453–473.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis.* John Wiley.

Broemeling, L. D., and Chin Choy, J. H. (1981). Detecting Structural Change in Linear Models. *Communications in Statistics A, Theory and Methods*, **10**, 2551–2561.

Guttman, I. (1982). *Linear Models: An Introduction.* New York: Wiley.

Guttman, I., and Srivastava, M. S. (1987). Bayesian Method of Detecting Change Point in Regression and Growth Curve Models. *I. B. MacNeill and*

G. J. Umphrey (eds.), Foundations of Statistical Inference, 73–91. Boston: Reidel.

Ma, X. (1999). Evaluation of a Commercial Microorganism Inoculum for the Bioremediation of 20W Motor Oil Contaminated Soil. Master Thesis, University of Wisconsin-Green Bay.

Moen, D. H., and Broemeling, L. D. (1984). Testing for a Change in the Regression Matrix of a Multivariate Linear Model. *Communications in Statistics A, Theory and Methods,* **13**, 1521–1532.

Montgomery, D. C. (2001). *Design and Analysis of Experiments,* Fifth Edition. New York: Wiley.

Smith, A. F. M. (1975). A Bayesian Approach about a Change-point in a Sequence of Random Variables. *Biometrika*, **62**, 407–416.

Srivastava, M. S., and Worsley, K. J. (1986). Likelihood Ratio Tests for a Change in the Multivariate Normal Mean. *Journal of the American Statistical Association,* **81**, 199–204.

Worsley, K. J. (1983). Testing for a Two-phase Multiple Regression. *Technometrics,* **25**, 35–42.

Appendix: Derivation of (3.4) and (3.5)

Derivation of (3.4) and (3.5) consists of two steps.

I. Derivation of the posterior $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t|\mathbf{y})$

For any fixed t, since ϵ_{1i} are independent $N(0, \sigma_1^2)$, the likelihood, given the data $\mathbf{y}_1$, is then (see (3.2))

$$\begin{aligned} l(\boldsymbol{\beta}_1, \sigma_1^2|\mathbf{y}_1) &\propto \frac{1}{(\sigma_1^2)^{1/2}} \exp\Big[-\frac{(y_1 - \boldsymbol{\beta}_1'\mathbf{x}_1)^2}{2\sigma_1^2}\Big] \cdots \frac{1}{(\sigma_1^2)^{1/2}} \exp\Big[-\frac{(y_t - \boldsymbol{\beta}_1'\mathbf{x}_t)^2}{2\sigma_1^2}\Big] \\ &\propto \frac{1}{(\sigma_1^2)^{t/2}} \exp\Big\{ -\frac{1}{2\sigma_1^2}[(y_1 - \boldsymbol{\beta}_1'\mathbf{x}_1)^2 + \cdots + (y_t - \boldsymbol{\beta}_1'\mathbf{x}_t)^2]\Big\}. \end{aligned}$$

But

$$\begin{aligned} & (y_1 - \boldsymbol{\beta}_1'\mathbf{x}_1)^2 + \cdots + (y_t - \boldsymbol{\beta}_1'\mathbf{x}_t)^2 \\ = \; & (\mathbf{y}_1 - X_1\boldsymbol{\beta}_1)'(\mathbf{y}_1 - X_1\boldsymbol{\beta}_1) \\ = \; & (\mathbf{y}_1 - X_1\hat{\boldsymbol{\beta}}_1)'(\mathbf{y}_1 - X_1\hat{\boldsymbol{\beta}}_1) + (\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)'X_1'X_1(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1), \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'\mathbf{y}_1$ is the least-squares estimator of $\boldsymbol{\beta}_1$. The last equality of the above comes from the Pythagorean theorem due to the geometry of the least-square solution. (See also the equation (3.59), Guttman (1987).) Thus one has

$$l(\boldsymbol{\beta}_1, \sigma_1^2|\mathbf{y}_1) \propto \frac{1}{(\sigma_1^2)^{t/2}} \exp\Big\{ -\frac{1}{2\sigma_1^2}[S_1 + (\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)'X_1'X_1(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)]\Big\}, \quad (5.1)$$

where $S_1 = (\mathbf{y}_1 - X_1\hat{\boldsymbol{\beta}}_1)'(\mathbf{y}_1 - X_1\hat{\boldsymbol{\beta}}_1)$.

Similarly, from the last $n-t$ equations of (3.2), one obtains the likelihood, given the data $\mathbf{y}_2$,

$$l(\boldsymbol{\beta}_2, \sigma_2^2|\mathbf{y}) \propto \frac{1}{(\sigma_2^2)^{(n-t)/2}} \exp\Big\{ -\frac{1}{2\sigma_2^2}[S_2 + (\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)'X_2'X_2(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)]\Big\}, \quad (5.2)$$

where $S_2 = (\mathbf{y}_2 - X_2\hat{\boldsymbol{\beta}}_2)'(\mathbf{y}_2 - X_2\hat{\boldsymbol{\beta}}_2)$, and $\hat{\boldsymbol{\beta}}_2 = (X_2'X_2)^{-1}X_2'\mathbf{y}_2$ is the least-squares estimator of $\boldsymbol{\beta}_2$.

Since ϵ_{1i} and ϵ_{2i} are independent, it follows from (3.2), (5.1) and (5.2) that the likelihood function of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1^2, σ_2^2, and t is,

$$\begin{aligned} l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t|\mathbf{y}) \propto \; & \tfrac{1}{(\sigma_1^2)^{t/2}} \exp\{-\tfrac{1}{2\sigma_1^2}[S_1 + (\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)'X_1'X_1(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)]\} \\ & \times \tfrac{1}{(\sigma_2^2)^{(n-t)/2}} \exp\{-\tfrac{1}{2\sigma_2^2}[S_2 + (\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)'X_2'X_2(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)]\}. \end{aligned}$$

Suppose the following uniform prior in $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1^2, σ_2^2, and t (see (3.3)):

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t) \propto \frac{1}{\sigma_1^2\sigma_2^2} \;\text{ for }\; p+1 \le t \le n-p-1.$$

Then the posterior

$$\begin{aligned} p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t|\mathbf{y}) \;\propto\; & p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t) l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t|\mathbf{y}) \\ \propto \; & \tfrac{1}{(\sigma_1^2)^{t/2+1}} \exp\{-\tfrac{1}{2\sigma_1^2}[S_1 + (\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)'X_1'X_1(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)]\} \\ & \times \tfrac{1}{(\sigma_2^2)^{(n-t)/2+1}} \exp\{-\tfrac{1}{2\sigma_2^2}[S_2 + (\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)'X_2'X_2(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)]\}. \end{aligned}$$

II. Derivation of $p(t|\mathbf{y})$

In order to obtain $p(t|\mathbf{y})$, we need to integrate out $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1^2, and σ_2^2 from the posterior $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t|\mathbf{y})$. This is done through successive integration processes.

Note that a normal density for the p-dimensional random vector $\mathbf{X}$ has the form

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\},$$

where $\boldsymbol{\mu}$ represents the expected value of the random variable $\mathbf{X}$ and Σ is the $p \times p$ variance-covariance matrix of $\mathbf{X}$. ¿From $\int f(\mathbf{x})d\mathbf{x} = 1$, we have

$$\int \exp\{-(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\}d\mathbf{x} = (2\pi)^{p/2}|\Sigma|^{1/2}.$$

Using this equality, one can integrate out $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ from $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t|\mathbf{y})$ to obtain

$$\begin{aligned}
p(\sigma_1^2, \sigma_2^2, t|\mathbf{y}) &\propto \int\int p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, t|\mathbf{y})d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2 \\
&\propto \int \tfrac{1}{(\sigma_1^2)^{t/2+1}} \exp\{-\tfrac{1}{2\sigma_1^2}[S_1 + (\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)'X_1'X_1(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)]\}d\boldsymbol{\beta}_1 \\
&\quad \times \int \tfrac{1}{(\sigma_2^2)^{(n-t)/2+1}} \exp\{-\tfrac{1}{2\sigma_2^2}[S_2 + (\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)'X_2'X_2(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)]\}d\boldsymbol{\beta}_2 \\
&\propto \tfrac{\exp(-S_1/2\sigma_1^2)}{(\sigma_1^2)^{t/2+1}} \int \exp[-\tfrac{1}{2\sigma_1^2}(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)'X_1'X_1(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)]d\boldsymbol{\beta}_1 \\
&\quad \times \tfrac{\exp(-S_2/2\sigma_2^2)}{(\sigma_2^2)^{(n-t)/2+1}} \int \exp[-\tfrac{1}{2\sigma_2^2}(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)'X_2'X_2(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2)]d\boldsymbol{\beta}_2 \\
&\propto \tfrac{\exp(-S_1/2\sigma_1^2)}{(\sigma_1^2)^{t/2+1}} (2\pi)^{p/2} |(X_1'X_1/2\sigma_1^2)^{-1}|^{1/2} \\
&\quad \times \tfrac{\exp(-S_2/2\sigma_2^2)}{(\sigma_2^2)^{(n-t)/2+1}} (2\pi)^{p/2} |(X_2'X_2/2\sigma_2^2)^{-1}|^{1/2} \\
&\propto \tfrac{\exp(-S_1/2\sigma_1^2)}{(\sigma_1^2)^{t/2+1}} (2\pi)^{p/2} (2\sigma_1^2)^{p/2} |X_1'X_1|^{-1/2} \\
&\quad \times \tfrac{\exp(-S_2/2\sigma_2^2)}{(\sigma_2^2)^{(n-t)/2+1}} (2\pi)^{p/2} (2\sigma_2^2)^{p/2} |X_2'X_2|^{-1/2} \\
&\propto \tfrac{(|X_1'X_1||X_2'X_2|)^{-1/2}}{(\sigma_1^2)^{(t-p)/2+1}(\sigma_2^2)^{(n-t-p)/2+1}} \exp\{-(S_1/2\sigma_1^2 + S_2/2\sigma_2^2)\}.
\end{aligned}$$

Now observe that if $X = aU$, where $a > 0$ is a constant and U^{-1} is distributed according to χ_m^2, the probability density function of X is then

$$f(x) = 2^{-\frac{m}{2}} \Big[\Gamma\Big(\frac{m}{2}\Big)\Big]^{-1} a^{\frac{m}{2}} x^{-\frac{m}{2}-1} \exp\Big(-\frac{a}{2x}\Big).$$

Thus $\int f(x)dx = 1$, so that $\int \exp(-\frac{a}{2x})x^{-m/2-1}dx = 2^{m/2}\Gamma(\frac{m}{2})a^{-m/2}$. Using this equality and integrating out σ_1^2 and σ_2^2 from $p(\sigma_1^2, \sigma_2^2, t|\mathbf{y})$ lead to

$$\begin{aligned}
p(t|\mathbf{y}) &\propto \int\int p(\sigma_1^2, \sigma_2^2, t|\mathbf{y})d\sigma_1^2 d\sigma_2^2 \\
&\propto (|X_1'X_1||X_2'X_2|)^{-1/2} \int \tfrac{\exp(-S_1/2\sigma_1^2)}{(\sigma_1^2)^{(t-p)/2+1}} d\sigma_1^2 \int \tfrac{\exp(-S_2/2\sigma_2^2)}{(\sigma_2^2)^{(n-t-p)/2+1}} d\sigma_2^2 \\
&\propto (|X_1'X_1||X_2'X_2|)^{-1/2} 2^{(t-p)/2}\Gamma(\tfrac{t-p}{2}) S_1^{-(t-p)/2} 2^{(n-t-p)/2}\Gamma(\tfrac{n-t-p}{2}) S_2^{-(n-t-p)/2} \\
&\propto 2^{n/2-p}(|X_1'X_1||X_2'X_2|)^{-1/2}\Gamma(\tfrac{t-p}{2})\Gamma(\tfrac{n-t-p}{2}) S_1^{-(t-p)/2} S_2^{-(n-t-p)/2}.
\end{aligned}$$

Therefore

$$p(t|\mathbf{y}) = K2^{\frac{n}{2}-p}(|X_1'X_1||X_2'X_2|)^{-\frac{1}{2}}\Gamma\Big(\frac{t-p}{2}\Big)\Gamma\Big(\frac{n-t-p}{2}\Big)S_1^{-\frac{t-p}{2}}S_2^{-\frac{n-t-p}{2}},$$

where the constant K is such that $\sum_{t=p+1}^{n-p-1} p(t|\mathbf{y}) = 1$, i.e.,

$$K^{-1} = 2^{\frac{n}{2}-p}\sum_{t=p+1}^{n-p-1}(|X_1'X_1||X_2'X_2|)^{-\frac{1}{2}}\Gamma\Big(\frac{t-p}{2}\Big)\Gamma\Big(\frac{n-t-p}{2}\Big)S_1^{-\frac{t-p}{2}}S_2^{-\frac{n-t-p}{2}}.$$

Received October 13, 2001; accepted April 25, 2002

Dechang Chen
Department of Natural and Applied Sciences
University of Wisconsin-Green Bay
2420 Nicolet Drive
Green Bay, WI 54311, USA
chend@uwgb.edu

Michael Fries
School of Computer science, Telecommunications and
Information Systems
DePaul University
243 South Wabash Avenue
Chicago, IL 60604, USA
mfries@cti.depaul.edu

John M. Lyon
Department of Natural and Applied Sciences
University of Wisconsin-Green Bay
2420 Nicolet Drive
Green Bay, WI 54311, USA
lyonj@uwgb.edu