On Bayesian Analysis, Multi-Stage Bayes Decisions, and Generalized Inverses

Shintaro SONO
Associate Professor
Faculty of Economics
Hokkaido University

Preface

This paper is based on Chapters 1, 4, 6, 7, & 8 in Sono [14], and the writer is most indebted to Professor Yukio Suzuki, whose critical and methodological advices always stimulates the writer's brain.

Section A & Appendix A are concise & critical notes on Bayesian analysis. Sections B to D & Appendix D are formal developments of Bayesian approaches to some multi-stage decision problems, and generalized inverses are used to derive the general solutions in Sections C, D, & Appendix D.

Section A. Notes on Bayesian Analysis

1. Introduction

Bayesian philosophy is one of the most important mathematical methodologies. The logical foundation of Bayesian methodology was seriously considered by some statisticians, especially, by Leonard J. Savage [11]. For the concise expositions of the foundation, see, for example, DeGroot [5], Chapters 6 & 7, and Miyasawa [9], Chapter 1. But it must be remarked that these instant courses to Bayesian thought are too convenient to understand its profound significance. (The formal development for the conceptions of personal or subjective probability and utility in DeGroot [5], is different from the Savage's manner. Perhaps it might be necessary for practical purposes to simplify the consequences and acts

in Savage [11]. See, for example, Savage [11], Section 5. 5, pp. 82-91.)

On the other hand Bayesian approaches are considered to be powerful methods which propose logical procedures to accomplish the reasonable statistical inference and decisions, where, of course, the mathematical representations of some organizations' knowledge states on crucial unknown factors are required rather than abstract mathematical optimizations. (In fact many Bayesians, explicitly or implicitly, regard Bayesian methodology as one logical foundation for Operations Research. See, for example, Aoki [1], DeGroot [5], Martin [8], Miyasawa [9], and Suzuki [15].)

2. Objective Bayesian Inference

The objective Bayesian inference like Box and Tiao [4] seems to have the most convincing power in Bayesian inference based on so-called noninformative priors. Because the derivation of the noninformative priors based on approximate data translated likelihoods has a more reasonable methodological interpretation than the formal use of some invariance criteria like Jeffreys' rule. (See Box and Tiao [4], Section 1. 3, pp.25-60.) But some Bayesians might criticize the Box and Tiao's approach. In fact Berger says; "Box and Tiao (1973) actually recommended only locally noninformative priors, i.e., priors which behave like noninformative priors locally and then tail off to zero. The rules they recommend are the same as the rules calculated from noninformative priors, however, so the difference seems solely cosmetic." (See Berger [3], Section 4. 3, p.99.) A possible answer to this criticism is the following; The actual construction of priors in Box and Tiao [4] is based on the data translated or approximate data translated likelihoods and the prior independence assumptions on the suitably transformed parameters. Hence the formal application of

Jeffreys' rule must be cautiously examined. (See, for example, Box and Tiao [4], Section 1. 3. 6, pp.46-58, Section 5. 2. 2, pp.251-252, Appendix 5. 5, pp.303-304, Appendix 5. 6, pp. 304-315, Section 10. 3. 1, pp.531-533, and Sono [12].) But the criticism in Berger [3], Section 4. 3, p.99, is convincing to other constructions of noninformative priors.

3. Selection on Priors

The selection of priors including noninformative priors is one of the main subjects in Bayesian analysis. (See, for example, Berger [3], Chapter 3.) The robustness to the misspecifications of priors have been considered by some Bayesians, who criticize the wide use of conjugate priors. (For concise exposition, see, for example, Berger [3], Section 3. 3, Example 2, p.85, and Section 4. 6. 3, pp.139-142.) In practice, however, actual parameter spaces have natural constraints, and conjugate priors implicitly take account of their constraints by the relatively tight tails. On the other hand so-called robust priors have rather loose tails and, hence, have no information for the natural constraints. If an extreme value is observed, it is natural to suspect the assumed data generating model. And if practical Bayesians are interested in the specifications of priors, they should consider the constraints of parameter spaces and the validities of the constraints must be checked by their posterior analyses.

4. Asymptotic Property of Posteriors

Bayesian analyses use implicitly or explicitly the following simple asymptotic property; If the constrained parameter space includes the unknown fixed true parameter, the posterior on the parameter space is concentrated around the true parameter for large sample size, and if not, the posterior accumulates near the boundary points

which have the smallest distance from the true parameter in the sense of the Kullback-Leibler information. (See, for example, many graphs and tables in Box and Tiao [4] and Appendix A in this paper.)

In practice Bayesian analysis is indifferent to the preexperimental asymptotic argument. Because the analysis is based on the posterior distributions and the computational analysis of the posteriors. (See, for example, many approximation formulas and numerical comuputations for posterior analyses in Box and Tiao [4].)

5. Bayes Decision Problems

It should be remarked that the loss or utility functions are not indispensable elements for Bayes decisions. For example, the Bayes tests based on H. P. D. (Highest Posterior Density) regions need no loss functions. But, of course, the loss or utility functions are widely used as convenient instruments to save the decision maker's labor for the examination of posteriors. However it might be more effective to use the characteristics of posteriors like modal values instead of the decisions under the convenient loss functions.

Bayesian approach to statistical decision problems has logical flexibilities to sequential or dynamical situations. (See, for example, Aoki [1], Martin [8], Suzuki [16], [17], and Sections B to D in this paper.) But the difficulty, which was named "curse of dimensionality" by Richard Bellman [2], Section 5. 16, p.94, Section 15. 2, p.197, of DP-algorithm requires approximate Bayes procedures and error estimations of the approximations. (See, for example, Aoki [1], pp.224-241, Berger [3], Section 7. 4, Sono [13], and Section B in this paper.)

6. Criticisms

The following criticisms, (6.1) and (6.2), have been

repeatedly proposed by nonBayesians;

(6.1) The results in Bayesian analysis are often inconsistent with the common sense in nonBayesian analysis. Therefore Bayesian analysis is unreasonable method.

(6.2) The logic in Bayesian analysis is consistent with the likelihood principle. Therefore Bayesian analysis is invalid in the situations under which the likelihood principle is not effective.

However the validity of these criticisms is quite questionable. To (6.1), many Bayesians already pointed out the illogicalities in nonBayesian common sense like Neyman-Pearson statistics. (See, for example, Berger [3], Section 1. 6, pp.17-30, and Jaynes [7].) To (6.2), it should be remarked that, in proper Bayesian analysis, the likelihood principle in weak form is supported only in the post-experimental situation. (See, for example, Berger [3], Section 1. 6. 2, pp.23-28, Section 7. 7. 1, pp.352-354.) Hence, the use of the likelihood in Bayesian analysis is quite different from the nonBayesian use of it. On the other hand, in objective Bayesian inference, data generating models are used as prior knowledge to specify noninformative priors. But the inference is accomplished by the posterior distributions. Hence, the pre-experimental use of the likelihood is not necessary. (See, for example, Box and Tiao [4], Section 1. 3, pp.44-46, Section 2. 8. 1, pp.123-124.)

The next criticism, (6.3), seems to be convincing in formal sense;

(6.3) Proper Bayesian analysis is based on the subjective probability constructed by the decision maker. However, the construction is too complex for practical applications. Moreover, the error estimation of the

construction procedure is quite vague. Hence, in numerical sense, the applicability of proper Bayesian approach is questionable.

An answer to (6.3) is the use of convenient priors like noninformative priors, conjugate priors, locally uniform priors as reference priors, and so on. Practical abilities of these priors have been illustrated by many numerical examples and experiments. But, from logical view point, the use of convenient priors is different from the use of the decision maker's subjective prior. Hence, the convenient priors should be interpreted as some imaginary decision makers' priors. (In some situations the imaginary decision makers might be "imaginary enemies" in Operations Research.) It might be possible that the Bayesian approach based on computational analysis and numerical experiments using convenient priors should be called "computational Bayesian analysis".

7. Conclusion

In practice the difficulty in "computational Bayesian analysis" is only the "curse of dimensionality" of DP-algorithm and numerical integrations on multi-dimensional parameter space. Because, in general, Bayesian analysis has so-called "inference robustness". (See, for example, Box and Tiao [4], Section 3. 2, pp.152-156, Section 4. 2. 3, pp.208-209.)

Appendix A. A Simple Asymptotic Property of the Posterior

1. Assumptions and Result

Consider the pair, $(\Theta, d; \theta_*, \Gamma)$, where Θ is a topological space with a metric, d, θ_* is an unknown fixed point $\varepsilon\Theta$, and Γ is a compact subset $\subset\Theta$. $f(x \mid \theta)$ is a probability density function, given θ, with respect to some fixed σ-finite measure, $\mu(dx)$, on a measurable space, $(X, \mathbb{F})$. $f(x|\theta)$ is assumed to be continuous on $\theta\varepsilon\Gamma$ for each $x\varepsilon X$. (Therefore $f(x|\theta)$ is measurable on $(X\times\Gamma, \mathbb{F}\times\mathbb{B}$

$(\Gamma))$, where $\mathcal{B}(\Gamma)$ is the topological Borel field on Γ.) $f(x|\theta)>0$ for all $(x,\theta)\in X\times(\Gamma\cup\{\theta_*\})$ is also assumed. The set, $\{\theta\in\Gamma;\ d(\theta,\theta_0)<\delta\}$ is written $U(\theta_0,\delta;\Gamma)$ for any $(\theta_0,\delta)\in\Gamma\times]0,\infty[$. The Kullback-Leibler information, if exists, is written $I(\theta_*;\theta)$, i.e., $I(\theta_*;\theta):=E(\log(f(x|\theta_*)/f(x|\theta))|\theta_*)$, where $E(\cdot|\theta_*)$ is the expectation with $f(x|\theta_*)\mu(dx)$. The conventions, $\min\emptyset:=+\infty$ and $\log 0:=-\infty$, are used.

Result: Assume $E(\sup_{\theta\in\Gamma\cup\{\theta_*\}}|\log f(x|\theta)||\theta_*)<\infty$, and put $\Gamma_*:=\{\theta_0\in\Gamma;\ \min_{\theta\in\Gamma}I(\theta_*;\theta)=I(\theta_*;\theta_0)\}$. Let $P(d\theta)$ be a σ-finite measure on $(\Theta,\mathcal{B}(\Theta))$ such that $0<P(\Gamma)<\infty$ and $P(U(\theta_0,\delta;\Gamma))>0$ for all $(\theta_0,\delta)\in\Gamma_*\times]0,\infty[$. Then, for any closed set, $F\subset\Gamma-\Gamma_*$, and any $(\theta_0,\delta)\in\Gamma_*\times]0,\infty[$,

$$(\int_F P(d\theta)\prod_{k=1}^n f(x_k|\theta))/(\int_{U(\theta_0,\delta;\Gamma)}P(d\theta)\prod_{k=1}^n f(x_k|\theta))$$

$$\xrightarrow[n\to\infty]{} 0,\ \text{a.s.}\ (\prod_{k=1}^\infty f(x_k|\theta_*)\mu(dx_k)=:P_{\theta_*}^{(\infty)})$$

2. Proof

Put $f(x^n|\theta):=\prod_{k=1}^n f(x_k|\theta)$, $x^n:=(x_k;\ k=1,\ \dots\ ,n)$, and $r(x;\delta):=2\sup(|\log f(x|\theta_1)-\log f(x|\theta_2)|;\ d(\theta_1,\theta_2)\le\delta\ \&\ (\theta_1,\theta_2)\in\Gamma\times\Gamma)$. The following inequality is easily obtained;

$$(1)\quad \limsup_{n\to\infty}\log(\int_H P(d\theta)f(x^n|\theta)/\int_K P(d\theta)f(x^n|\theta))\le$$

$$E(r(x;\delta)|\theta_*)-\min_{\theta\in F}I(\theta_*;\theta)+\min_{\theta\in\Gamma}I(\theta_*;\theta),\ \text{a.s.}\ P_{\theta_*}^{(\infty)},$$

where $H\subset F\subset\Gamma-\Gamma_*$, F is closed in Θ, and $K\subset\Gamma$ satisfying $K\cap\Gamma_*\neq\emptyset\ \&\ P(K)>0$, and $\max(\mathrm{diam}(K),\mathrm{diam}(H))\le\delta$ $(\delta>0)$. (In general $\mathrm{diam}(S):=$

$\sup(d(s_1,s_2);(s_1,s_2)\in S\times S)$ for $S\neq\emptyset$, and $\mathrm{diam}(\emptyset):=-\infty$.)

From $E(r(x;\delta_0)|\theta_*)\to 0$ $(\delta_0\to 0)$, there exists δ_0 such that the right-hand side of $(1)<0$, for all $\delta\le\delta_0$, and, for $\delta\le\delta_0$, the inequality, (1), implies

$$(2)\quad \int_H P(d\theta)f(x^n|\theta)/\int_K P(d\theta)f(x^n|\theta)\xrightarrow[n\to\infty]{}0,\ \text{a.s.}\ P_{\theta_*}^{(\infty)}.$$

Take the finite covering of F, $\{H_i; i\in I\}$, satisfying $\mathrm{diam}(H_i)\leq\delta_0$, and put $K:=U(\theta_0, \delta;\Gamma)$ for any $(\theta_0, \delta)\in\Gamma_*\times]0, \delta_0]$. Then, using (2) and $\int_F P(d\theta)f(x^n|\theta)\leq\Sigma_{i\in I}\int_{H_i} P(d\theta)f(x^n|\theta)$, the result is obtained.

Section B. An approximate Bayes procedure for some Markov chains

(It is remarked that an idea in Aoki [1], pp.224-241, and Sono [13] is applicable to a multi-stage decision problem of some Markov chains with unknown transition probabilities.)

1. System and DP-algorithm

Consider the temporally homogeneous finite state Markov chain, $(X_k; k=0, \ldots, N+1)$, in which the transition probability from the k th stage to the k+1 th stage depends on the unknown but fixed parameter, θ, and the decision maker's k th act, a_k. Let $p(x_k, x_{k+1}|\theta, a_k)$ be the transition probability from $x_k\in S$ to $x_{k+1}\in S$ given $\theta\in\Theta$ and $a_k\in A$, where S, Θ, and A are the finite state space, the parameter space, and the finite action space, respectively. In the following discussion, in general, the sequence of symbols like $(s_i; i=0, 1, \ldots, k)$ is simply written s^k. The sequence of A-valued functions, $(\hat{a}_k(x^k, a^{k-1}); k=0, \ldots, N)$, where each $\hat{a}_k(\cdot)$ depends only on $(x^k, a^{k-1})\in S^{k+1}\times A^k$, is called a policy. If the prior on Θ, $P(d\theta)$, and the loss function on the Markov chain, $L(x^{N+1})$, are given, then, for each $n=0, 1, \ldots, N$, the conditional expected loss of the policy, $(\hat{a}_k(x^k, a^{k-1}); k=0, \ldots, N)$, given (x^n, a^n), is obtained by the formula,

$$(1.1)\quad E(L(X^{N+1})|x^n, a^n)=\int_\Theta P(d\theta|x^n, a^{n-1})E(L(X^{N+1})|\theta, x^n, a^n)$$

$$=\int_\Theta P(d\theta|x^n, a^{n-1})\Sigma_{x_{n+1}\in S}p(x_n, x_{n+1}|\theta, a_n)\Sigma_{x_{n+2}\in S}$$

$$p(x_{n+1}, x_{n+2}|\theta, \hat{a}_{n+1})\Sigma_{x_{n+3}\in S}p(x_{n+2}, x_{n+3}|\theta, \hat{a}_{n+2})\Sigma_{x_{n+4}\in S}$$

$$\cdots \Sigma_{x_{N+1} \in S} p(x_N, x_{N+1} | \theta, \hat{a}_N) L(x^{N+1}),$$

where $\hat{a}_k = \hat{a}_k(x^k, a^n, \hat{a}_{n+1}, \hat{a}_{n+2}, \cdots, \hat{a}_{k-1})$, $k=n+1, n+2, \cdots, N$, and $P(d\theta | x^n, a^{n-1})$ is the posterior given (x^n, a^{n-1}).

The optimal policy of the Bayes decision problem with the additive loss function, $\Sigma_{k=0}^{N} L_k$, where $L_k := L_k(x_k, x_{k+1})$ depends only on the k th transition, (x_k, x_{k+1}), for each $k=0, \cdots, N$, is constructed by the backward induction or DP-algorithm, i.e., by the recursive formula;

(1.2) $J_{N+1}^* := 0$,

$J_k^* := \min_{a_k \in A} E(L_k + J_{k+1}^* | x^k, a^k)$, and let $a_k^*(x^k, a^{k-1})$ be the minimizing a_k, then $(a_k^*(x^k, a^{k-1}); k=0, 1, \cdots, N)$ is the optimal policy.

In Martin [8] (1.2) is discussed under the assumptions; N is sufficiently large, the prior is a natural conjugate family, and the rewards are discounted by a known constant rate. However, in some Bayesian situations, the priors and losses are somewhat arbitrarily defined by the decision maker.

2. Approximation and Error Estimation

The approximation procedure is derived under the assumption; There exists a known function from $S \times S$ to some finite set Y, written $y(\cdot)$, such that the loss on the k th transition, L_k, is the function of $y(\cdot)$, i.e., $L_k = L_k(x_k, x_{k+1}) = L_k(y(x_k, x_{k+1}))$ for some function on Y, $L_k(\cdot)$, and the probability, $p(y|\theta, a) := \text{Prob}(y(x_k, X_{k+1}) = y | \theta, x^k, a^{k-1}, a_k = a) = \Sigma_{x \in \{x_{k+1} \in S;\ y(x_k, x_{k+1}) = y\}} p(x_k, x | \theta, a)$, depends only on $(\theta, a) \in \Theta \times A$ and $y \in Y$. For simplicity, the non-negativity of each L_k is also assumed. If $S = \{0.1, \cdots, M\}$ and the stochastic and loss matrices for each transition are cyclic, then the function, $y(\cdot)$, is always constructed by taking $Y := S$ and $y(x_k, x_{k+1}) := r \in Y$, where $x_{k+1} - x_k = r + q \cdot (M+1)$ for some integer q.

Take any policy, $(\hat{a}_k(x^k, a^{k-1}); k=0, 1, \cdots, N)$, and define

(2.1) $\hat{J}_{N+1}(\theta) := 0$,

$\hat{J}_k(\theta) := E(L_k + \hat{J}_{k+1}(\theta) | \theta, x^k, a^{k-1}, a_k = \hat{a}_k)$, and $\hat{J}_k := E(\hat{J}_k(\theta) | x^k, a^{k-1}, a_k = \hat{a}_k)$, $k=0, \ldots, N$.

To estimate the J^*'s from below, the following J^o's are employed;

(2.2) $J^o_{N+1}(\theta) := 0$,

$J^o_k(\theta) := \min_{a_k \in A} E(L_k + J^o_{k+1}(\theta) | \theta, x^k, a^k)$.

Let $a^o_k(\theta)$ be the minimizing a_k, and put $J^o_k := E(J^o_k(\theta) | x^k, a^{k-1})$, $k=0, \ldots, N$. From the assumption $a^o_k(\theta)$ is the minimizing $a \in A$ in $\min_{a \in A} \Sigma_{y \in Y} L_k(y) p(y | \theta, a)$ and $J^o_k(\theta)$ depends only on $(k, 0)$. Hence, from (2.2),

(2.3) $J^o_k(\theta) = \Sigma_{y \in Y} L_k(y) p(y | \theta, a^o_k(\theta)) + J^o_{k+1}(\theta)$, $k=0, \ldots, N$.

It is clear that $J^o_k \leq J^*_k \leq \hat{J}_k$ for each $k=0, \ldots, N+1$. In practice the relative errors, $(\hat{J}_k - J^*_k)/\hat{J}_k$, $k=0, \ldots, N$, might be needs, and, hence, $\hat{J}_k - J^o_k$, $k=0, \ldots, N$, should be assessed. Using (2.1) and (2.3), the recursive formulas, (2.4) and (2.5), are obtained;

(2.4) $\hat{\Delta} J_k(\theta) = \hat{D}_k(\theta) + E(\hat{\Delta} J_{k+1}(\theta) | \theta, x^k, a^{k-1}, a_k = \hat{a}_k(x^k, a^{k-1}))$,

where $\hat{\Delta} J_k(\theta) := \hat{J}_k(\theta) - J^o_k(\theta)$, and

$\hat{D}_k(\theta) := \sum_{y \in Y} L_k(y) \cdot (p(y | \theta, \hat{a}_k) - p(y | \theta, a^o_k(\theta)))$, $k=0, \ldots, N$.

(2.5) $\hat{\Delta} J_k = \hat{D}_k + E(\hat{\Delta} J_{k+1} | x^k, a^{k-1}, a_k = \hat{a}_k)$,

where $\hat{\Delta} J_k := \hat{J}_k - J^o_k$, and $\hat{D}_k := E(\hat{D}_k(\theta) | x^k, a^{k-1}, a_k = \hat{a}_k)$, $k=0, \ldots, N$. The $\hat{D}_k$, $k=0, \ldots, N$, are roughly estimated by (2.6);

(2.6) $\hat{D}_k = \sum_{a \in A} \int_{\Theta(a^o_k = a)} P(d\theta | x^k, a^{k-1}) \sum_{y \in Y} (p(y | \theta, \hat{a}_k) - p(y | \theta, a))$

$\leq (1 - P(\Theta(a^o_k = \hat{a}_k) | x^k, a^{k-1})) \sum_{y \in Y} L_k(y)$,

where $\Theta(a^o_k = a) := \{\theta \in \Theta;\ a^o_k(\theta) = a\}$, $k=0, \ldots, N$. From (2.5) &

(2.6), the formula, (2.7), is easily derived;

$$(2.7)\quad \Delta J_k = \hat{D}_k + \sum_{i=k+1}^{N} E(\hat{D}_i \mid x^k,\ a^{k-1},\ a_k = \hat{a}_k)$$
$$\leqq \sum_{i=k}^{N} (1 - E(P(\ominus(a_i^o = \hat{a}_i) \mid x^i,\ a^{i-1}) \mid x^k,\ a^{k-1}))\ \sum_{y \in Y} L_i(y),$$
$$k = 0, \ldots, N.$$

3. Simple Application

Let $\hat{a}_k^{(1)}$ and $\hat{a}_k^{(2)}$ be the minimizing and maximizing $\hat{a}_k$ in $\min(\hat{D}_k;\ \hat{a}_k \in A)$ and $\max(P(\ominus(a_k^o = \hat{a}_k) \mid x^k,\ a^{k-1});\ \hat{a}_k \in A)$, respectively. For each $k = 0, 1, \ldots, N$, take the dichotomous partition of S, $(I_k^{(1)}, I_k^{(2)})$, where I_k's depend only on (x^{k-1}, a^{k-1}), and define the policy, $(\hat{a}_k^{(3)};\ k = 0, 1, \ldots, N)$, by (3.1);

$$(3.1)\quad \hat{a}_k^{(3)} := \hat{a}_k^{(1)} \text{ for } x_k \in I_k^{(1)}, \text{ and } \hat{a}_k^{(2)} \text{ for } x_k \in I_k^{(2)},$$
$$k = 0, \ldots, N.$$

Using the formula, (2.6), we obtain

$$(3.2)\quad \hat{D}_k^{(3)} \leqq (1 - P(\ominus(a_k^o = \hat{a}_k^{(2)}) \mid x^k,\ a^{k-1}))\ \sum_{y \in Y} L_k(y),$$
$$k = 0, \ldots, N,$$

where, in general, the superscript, (3), denotes the use of the policy, $(\hat{a}_k^{(3)};\ k = 0, \ldots, N)$. From (3.2) and the inequalities, $E(P(\ominus(a_i^o = \hat{a}_i^{(2)}) \mid x^i,\ a^{i-1}) \mid x^k,\ a^{k-1}) \geq \max_{a \in A} E(P(\ominus(a_i^o = a) \mid x^i,\ a^{i-1}) \mid x^k,\ a^{k-1}) = \max_{a \in A} P(\ominus(a_i^o = a) \mid x^k,\ a^{k-1})$, $i = k, k+1, \ldots, N$, the rough estimation, (3.3), is derived;

$$(3.3)\quad \hat{\Delta}^{(3)} J_k \leqq \sum_{i=k}^{N} (1 - \max_{a \in A} P(\ominus(a_i^o = a) \mid x^k,\ a^{k-1}))\ \sum_{y \in Y} L_i(y),$$
$$k = 0, \ldots, N.$$

From (2.3) and (3.3), using $\hat{\Delta} J_k / \hat{J}_k \leqq \hat{\Delta} J_k / J_k^o$ and $\hat{J}_k - J_k^* \leqq \hat{\Delta} \hat{J}_k$, $k = 0, \ldots, N$, we obtain

(3.4) $(\hat{J}_k - J_k^*)/\hat{J}_k \leq$

$$\frac{\sum_{i=k}^{N} (1 - \max_{a\in A} P(\ominus(a_i^o=a)|x^k, a^{k-1})) \sum_{y\in Y} L_i(y)}{\sum_{i=k}^{N} \sum_{y\in Y} L_i(y) E(p(y|\theta, a_i^o(\theta))|x^k, a^{k-1})}, \quad k=0,\ldots, N.$$

For the simplest case, $S=\{0,1\}$, $A=\{a_0, a_1\}$, $p(0|\theta, a)= p(0,0|\theta,a)= p(1,1|\theta,a)$, $p(1|\theta,a)=p(0,1|\theta,a)=p(1,0|\theta,a)$, $L_k(1)>L_k(0):=0$, $k=0, \ldots, N$, the right-hand side of (3.4) is reduced to $(1-\max_{a\in A} P(\ominus(a^o=a)|x^k, a^{k-1})) / (1-E(p(0 | \theta, a^o(\theta))|x^k, a^{k-1}))$, $k=0, \ldots, N$, where $a^o(\theta):=a_0$ for $p(0|\theta, a_0)>p(0|\theta, a_1)$, a_1 for $p(0|\theta, a_1)> p(0|\theta, a_0)$, and arbitrary for $p(0|\theta, a_0)=p(0|\theta, a_1)$.

In practice, instead of the conditioning variables, (x^k, a^{k-1}), $k=0, \ldots, N$, some sufficient statistics of them are used. In fact the statistics are recursively defined by $M_0:=0$ and $M_{k+1}:=M_k+T_k$, $k=0,\ldots, N$, where M's and T's are $S\times S\times A$ matrices and, if the k th transition, (x_k, x_{k+1}), occurs under the k th act, a_k, then the (x_k, x_{k+1}, a_k) element of T_k is one and all other elements are zeros. (See, for example, Martin [8], Chapter 2.) But, for some formal discussions like above, the use of the sufficient statistics is not necessary.

Section C. Notes on Least Squares Method and Bayesian Method

Abstract

In some statistical models least squares method and Bayesian method give some similar results for the analysis of the models, and many Bayesians remark that such similarity is in numerical sense and not in logical sense and, often, the results proposed by Bayesian method are more natural than the results proposed by least squares method. In this Section one of such statistical models, which is a linear dynamical system with two kinds of

noises which are called plant and observation noises, is discussed.

1. System

Consider the model,

(1.1) $x_{i+1} = A_i x_i + q_i$, and

(1.2) $y_i = H_i x_i + r_i$,

where i=0, ..., N-1 (N is a positive integer), state vectors x's and observation vectors y's are real valued n and m dimensional column vectors, respectively. A's and H's are deterministic real valued (n,n) and (m,n) matrices. Plant and observation noises, q's and r's, are n and m dimensional vector valued random variables on some probability space. The initial state vector x_0 is unknown but fixed. And, for simplicity, it is assumed that each q_i and r_i have zero mean vectors and finite variance matrices written as Q_i and R_i, respectively, and the family of all q's and r's is stochastically independent. Remark that each A_i, Q_i, and R_i may be singular and each H_i may be not full rank. Therefore elementary properties of generalized inverse of matrices are used in the discussion.
(See, for example, Iri and Kan [6], Chapter 8, or Rao [10], Chapter 1.)

M. Aoki considers the least squares and Bayes estimations of the states, x_i, i=0, ..., N-1, under the assumptions that all A's & R's are non-singular and all q's are zero vectors, or that all Q's and R's are non-singular, respectively. He suggests the relation between the two methods in general case, (1.1) & (1.2), in vague terms (see, Aoki [1], pp.155-161, pp.173-179, p.162, in Chapter V).

In the following the procedures which are derived by the two methods for the estimation of x's are discussed

from the view point of the statistical interpretations of the procedures in general and rigorous manner. The notation, (1.3), is used:

$$(1.3)\quad \langle\langle a,b\rangle\rangle_M := a'Mb,$$

where a, b, and M are column vectors and matrix, respectively, assuming the multiplication is well-defined. Especially if M is symmetric positive definite, then (1.3) is identical with the inner product of (a, b) with respect to the metric M. In the following discussion, usually, M is not positive but nonnegative definite and, then (1.3) is only pseudo-inner product of (a, b). If a=b, then (1.3) is written as only $||a||_M^2$. In this Section, unless otherwise stated, all vectors are real valued column vectors and all matrices are also real. $\mathbb{R}^n$ is the n dimensional Euclidean space, E_n is the (n,n) unit matrix, and, in general, Im(M) & Ker(M) is the image & kernel of (m,n) matrix M, i.e., $Im(M):=\{Mx;\ x\in\mathbb{R}^n\}$ & $Ker(M):=\{x;\ Mx=0\}$. M^- represents any generalized inverse of M. The sequence $(y_j)_{j=0}^i$ is written y^i, and put $y^{-1}:=0$.

2. Estimation by Least Squares Method

Consider the estimation problem of the states, x_i, $i=0,\ldots,N-1$, as the following type;

$$(2.1)\quad J:=\sum_{i=0}^{N-1}\left(||x_i-A_{i-1}x_{i-1}||_{T_{i-1}}^2+||y_i-H_ix_i||_{D_i}^2\right)$$

where all T's and D's are symmetric nonnegative definite matrices and $T_{-1}:=0$, and J is minimized with respect to $x_i\in\mathbb{R}^n$, $i=0,\ldots,N-1$, and the minimizing values are the estimates of x's, i.e., the least squares estimation of x_i, $i=0,\ldots,N-1$, based on the data, y_i, $i=0,\ldots,N-1$, is considered. It is well-known that this type of minimization problem is systematically solved by the backward induction or, so-called, DP-algorithm, i.e., the functional equation,

(2.2) $J_i := ||x_{i-1} - A_{i-2}x_{i-2}||^2_{T_{i-2}} + ||y_{i-1} - H_{i-1}x_{i-1}||^2_{D_{i-1}} + \hat{J}_{i+1}$,

where $\hat{J}_i := \min \{J_i ; x_{i-1} \in \mathbb{R}^n\}$, $J_{N+1} := 0$, and $i=1,\ldots, N$.

Remark that $\hat{x}_{i-1}$ which is minimizing J_i is the function of x_j, $j=0,\ldots, i-2$, and the $\hat{x}_i$, $i=0,\ldots, N-1$, minimize $\hat{J}_1$ by substituting $\hat{x}_j$, $j=0,\ldots, i-2$, for x's in $\hat{x}_{i-1}$ and the minimized $\hat{J}_1$ is equal to min J and, therefore, the reason why the functional equation, (2.2), gives the solution of the minimization problem of (2.1) is almost clear. (2.2) is solved as the following: If $i=N$ and x_j, $j=0,\ldots, N-2$, are given, then

(#.N) $J_N = ||x_{N-1} - \hat{x}_{N-1}||^2_{S_{N-1}} + \hat{J}_N$,

where

$\hat{x}_{N-1} := S^-_{N-1}(T_{N-2}A_{N-2}x_{N-2} + H'_{N-1}D_{N-1}y_{N-1})$,

$S_{N-1} := T_{N-2} + H'_{N-1}D_{N-1}H_{N-1}$,

$\hat{J}_N := ||A_{N-2}x_{N-2} - z_{N-2}||^2_{I_{N-2}} + R_{N-1}$,

$I_{N-2} := T_{N-2} - T_{N-2}S^-_{N-1}T_{N-2}$,

$z_{N-2} := I^-_{N-2}T_{N-2}S^-_{N-1}H'_{N-1}D_{N-1}y_{N-1}$,

$\hat{R}_{N-1} := ||y_{N-1}||^2_{D^{(N-1)}_{N-1}}$,

$D^{(N-1)}_{N-1} := D_{N-1} - D_{N-1}H_{N-1}S^-_{N-1}(S_{N-1} + T_{N-2}I^-_{N-2}T_{N-2})S^-_{N-1}H'_{N-1}D_{N-1}$,

and it is easily shown that I_{N-2} and $D^{(N-1)}_{N-1}$ are symmetric nonnegative definite matrices and, excepting $\hat{x}$ & z, these quantities are uniquely determined independently of the selections of the generalized inverses. (The derivation of (#.N) is the work in matrix analysis, using generalized inverses, $Im(S_{N-1}) = Im(T_{N-2}) + Im(H'_{N-1}D_{N-1}H_{N-1})$, and completion and combination of quadratic forms.)

In general, if

$$\hat{J}_{i+1} = \| A_{i-1} x_{i-1} - z_{i-1} \|^2_{I_{i-1}} + \hat{R}_i \ ,$$

where z_{i-1} , I_{i-1} , and $\hat{R}_i$ are including no terms of x's, z & $\hat{R}$ are including y's, and I_{i-1} is symmetric nonnegative definite, then

$$(\#.i) \quad J_i = \| x_{i-1} - \hat{x}_{i-1} \|^2_{S_{i-1}} + \hat{J}_i \ ,$$

where

$$\hat{x}_{i-1} := S^{-}_{i-1} (T_{i-2} A_{i-2} x_{i-2} + H'_{i-1} D_{i-1} y_{i-1} + A'_{i-1} I_{i-1} z_{i-1}) \ ,$$

$$S_{i-1} := T_{i-2} + H'_{i-1} D_{i-1} H_{i-1} + A'_{i-1} I_{i-1} A_{i-1} \ ,$$

$$\hat{J}_i := \| A_{i-2} x_{i-2} - z_{i-2} \|^2_{I_{i-2}} + R_{i-1} \ ,$$

$$I_{i-2} := T_{i-2} - T_{i-2} S^{-}_{i-1} T_{i-2} \ ,$$

$$z_{i-2} := I^{-}_{i-2} T_{i-2} S_{i-1} (H'_{i-1} D_{i-1} y_{i-1} + A'_{i-1} I_{i-1} z_{i-1}) \ ,$$

$$\hat{R}_{i-1} := \hat{R}_i + \| y_{i-1} \|^2_{D^{(i-1)}_{i-1}} + \| z_{i-1} \|^2_{I^{(i-1)}_{i-1}} - 2 \langle\langle z_{i-1}, L_{i-1} y_{i-1} \rangle\rangle E_n \ ,$$

$$D^{(i-1)}_{i-1} := D_{i-1} - D_{i-1} H_{i-1} S^{-}_{i-1} (S_{i-1} + T_{i-2} I^{-}_{i-2} T_{i-2}) S^{-}_{i-1} H'_{i-1} D_{i-1} \ ,$$

$$I^{(i-1)}_{i-1} := I_{i-1} - I_{i-1} A_{i-1} S^{-}_{i-1} (S_{i-1} + T_{i-2} I^{-}_{i-2} T_{i-2}) S^{-}_{i-1} A'_{i-1} I_{i-1} \ ,$$

$$L_{i-1} := I_{i-1} A_{i-1} S^{-}_{i-1} (S_{i-1} + T_{i-2} I^{-}_{i-2} T_{i-2}) S^{-}_{i-1} H'_{i-1} D_{i-1} \ .$$

It is easily shown that I_{i-2} , $D^{(i-1)}_{i-1}$, and $I^{(i-1)}_{i-1}$ are symmetric nonnegative definite and, excepting $\hat{x}$ & z, these quantities are uniquely determined independently of the selections of the generalized inverses. (The derivation of (#.i) is the work in matrix analysis, using generalized inverses, $\mathrm{Im}(S_{i-1}) = \mathrm{Im}(T_{i-2}) + \mathrm{Im}(H'_{i-1} D_{i-1} H_{i-1}) + \mathrm{Im}(A'_{i-1} I_{i-1} A_{i-1})$, and combination of quadratic forms, etc..)

Therefore, from (#.N) & (#.i) (i=1,..., N), (by the backward induction), the functional equation, (2.2), is recursively solved.

Remark that the sequential estimates of x's by least

squares method are obtained by replacing N-1 in (#.N) by i=0,..., N-1, as

$$(2.3)\quad \ddot{x}_i := \hat{S}_i^{-} (T_{i-1} A_{i-1} \ddot{x}_{i-1} + H_i' D_i y_i) \ , \quad i=0,\ldots, N-1,$$

especially,

$$\ddot{x}_0 := (H_0' D_0 H_0)^{-} H_0' D_0 y_0 \ ,$$

where $\hat{S}_i := T_{i-1} + H_i' D_i H_i$, because $\ddot{x}_i$ in (2.3) is depending only on y_i and already determined estimates, $\ddot{x}_j$, $j=0,\ldots,$ i-1, and, by replacing N-1 in $\hat{x}_{N-1}$ by j, $\hat{x}_{N-1}$ is identical with $\ddot{x}_j$, $j=0,\ldots,$ i.

3. Estimation by Bayes Method

Consider the model, (1.1) & (1.2), under the assumption that all q's and r's are Gaussian. (There is no assumption such that the variance matrices of the noises are non-singular, therefore these Gauss distributions may be degenerate.) It is well-known and easily proved that, if

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{matrix} \} p_1 \\ \} p_2 \end{matrix} \sim N_{p_1+p_2} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \begin{matrix} \} p_1 \\ \} p_2 \end{matrix} , \ \begin{pmatrix} \underbrace{\Sigma_{11}}_{p_1} & \underbrace{\Sigma_{12}}_{p_2} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{matrix} \} p_1 \\ \} p_2 \end{matrix} \right),$$

then $x_2 | x_1 \sim N_{p_2} (\mu_{2.1}, \Sigma_{22.1})$, where

$$\mu_{2.1} := \mu_2 + \Sigma_{21} \Sigma_{11}^{-} (x_1 - \mu_1),$$

$$\Sigma_{22.1} := \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-} \Sigma_{12} \ ,$$

and $\mu_{2.1}$ and $\Sigma_{22.1}$ are uniquely determined because $\mathrm{Ker}(\Sigma_{21}) \supset \mathrm{Ker}(\Sigma_{11})$ (i.e., $\mathrm{Im}(\Sigma_{11}) \supset \mathrm{Im}(\Sigma_{12})$ and $x_1 - \mu_1 \in \mathrm{Im}(\Sigma_{11})$ a.s.). (In the estimation by Bayes method this proposition is used to compute the posterior distributions of x's.)

For Bayes procedure the natural conjugate prior distribution of "unknown but fixed parameter" x_0 is introduced, i.e.,

$x_0 \sim N_n(\hat{\mu}_0, \hat{\Sigma}_0)$,

where x_0 is stochastically independent of all q's and r's and $\hat{\Sigma}_0$ is nonnegative definite. (In mathematical manner the product probability space of the probability space, on which all q's and r's are defined, and the probability space, ($\mathbb{R}^n$, topological Borel field of $\mathbb{R}^n$, $N_n(\hat{\mu}_0, \hat{\Sigma}_0)$), is introduced. But the statement as the above is usually used in Bayesian analysis. (Of course the expectation operator, E, in the following discussion is defined on this product probability space.))

Consider the sequential estimation of x_i, i=0,..., N-1, which is equivalent to the recursive computation of $p(x_i | y^i)$, i=0,..., N-1. (The posterior mean of x_i is the estimate of x_i.) $p(x_i|y^i)$'s are derived as the following: (Put $A_{-1}\mu_{-1} := \hat{\mu}_0$, i=0,..., N-1.): From the definition of the system

$$(3.1)\quad E\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} \middle| y^{i-1}\right) = \begin{pmatrix} A_{i-1}\mu_{i-1} \\ H_i A_{i-1}\mu_{i-1} \end{pmatrix} \text{ and}$$

$$(3.2)\quad V\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} \middle| y^{i-1}\right) = \begin{pmatrix} \hat{\Sigma}_i & \hat{\Sigma}_i H_i' \\ H_i\hat{\Sigma}_i & H_i\hat{\Sigma}_i H_i' + R_i \end{pmatrix},$$

where $\mu_{i-1} := E(x_{i-1} | y^{i-1})$ and

$$(3.3)\quad \hat{\Sigma}_i := A_{i-1}\Sigma_{i-1}A_{i-1}' + Q_{i-1} \text{ and}$$

$$\Sigma_{i-1} := V(x_{i-1} | y^{i-1}), \text{ therefore,}$$

$$(3.4)\quad \mu_i := E(x_i | y^i) = A_{i-1}\mu_{i-1} + K_i(y_i - H_i A_{i-1}\mu_{i-1})$$

and

$$(3.5)\quad \Sigma_i := V(x_i | y^i) = \hat{\Sigma}_i - \hat{\Sigma}_i H_i'(H_i\hat{\Sigma}_i H_i' + R_i)^{-} H_i\hat{\Sigma}_i \quad ,$$

where $K_i := \hat{\Sigma}_i H_i'(H_i\hat{\Sigma}_i H_i' + R_i)$.

Consider the estimation of x_i, i=0,..., N-1, based on the data, y^{N-1}, which is equivalent to the recursive

computation of $p(x_i|y^{N-1})$, i=0,..., N-1. ($E(x_i|y^{N-1})$ is the estimate of x_i.) $p(x_i|y^{N-1})$'s are derived as the following: (Put $A_{-1}\mu_{-1} := \hat{\mu}_0$, i=0,..., N-1.) From the definition of the system

$$(3.6)\quad E\left(\begin{pmatrix} x_{i-1} \\ x_i \end{pmatrix} | y^{i-1}\right) = \begin{pmatrix} \mu_{i-1} \\ A_{i-1}\mu_{i-1} \end{pmatrix} \quad \text{and}$$

$$(3.7)\quad V\left(\begin{pmatrix} x_{i-1} \\ x_i \end{pmatrix} | y^{i-1}\right) = \begin{pmatrix} \Sigma_{i-1} & \Sigma_{i-1}A_{i-1}' \\ A_{i-1}\Sigma_{i-1} & \hat{\Sigma}_i \end{pmatrix},$$

therefore,

$$(3.8)\quad E(x_{i-1}|x_i, y^{i-1}) = \mu_{i-1} + \Sigma_{i-1}A_{i-1}'\hat{\Sigma}_i^{-}(x_i - A_{i-1}\mu_{i-1})$$

and

$$(3.9)\quad V(x_{i-1}|x_i, y^{i-1}) = \Sigma_{i-1} - \Sigma_{i-1}A_{i-1}'\hat{\Sigma}_i^{-}A_{i-1}\Sigma_{i-1},$$

and, using (3.8), (3.9), and "$((q_j, r_j))_{j=i}^{N-1}$ is stochastically independent of $(x_0, (q_j, r_j)_{j=0}^{i-1})$",

$$(3.10)\quad \mu_{i-1,N-1} := E(x_{i-1}|y^{N-1})$$
$$= \mu_{i-1} + \Sigma_{i-1}A_{i-1}'\hat{\Sigma}_i^{-}(\mu_{i,N-1} - A_{i-1}\mu_{i-1}),$$

$$(3.11)\quad \Sigma_{i-1,N-1} := V(x_{i-1}|y^{N-1})$$
$$= \Sigma_{i-1} - \Sigma_{i-1}A_{i-1}'\hat{\Sigma}_i^{-}(\hat{\Sigma}_i - \Sigma_{i,N-1})\hat{\Sigma}_i^{-}A_{i-1}\Sigma_{i-1},$$

$$(3.12)\quad \hat{\Sigma}_i := V(x_i|y^{i-1})$$
$$= A_{i-1}\Sigma_{i-1}A_{i-1}' + Q_{i-1} = \Sigma_{i,N-1} + V(\mu_{i,N-1}|y^{i-1}),$$

where $\mu_{i,N-1} := E(x_i|y^{N-1})$ and $\Sigma_{i,N-1} := V(x_i|y^{N-1})$ (Remark the formulas, $E(x|y)=E(E(x|y, z)|y)$ and $V(x|y)=V(E(x|y, z)|y) + E(V(x|y, z)|y)$.)

4. Criticism

The procedure of least squares method in Subsection 2 is, of course, applicable to the Gaussian system in Section 3. But this procedure is only the mathematical

optimization and, hence, seems to be too general to propose the reasonable statistical representations for the knowledge states of unknown variables.

In the case of the sequential estimation the two methods propose quite similar results, i.e., from (2.3)

$$(4.1)\quad \ddot{x}_i = \hat{S}_i^{-} \hat{S}_i A_{i-1} \ddot{x}_{i-1} + \hat{S}_i^{-} H_i' D_i (y_i - H_i A_{i-1} \hat{x}_{i-1}),\ i=1,\ldots,N-1,$$

$$(4.2)\quad \ddot{x}_0 = (H_0' D_0 H_0)^{-} H_0' D_0 y_0, \text{ and, from (3.4),}$$

$$(4.3)\quad \mu_i = \hat{\mu}_i + K_i (y_i - H_i \hat{\mu}_i),\ i=0,\ldots,N-1, \text{ where } \hat{\mu}_i := A_{i-1}\mu_{i-1} \text{ for } i=1,\ldots,N-1.$$

(It is not necessary that the generalized inverses of $\hat{S}_i$ in (4.1) are identical.)

The two procedures, (4.1) and (4.2), are essentially equivalent excepting the initial estimates. In (4.3) $\hat{\mu}_0$ is specified by the assumption of the prior knowledge, for example, if the noninformative case is considered, then let $\hat{\Sigma}_0 = \hat{m} E_n (\hat{m} \to \infty)$ ($\hat{\mu}_0$ is fixed), and in this case, under the assumption of the positive definiteness of R_0 and $H_0' R_0^{-1} H_0$, when $\hat{m} \to \infty$,

$$(4.4)\quad \mu_0 \to (H_0' R_0^{-1} H_0)^{-1} H_0' R_0^{-1} y_0,$$

$$(4.5)\quad \Sigma_0 \to (H_0' R_0^{-1} H_0)^{-1}.$$

(4.4) and (4.5) give a Bayesian justification of (4.2).

In the case of the estimation based on y^{N-1}, from $\hat{x}_{i-1}$ and (4.6),

$$(4.6)\quad \hat{x}_{i-1} = \ddot{x}_{i-1} + (S_{i-1}^{-} - \hat{S}_{i-1}^{-})(T_{i-2} A_{i-2} \hat{x}_{i-2} + H_{i-1}' D_{i-1} y_{i-1}) + S_{i-1}^{-} A_{i-1}' I_{i-1} z_{i-1},$$

and, from (3.10),

$$(4.7)\quad \mu_{i-1,N-1} = \mu_{i-1} + \Sigma_{i-1} A_{i-1}' \hat{\Sigma}_i^{-} (\mu_{i,N-1} - \hat{\mu}_i),$$

$i=1,\ldots, N-1$, (Remark $\mu_{N-1,N-1}= \mu_{N-1}$.)

In (4.7) the relation among the estimate based on y^{N-1}, $\mu_{i-1,N-1}$, the sequential estimate, μ_{i-1}, and the predictor, $\hat{\mu}_i$, is represented in a transparent manner, but in (4.6) the relation is not clear. And in Bayes method the precisions are systematically estimated by using (3.3), (3.5), and (3.11), i.e., by using the posterior variance matrices. But in least squares method the precisions of the estimates and the predictor of each state is not clear. (For example the $\hat{J}$'s in (2.2) are too complex to have statistical interpretations.)

Section D. A Note on a Multi-Stage Decision Problem with Matric-Variate Loss Functions

1. System and Assumptions

Formulas and notations in Appendix D are used freely. The coefficient field is assumed to be the real or complex number field. In general, for the sets of matrices, M_1, M_2, and M_3, $M_1+M_2M_3$ means the set, $\{m_1+m_2m_3;(m_1, m_2, m_3)\varepsilon M_1\times M_2\times M_3\}$. The set of all Hermite and nonnegative definite matrices in M(n) is written HNND(n). (In general, M(q, p) is the set of all (q, p) matrices with the assumed number field elements, and M(p):=M(p, p). See Appendix D for other notations.)

The model is given by the equation, (1.1);

$$(1.1)\quad X_{i+1} = A_iX_i + B_iU_i + C_iW_i,\quad i=0,1,\ldots, N,$$

where X's are the M(n, n′) valued state variables, A_i, B_i, and C_i are M(n), M(n, p), and M(n, q) valued random variables, respectively, W's are M(q, n′) valued disturbances, and U_i is the M(p, n′) valued i th stage decision for each i=0,1,..., N. The observation equations are given by (1.2);

$$(1.2)\quad Y_i= F_i(X_i, Z_i),\quad i=0,1,\ldots, N,$$

where Y's are the M(m, m′) valued observations, Z's are M(r, r′) valued random variables, and F_i is the function of (X_i, Z_i) for each $i=0,1,\ldots, N$. The loss for the i th transition, L_i, is given by (1.3);

$$(1.3)\quad L_i = ||X_{i+1}||^2_{V_{i+1}} + ||U_i||^2_{P_i}, \quad i=0,1,\ldots, N,$$

where $V_{i+1} \varepsilon$ HNND(n) and $P_i \varepsilon$ HNND(p), $i=0,1,\ldots, N$. (In general, $((A, B))_C := A^*CB$ and $||A||^2_D := ((A, A))_D$. See Appendix D.) Hence L_i is the HNND(n′) valued loss for each i=0,1, ... , N.

DP-algorithm, (1.4), is considered;

$$(1.4)\quad J^{(*)}_{N+1} = J_{N+1} = 0,$$

$$J^{(*)}_i := \min_{U_i \varepsilon M(p, n')} J_i(U_i),$$

$$J_i(U_i) := E(L_i + J^{(*)}_{i+1} \mid Y^i, U^i), \quad i=0, 1,\ldots, N,$$

where the minimizing U_i, written $U^{(*)}_i$, is defined by (1.5);

$$(1.5)\quad J_i(U_i) - J_i(U^{(*)}_i) \varepsilon \text{ HNND}(n') \text{ for all } U_i \varepsilon M(p, n'),$$

and, in general, the sequence of symbols like $(M_i; i=0, 1, \ldots, j)$ is written M^j. (Hence $Y^i = (Y_k; k=0, \ldots, i)$ and $U^i = (U_k; k=0, \ldots, i)$.) $(U^{(*)}_i; i=0,1,\ldots, N)$ is the optimal policy for the loss function, $\Sigma^N_{i=0} L_i$. Hence, the policy also minimizes the trace and the maximum eigenvalue of $E(\Sigma^N_{i=0} L_i \mid Y_0)$.

The optimal policies are constructed under the assumptions, (1.6), (1.7), and (1.8);

(1.6) X_0, $((A_i, B_i, C_i, W_i); i=0,1,\ldots, N)$, and $(Z_i; i=0,1, \ldots, N)$ are stochastically independent.

(1.7) $((A_i, B_i, C_i, W_i); i=0,1,\ldots, N)$ is the stochastically independent sequence.

(1.8) $E(||X_i - E(X_i|Y^i)||^2_{Q_i}|Y^i)$, where Q_i is recursively defined in Subsection 2, is independent of values of Y^i for each $i=0,1,\ldots, N$.

2. Construction

The DP-algorithm, (1.4), is solved by the backward induction, and, hence, the optimal policies are constructed. (See, for example, Aoki [1] Chapter II.) In fact, using the formulas in Appendix D, the following results, (2.1)-(2.18), are derived;

$$(2.1)\quad J_i(U_i) = ||U_i + S_i^- N_i||^2_{S_i} + E(||X_i + I_i^- T_i||^2_{I_i}|Y^i) + R_i, \quad i=0,\ldots, N,$$

$$(2.2)\quad J_i^{(*)} = E(||X_i + I_i^- T_i||^2_{I_i}|Y^i) + R_i, \quad i=0,\ldots, N,$$

where

$$(2.3)\quad S_N := 0,$$

$$S_i := P_i + E(||B_i||^2_{V_{i+1}+I_{i+1}}), \quad i=0,\ldots, N,$$

$$(2.4)\quad I_N := 0,$$

$$I_i := E(||A_i||^2_{V_{i+1}+I_{i+1}}) - Q_i$$

$$= E(||A_i - B_i S_i^- E(B_i^*(V_{i+1}+I_{i+1})A_i)||^2_{V_{i+1}+I_{i+1}})$$

$$+ ||E(B_i^*(V_{i+1}+I_{i+1})A_i)||^2_{P_i}, \quad i=0,\ldots, N,$$

$$(2.5)\quad Q_N := 0,$$

$$Q_i := ||E(B_i^*(V_{i+1}+I_{i+1})A_i)||^2_{S_i^-}, \quad i=0,\ldots, N,$$

$$(2.6)\quad T_N := 0,$$

$$T_i := E(A_i^*(V_{i+1}+I_{i+1})W_i^{(i)})$$

$$- E(A_i^*(V_{i+1}+I_{i+1})B_i)S_i^- E(B_i^*(V_{i+1}+I_{i+1})W_i^{(i)})$$

$$= E((A_i - B_i(S_i^-)^* E(B_i^*(V_{i+1} + I_{i+1})A_i))^* (V_{i+1} + I_{i+1})W_i^{(i)}),$$

$$i=0,\ldots, N,$$

(2.7) $W_N^{(N)} := 0,$

$$W_i^{(i)} := C_i W_i + (V_{i+1} + I_{i+1})^- T_{i+1}, \quad i=0,\ldots, N,$$

(2.8) $R_N^{(N)} := 0,$

$$R_i^{(i)} := ||T_{i+1}||^2_{I_{i+1}^-} - (V_{i+1} + I_{i+1})^- + R_{i+1},$$

$$i=0,\ldots, N,$$

(2.9) $R_N := 0,$

$$R_i := E(||W_i^{(i)}||^2_{V_{i+1} + I_{i+1}}) - D_i$$

$$+ E(||X_i - E(X_i|Y^i)||^2_{Q_i}|Y) + R_i^{(i)}, \quad i=0,\ldots, N,$$

(2.10) $D_N := 0,$

$$D_i := ||E(B_i^*(V_{i+1} + I_{i+1})W_i^{(i)})||^2_{S_i^-} - ||T_i||^2_{I_i^-},$$

$$i=0,\ldots, N,$$

(2.11) $N_i := E(B_i^*(V_{i+1} + I_{i+1})A_i)M_i$

$$+ E(B_i^*(V_{i+1} + I_{i+1})W_i^{(i)}), \quad i=0,\ldots, N,$$

(2.12) $M_i := E(X_i|Y^i) := E(X_i|Y^i, U^{i-1}), \quad i=0,\ldots, N.$

(The optimal policies are obtained from the formulas, (2.13)-(2.18).)

(2.13) $U_i^{(*)} = -S_i^- N_i + K_i$

$$= -(\Lambda_i M_i + \Delta_i) + K_i, \quad i=0,\ldots, N,$$

(2.14) $\mathrm{vect}(U_i^{(*)}) = -(S_i^{-(1)} + S_i^{-(2)} \dot{+} \ldots \dot{+} S_i^{-(n)})\mathrm{vect}(N_i),$

$$i=0,\ldots,N,$$

$$(2.15)\quad \{U_i^{(*)}\ ;\ \mathrm{Im}(U_i^{(*)})\cap\mathrm{Ker}(S_i)=\{0\}\}=-\mathrm{GIN}(S_i)N_i$$

$$=-(\hat{\Lambda}_i M_i+\hat{\Delta}_i),\quad i=0,\ldots,N,$$

where

$$(2.16)\quad \Lambda_i-\hat{\Lambda}_i:=\mathrm{GIN}(S_i)E(B_i^*(V_{i+1}+I_{i+1})A_i),$$

$$(2.17)\quad \Delta_i-\hat{\Delta}_i:=\mathrm{GIN}(S_i)E(B_i^*(V_{i+1}+I_{i+1})W_i^{(i)}),$$

$$(2.18)\quad K_i\varepsilon\{M\varepsilon M(p,n');\ \mathrm{Im}(M)\subset\mathrm{Ker}(S_i)\},\quad i=0,\ldots,N.$$

Appendix D. A Note on Generalized Inverses

1. Notations

The following notations are used; $L(V, W):=L(V, W;K):=$ the set of all linear mappings from the vector space, V, to the vector space, W, over the coefficient field, K. For each $f\varepsilon L(V, W)$, the image and the kernel of f are written $\mathrm{Im}(f)$ and $\mathrm{Ker}(f)$, respectively, i.e., $\mathrm{Im}(f):=f(V)$ and $\mathrm{Ker}(f):=f^{-1}(0)$. The set of all (n, m) matrices over K is written $M(n, m; K)$ or, simply, $M(n, m)$, and if $n=m$, then put $M(n):=M(n, m)$. I_V is the identity mapping on V, and E_n is the unit matrix in $M(n)$. The direct sums of the vector spaces, $(V_k;k=1, 2, \ldots, L)$, and the matrices, $(M_k\ ;\ k=1, 2,\ldots, L)$, are written $V_1+V_2\dot{+}\ldots\dot{+}V_L$, and $M_1\dot{+}M_2\dot{+}\ldots\dot{+}M_L$, respectively. If $V=V_1+V_2$, then define $\mathrm{proj}(V_1/V_2)(x):=x_1$, where $x=x_1+x_2\varepsilon V$, $x_1\varepsilon V_1$, and $x_2\varepsilon V_2$. The restriction on the mapping, f, to the set, H, is written $f|H$, i.e., $(f|H)(x)=f(x)$, $x\varepsilon H$.

For each $f\varepsilon L(V, W)$, the family of generalized inverses of f, $\mathrm{GIN}(f)$, is defined by (1.1);

$$(1.1)\quad \mathrm{GIN}(f):=\{(f|V_1)^{-1}\circ\mathrm{proj}(\mathrm{Im}(f)/W_1)+g\circ\mathrm{proj}(W_1/\mathrm{Im}(f));$$

$$V=\mathrm{Ker}(f)\dot{+}V_1,\ W=\mathrm{Im}(f)\dot{+}W_1,\ \&\ g\varepsilon L(W_1, \mathrm{Ker}(f))\}.$$

The element of GIN(f) is called the generalized inverse of f, and written f^-.

From (1.1), (1.2)-(1.4) are easily derived;

(1.2) $GIN(f)= \{h\varepsilon L(W, V);\ f\circ h\circ f=f\}$.

(1.3) $\{ I_V - h\circ f;\ h\varepsilon GIN(f)\} = \{ proj(Ker(f)/V_1);\ V= Ker(f)\dot{+} V_1\}$.

(1.4) $\{f\circ h;\ h\varepsilon GIN(f)\}= \{proj(Im(f)/W_1);\ W= Im(f) \dot{+} W_1\}$.

For elementary properties of generalized inverses see, for example, Iri and Kan [6], Chapter 8, or Rao [10], Chapter 1.

In Subsection 2 the coefficient field is assumed to be the real or complex number field. In general, the elements in GIN(f) are written as $f^{-(1)}$, $f^{-(2)}$, $f^{-(3)}$,... etc.

2. Formulas

The following symbols, (2.1)-(2.3), are used;

(2.1) $((A, B))_C := A^* CB$,

where A^* is the conjugate transposed matrix of A, i.e., $A^* := \bar{A}^t$, and the product of matrices is assumed to be well-defined.

(2.2) $||A||_D^2 := ((A, A))_D$.

(2.3) $\{ X, Y\}_Z := (((X, Y))_Z + ((Y, X))_Z)/2$.

The formulas, (2.4)-(2.7), are easily obtained;

(2.4) $||X + Y||_Z^2 = ||X||_Z^2 + 2\{X, Y\}_Z + ||Y||_Z^2$,

where $(X, Y, Z)\varepsilon M(n, m)\times M(n, m)\times M(n)$.

(2.5) $((A, B))_{C-(1)} = ((A, B))_{C-(2)}$,

where $Im(A)\subset Im(C^*)$, $Im(B)\subset Im(C)$, and $(C^{-(1)}, C^{-(2)})\varepsilon GIN(C)^2$.

$$(2.6)\quad ||JX + A||_V^2 + ||HX + B||_W^2 =$$

$$||X + (J^*VJ + H^*WH)^{-(1)}(J^*VA + H^*WB)||^2_{J^*VJ + H^*WH}$$

$$- ||J^*VA + H^*WB||^2_{(J^*VJ + H^*WH)^{-(2)}} + ||A||_V^2 + ||B||_W^2 ,$$

where V and W are Hermite nonnegative definite matrices.

$$(2.7)\quad ||X + A||_V^2 + ||X + B||_W^2 =$$

$$||X + (V+W)^{-(1)}(VA+WB)||^2_{V+W} + ((V(A-B), W(A-B)))_{(V+W)^{-(2)}},$$

where V and W are Hermite nonnegative definite matrices.

Propositions, 2.1 and 2.2, are also easily obtained.

Proposition 2.1. Consider the M(m) valued function on M(n, m) defined by

$$I(X) := ||X||_S^2 + 2\cdot\{X, TY\}_{E_n} + ||Y||_R^2 ,$$

where $X\varepsilon M(n, m)$, $S\varepsilon M(n)$, $T\varepsilon M(n, 1)$, $Y\varepsilon M(1, m)$, $R\varepsilon M(1)$, and S is nonnegative definite and $S^*=S$. Put

$$MP := \{X_0\varepsilon M(n, m);\ I(X) - I(X_0) \text{ is nonnegative definite for any } X\varepsilon M(n, m)\}.$$

If $Im(TY)\subset Im(S)$, then (2.8)-(2.10) are derived;

$$(2.8)\quad I(X) = ||X + S^{-(1)}TY||_S^2 + ||Y||^2_{R-T^*S^{-(2)}T} .$$

$$(2.9)\quad vect(MP) = \{(S^{-(1)} \dotplus S^{-(2)} \dotplus \ldots \dotplus S^{-(m)})vect(TY);$$

$$(S^{-(i)};\ i=1, 2, \ldots, m)\varepsilon GIN(S)^m\} ,$$

where, in general, $vect(M) := (m_1^t, m_2^t, \ldots, m_p^t)^t$ for $M=(m_1, m_2, \ldots, m_p)\varepsilon M(q, p)$.

$$(2.10)\quad \{X\varepsilon MP;\ Im(X)\cap Ker(S) = \{0\}\} = \{S^-TY;\ S^-\varepsilon GIN(S)\} .$$

If $MP\neq\emptyset$, then $Im(TY)\subset Im(S)$.

Proposition 2.2. Consider the probability space, $(\Omega, \mathbb{F}, P)$, and the expectation operation with respect to P, $E(\cdot)$. Then the formulas, (2.11) and (2.12), are obtained;

$$(2.11)\quad E(||X||_Q^2) = ||E(X)||_Q^2 + E(||X-E(X)||_Q^2),$$

where X is a M(n, m) valued random variable on $(\Omega, \mathbb{F})$, and $Q \in M(n)$.

$$(2.12)\quad E(||A||_V^2) - ||E(B^*VA)||_W^2 = E(||A-BWE(B^*VA)||_V^2) + ||E(B^*VA)||_W^2,$$

where A and B are M(n, m) and M(n, 1) valued random variables on $(\Omega, \mathbb{F})$, respectively, and $V \in M(n)$ and $W \in M(1)$ are Hermite and nonnegative definite matrices.

References

[1] Aoki, M. (1967). *Optimization of Stochastic Systems*, Academic Press, New York.
[2] Bellman, R. (1961). *Adaptive Control Processes(A guided tour)*, Princeton University Press, Princeton.
[3] Berger, J.O. (1980). *Statistical Decision Theory (Foundations, concepts, and methods)*, Springer-Verlag, New York.
[4] Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Massachusetts.
[5] DeGroot, M.H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, New York.
[6] Iri, M, and Kan, T. (1977). *Linear Algebra (Matrices and their normal forms)*, Kyôiku Shuppan, Tokyo, (in Japanese).
[7] Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol.II*, (Harper and Hooker, editors), D. Reidel Publishing Company, Dordrecht-Holland, 175-257.
[8] Martin, J. J. (1967). *Bayesian Decision Problems and Markov Chains*, John Wiley & Sons, Inc., New York, reprinted by Robert E. Krieger Publishing Company, Huntington, New York, in 1975).
[9] Miyasawa, K. (1971). *An Introduction to Information and Decision Theory*, Iwanami Shoten, Tokyo, (in Japanese).
[10] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications, 2nd edition*, John Wiley & Sons, Inc., New York.
[11] Savage, L.J. (1972). *The Foundations of Statistics, 2nd revised edition*, Dover Publications, Inc., New York,

(a revised and enlarged version of the work originally published by John Wiley & Sons, Inc., New York, in 1954.)

[12] Sono, S.(1983). On a noninformative prior distribution for Bayesian inference of multinomial distribution's parameters, *Annals of the Institute of Statistical Mathematics*, Vol.35, No.2, A, 167-174.

[13] Sono, S. (1983). On an approximation for a multi-stage decision problem, *Annals of the Institute of Statistical Mathematics*, Vol.35, No.2, A, 185-191.

[14] Sono, S. (May, 1985). On Bayesian inference and Bayes decision problems, Thesis for Doctor of Science, Department of Mathematics, Faculty of Science, Science, Tokyo University.

[15] Suzuki, Y. (1978). *Statistical Analysis* , Chikuma Shobo, Tokyo, (in Japanese).

[16] Suzuki, Y. (1980). A Bayesian approach to some empirical Bayes models, *Recent Developments in Statistical Inference and Data analysis, Proceedings of the International Conference in Statistics in Tokyo,* Matsusita, K., editor, North-Holland Publishing Company, Amsterdam, 269-286.

[17] Suzuki, Y. (1981). Sequential estimation, *The Journal of Economics*, Tokyo University, 47-2, pp.11-24, (in Japanese).

(Dec.29, Sun. 1985)