An EM-Based Scheme for Record Value Statistics Models in Software Reliability Estimation

Hiroyuki Okamura and Tadashi Dohi
Graduate School of Engieering, Hiroshima University
okamu@rel.hiroshima-u.ac.jp, dohi@rel.hiroshima-u.ac.jp

Abstract

This paper considers an EM (expectation-maximization) based scheme for record value statistics (RVS) models in software reliability estimation. The RVS model provides one of the generalized modeling frameworks to unify several of existing software reliability models described as non-homogeneous Poisson processes (NHPPs). The proposed EM algorithm gives a numerically stable procedure to compute the maximum likelihood estimates of RVS models. In particular, this paper focuses on an RVS model based on a mixture of exponential distributions. As an illustrative example, we also derive a concrete EM algorithm for the well-known Musa-Okumoto logarithmic Poisson execution time model by applying our result, and discusses the effectiveness of the EM-based scheme for RVS models with a simple numerical example.

Keywords: *software reliability, non-homogeneous Poisson process, record value statistics, parameter estimation, EM algorithm*

1. Introduction

Software reliability is one of the most significant attributes for measuring software quality. The software reliability is quantitatively defined as the probability that there is no failure for a certain time period in operation. Thus, probabilistic models are applied to estimating software reliability with the field data. The software reliability community has developed a number of software reliability models (SRMs) from various points of view during the last four decades. Specifically, non-homogeneous Poisson process (NHPP) based SRMs have played a central role to estimate the number of remaining faults as well as the software reliability [1,2] due to their mathematically tractable properties.

Generally speaking, NHPP-based SRMs can be classified into finite and infinite failure models. The finite failure SRMs assume that there are a finite number of failure-causing faults in software, and that the expected total number of failures is always bounded even if a software lifetime goes to infinity. Since this property might be plausible to represent debugging activities for real software development, a large number of NHPP-based SRMs belonging to this category have been proposed and extensively discussed in the literature. On the other hand, the infinite failure SRMs are stochastic models whose expected total numbers of failures go to infinity as time elapses. The infinite failure SRMs can be regarded as limiting models derived from the finite failure models. The well-known SRM belonging to this category is Musa-Okumoto logarithmic Poisson execution time model [3], which is also known as one of the most useful NHPP-based SRMs.

To understand the modeling assumption in a unified way, there are a few generalized modeling frameworks (meta-modeling frameworks) which provide either of existing finite or

infinite NHPP-based SRMs [4,5,6]. Such a meta-modeling framework is quite important to discuss the parameter estimation and the model selection in the practical point of view. To the best of our knowledge, generalized order statistics (GOS) and record value statistics (RVS) models are two major modeling frameworks. In particular, GOS models have attractively been studied in terms of parameter estimation and model selection [7,8,9], since GOS models contain all of finite failure SRMs used in software reliability estimation. In contrast, RVS models have not fully discussed from the theoretical point of view.

This paper concerns a parameter estimation problem, and particularly proposes an EM (expectation-maximization) based scheme for RVS models in software reliability estimation. The EM algorithms for GOS models were proposed in [10, 13-16], where the authors achieved practically more stable estimation procedures than general-purpose numerical algorithms. The proposed EM algorithm here also gives a numerically stable procedure to compute the maximum likelihood estimates of RVS models. In particular, we focus on an RVS model based on a mixture of exponential distributions. As an illustrative example, we also derive a concrete EM algorithm for Musa-Okumoto NHPP SRM [3] by applying our generalized EM framework.

2. Software reliability modeling

NHPP-based SRMs are widely accepted as stochastic models to estimate software reliability. Let $\{N(t);\ t \geq 0\}$ be a stochastic process which represents the number of software failures experienced before time t. The following assumptions are made on the cumulative number of software failures:

(i) $N(0) = 0$,

(ii) $\{N(t);\ t \geq 0\}$ has independent increments,

(iii) $P(N(t + \otimes t) - N(t) = 1) = \lambda(t) \otimes t + o(\otimes t)$,

(iv) $P(N(t + \otimes t) - N(t) \geq 2) = o(\otimes t)$,

where $\lambda(t)$ is a failure intensity function and $o(\otimes t)$ is the second or higher order term of $\otimes t$. According to the above assumptions, $N(t)$ follows an NHPP having the probability mass function (p.m.f.):

$$P(N(t) = n) = \frac{\Lambda(t)^n}{n!} \exp(-\Lambda(t)), \quad t \geq 0,\ n = 0, 1, \cdots,$$

where

$$\Lambda(t) = \int_0^t \lambda(s) ds$$

is called a mean value function representing the expected cumulative number of software failures experienced before time t.

A numerous number of NHPP-based SRMs have been proposed to estimate software reliability, and characterized by mean value functions. For example, when the mean value function is given by

$$\Lambda(t) = \omega(1 - e^{-\beta t}),$$

the corresponding NHPP-based SRM is the exponential SRM by Goel and Okumoto [11].

On the other hand, there are two generalized modeling frameworks that contain almost all existing NHPP-based SRMs. The first framework is based on generalized order statistics (GOS), which is called the GOS-NHPP model in this paper. The GOS-NHPP model is made on the following assumptions [4]:

- The number of failure-causing faults is finite with a Poisson distributed random variable.
- All software failure times are mutually independent random variables with an identical probabilistic law.

Let ω and $F(t)$ be an average number of failure-causing faults and the cumulative distribution function (c.d.f.) of a software failure time, respectively. Since the number of failure-causing faults follows the Poisson distribution with mean ω, the p.m.f. of the number of failures is given by

$$P(N(t) = n) = \frac{\{\omega F(t)\}^n}{n!} e^{-\omega F(t)}, \quad n = 0, 1, \ldots. \tag{1}$$

Equation (1) is equivalent to the p.m.f of NHPP-based SRMs. Substituting well-known statistical distributions into $F(t)$ yields many of existing NHPP-based SRMs in the GOS-NHPP model.

The second framework is based on record value statistics (RVS). The RVS model is a point process consisting of record-breaking times. Let S_1, S_2, ... be IID (independently and identically distributed) random variables drawn from a p.d.f. $f(t)$ = d$F(t)$/dt. Then the sequence of record-breaking times are defined as follows.

$$\begin{aligned} &R_1 = 1, \\ &R_k = \min\{i; S_i > S_{R_{k-1}}\}, \quad \text{for } k = 2, 3, \ldots, \\ &T_k = S_{R_k}, \quad \text{for } k = 1, 2, \ldots, \end{aligned}$$

where T_k is the k-th record-breaking time. In the RVS model, T_k corresponds to the k-th software failure time, and then the p.m.f. of the number of failures experienced before t becomes an NHPP with the mean value function [12]:

$$\Lambda(t) = -\log \int_t^\infty f(s)ds = -\log \overline{F}(t),$$

where

$$\overline{F}(t) = 1 - F(t)$$

is a survival function of $f(t)$. Similar to the GOS-NHPP model, we can represent a variety of NHPPs by substituting a statistical distribution defined on the positive domain into $f(t)$. For example, when $f(t)$ is given by an exponential density function, the resulting RVS model is coincide with a homogeneous Poisson process. If $f(t)$ is a Pareto distribution of the second kind;

$$f(t) = \frac{ab^a}{(b+t)^{a+1}}, \quad t > 0, \quad a, b > 0,$$

the corresponding RVS model is equivalent to the well-known Musa-Okumoto NHPP SRM [3]. The main difference between GOS-NHPP and RVS models is the expected number of total failures as $t \to \infty$. Since the GOS-NHPP model assumes a finite number of failure-causing faults, the expected number of total failures is also bounded, i.e., $\Lambda(\infty) < \infty$. On the other hand, RVS model gives the NHPP with unbounded failures, i.e., $\Lambda(\infty) \to \infty$.

This paper focuses on a class of the RVS model whose distribution is given by a mixture of exponential (ME) distributions [13]. Concretely, the p.d.f. of an ME distribution is generally given by

$$f(t) = \int_0^\infty ue^{-ut} g(u) du,$$

where $g(u)$ is a mixture ratio distribution. In the statistical sense, the ME distribution consists of exponential p.d.f.'s with different rates \$u\$. If the ratio is a deterministic value, i.e., $g(u)$ is the delta function $\delta(u - u_0)$, the corresponding RVS model is reduced into a homogeneous Poisson process with rate u_0. Also, when $g(u)$ is a gamma density;

$$g(u) = \frac{b^a u^{a-1} e^{-bu}}{\Gamma(a)}, \quad u > 0, \quad a, b > 0,$$

the corresponding model is Musa-Okumoto NHPP SRM, where $\Gamma(\cdot)$ is the standard gamma function. This paper abbreviates the RVS model with an ME distribution as the ME-RVS model.

3. Parameter estimation

In this section, we discuss parameter estimation for the ME-RVS model. The commonly used method for parameter estimation is the maximum likelihood (ML) estimation.

Define the software failure time data as $D = (t_1, ..., t_m; t_e)$, where t_i denotes the i-th ordered failure time and t_e is the last of observation period, i.e., $N(t_e) = m$. Then the ML estimates of model parameters $\boldsymbol{\theta}_{ML}$ are determined as the values maximizing the log-likelihood function (LLF):

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}),$$
$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \sum_{i=1}^{m_e} \log \lambda(t_i; \boldsymbol{\theta}) - \Lambda(t_e; \boldsymbol{\theta}),$$

where is a parameter vector of NHPP-based SRM. To solve the above maximization problem, numerical techniques are needed since we cannot obtain closed-form solutions. However, general-purpose numerical methods like Newton-type methods are unstable to find the maximum due to their local convergence property. Therefore, in order to compute ML estimates stably, EM (expectation-maximization) algorithms are useful. In fact, the literature [10, 14-16] developed the EM algorithms for GOS-NHPP models. This paper discusses the EM algorithm for the ME-RVS model.

The EM algorithm is an iterative method for computing ML estimates with incomplete data [17, 18]. Let D and U be observable and unobservable data vectors, respectively, and we wish to estimate a model parameter vector $\boldsymbol{\theta}$ from only the observable data vector D. The problem corresponds to finding a parameter vector that maximizes a marginal LLF:

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}),$$
$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \log p(\mathcal{D}; \boldsymbol{\theta}) = \log \int p(\mathcal{D}, \mathcal{U}; \boldsymbol{\theta}) d\mathcal{U},$$

where $p(\cdot)$ is any appropriate p.d.f. or p.m.f. The EM algorithm consists of E-step and M-step. E-step computes a conditional expected LLF with respect to the complete data vector (D, U) using the posterior distribution for unobservable data vector with provisional parameter vector $\boldsymbol{\theta}$, i.e., the conditional expected LLF is given by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathrm{E}[\log p(\mathcal{D}, \mathcal{U}; \boldsymbol{\theta}) | \mathcal{D}; \boldsymbol{\theta}']$$
$$= \int p(\mathcal{U}|\mathcal{D}; \boldsymbol{\theta}') \log p(\mathcal{D}, \mathcal{U}; \boldsymbol{\theta}) d\mathcal{U}.$$

The posterior distribution for unobservable data can be obtained from Bayes theorem:

$$p(\mathcal{U}|\mathcal{D}; \boldsymbol{\theta}) = \frac{p(\mathcal{D}, \mathcal{U}; \boldsymbol{\theta})}{\int p(\mathcal{D}, \mathcal{U}; \boldsymbol{\theta}) d\mathcal{U}}.$$

In M-step, we find a new parameter vector $\boldsymbol{\theta}'$ that maximizes the expected LLF:

$$\boldsymbol{\theta}'' := \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}'),$$

and $\boldsymbol{\theta}'$ becomes a provisional parameter vector at the next E- and M-steps. The E- and M-steps are repeatedly executed until the parameters converge to ML estimates. Here it should be noted that concrete EM algorithms depend on both model and unobserved data structures.

To develop the EM algorithm for ME-RVS model, we consider a pair of IID random variates (S_i, U_i) which are jointly drawn from the exponential distribution with rate U_i and the mixture ratio distribution $g(u)$. In addition, R_i denotes an index of the i-th record-breaking value. Then the LLF for (S_i, U_i) and R_i gives a complete LLF:

$$\log p(\mathcal{D},\mathcal{U};\boldsymbol{\theta}) = \sum_{i=1}^{R_{m+1}} (\log U_i - U_i S_i) + \sum_{i=1}^{R_{m+1}} \log g(U_i;\boldsymbol{\theta}).$$

Then the E-step is represented as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathrm{E}\left[\sum_{i=1}^{R_{m+1}} \log g(U_i;\boldsymbol{\theta}) \middle| \mathcal{D};\boldsymbol{\theta}' \right].$$

Here we can derive a useful formula to compute the above Q-function: For any function $h(\cdot)$, the expected value can be computed as follows.

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{R_{m+1}} h(U_i) \middle| \mathcal{D};\boldsymbol{\theta} \right] &= \frac{\int_0^\infty h(u) u e^{-ut_1} g(u;\boldsymbol{\theta}) du}{\int_0^\infty u e^{-ut_1} g(u;\boldsymbol{\theta}) du} \\
&+ \sum_{k=2}^{m} \left(\frac{\int_0^\infty h(u) u e^{-ut_k} g(u;\boldsymbol{\theta}) du}{\int_0^\infty u e^{-ut_k} g(u;\boldsymbol{\theta}) du} + \frac{\int_0^\infty h(u)(1-e^{-ut_{k-1}}) g(u;\boldsymbol{\theta}) du}{\int_0^\infty e^{-ut_{k-1}} g(u;\boldsymbol{\theta}) du} \right) \\
&+ \frac{\int_0^\infty h(u) e^{-ut_e} g(u;\boldsymbol{\theta}) du}{\int_0^\infty e^{-ut_e} g(u;\boldsymbol{\theta}) du} + \frac{\int_0^\infty h(u)(1-e^{-ut_m}) g(u;\boldsymbol{\theta}) du}{\int_0^\infty e^{-ut_m} g(u;\boldsymbol{\theta}) du}.
\end{aligned}$$

On the other hand, since we find the parameter maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta})$, the M-step procedure can be derived from a usual ML procedure for $g(u)$. For example, $g(u)$ belongs to the exponential family, so that the ML estimates in the M-step are given in closed forms.

4. Illustration of EM procedure for Musa-Okumoto SRM

We present concrete EM-step formulas for Musa-Okumoto NHPP SRM in this section. Suppose that the mixture ratio distribution is given by a gamma distribution with parameters *a* and *b*. Then the corresponding ME-RVS model becomes Musa-Okumoto NHPP SRM. The complete LLF for IID rate samples $\Lambda_1, ..., \Lambda_{Rm+1}$ can be written in the form:

$$\log p(\mathcal{D}, \mathcal{U}; (a,b)) = R_{m+1} a \log b + (a-1) \sum_{i=1}^{R_{m+1}} \log \Lambda_i - b \sum_{i=1}^{R_{m+1}} \Lambda_i - R_{m_e+1} \log \Gamma(a).$$

Based on the EM formulas in the previous section, we have

$$\mu_1(a,b) := \mathrm{E}[R_{m+1} | \mathcal{D}; (a,b)] = 1 + \sum_{k=1}^{m} \left(\frac{b}{b t_k} \right)^{-a},$$

$$\mu_\lambda(a,b) := \mathrm{E}\left[\sum_{i=1}^{R_{m+1}} \Lambda_i \middle| \mathcal{D}; (a,b) \right] = \sum_{k=1}^{m} \left(\frac{1}{b+t_k} + \frac{a}{b} \left(\frac{b}{b+t_k} \right)^{-a} \right) + \frac{a}{b+t_e},$$

$$\mu_{\log\lambda}(a,b) := \mathrm{E}\left[\sum_{i=1}^{R_{m+1}} \log \Lambda_i \middle| \mathcal{D}; (a,b) \right] = \sum_{k=1}^{m} \left(\frac{1}{a} + (\psi(a) - \log b) \left(\frac{b}{b+t_k} \right)^{-a} \right) + \psi(a) - \log(b+t_e),$$

where $\psi(\cdot)$ is the digamma function $\psi(a) = d\log\Gamma(a)/da$. Using $\mu_1(a, b)$, $\mu_{\lambda(a,b)}$ and $\mu_{\log\lambda(a,b)}$, the M-step for Musa-Okumoto NHPP SRM is given by

$$a = \inf \left\{ a > 0; \ \log a - \psi(a) = \log \frac{\mu_\lambda(a',b')}{\mu_1(a',b')} - \frac{\mu_{\log\lambda}(a',b')}{\mu_1(a',b')} \right\},$$

$$b = \frac{a \mu_1(a',b')}{\mu_\lambda(a',b')}.$$

5. Discussions and future research

The proposed EM algorithm provides more stable procedures to compute ML estimates than general-purpose numerical methods, since the EM algorithm has a global convergence property. However, it should be noted that the convergence is slower than other methods like Newton-type methods. In general, the convergence rate of EM algorithm is known as a liner

function of the number of iterations, whereas the convergence rate of Newton-type methods is a quadratic function of the number of iteration. In addition, the convergence speed of EM algorithm depends on the amount of unobservable information. In our EM framework for RVS models, we define unobservable data as the samples that do not break the current record value. Indeed, there are a huge number of such unobservable samples in our framework, except for the case where the total number of failures is small. Thus the proposed EM procedure is expected to be slow for the convergence when the number of failures becomes large, so that the EM algorithm for RVS models may not be applied to estimate software reliability from observed data. Figure 1 depicts an illustrative example of the convergence behavior of log-likelihood function under the EM algorithm. In this case, we apply the proposed EM algorithm for Musa-Okumoto NHPP SRM. From this result, we see that the log-likelihood does not fully converge to the maximum value even when the numer of iteration is large. On the other hand, we also developed EM algorithms for GOS-NHPP models in [10, 14-16] which gave a practically rapid convergence speed unlike the RVS case. Since several acceleration methods for EM algorithms have been proposed, e.g., [19], we will try to apply such acceleration techniques to improve the convergence speed of EM algorithm for RVS models in the future.e present concrete EM-step formulas for Musa-Okumoto NHPP SRM in this section.

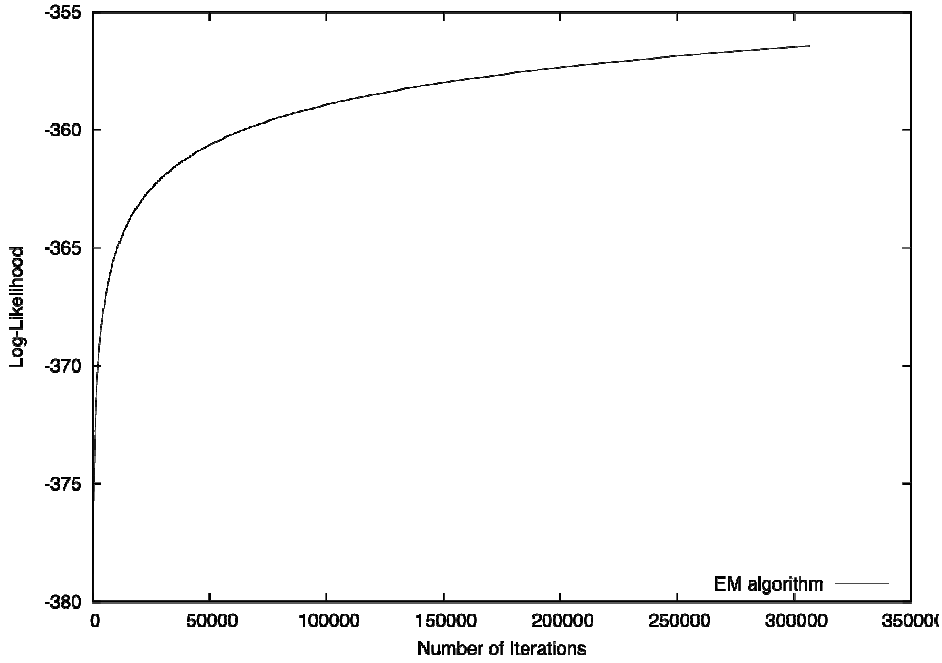

Figure 1. Convergence behavior of log-likelihood function for Musa-Okumoto NHPP SRM.

6. References

[1] Musa, J.D., Iannino, A., Okumoto, K.: Software Reliability, Measurement, Prediction, Application. McGraw-Hill, New York (1987)

[2] Lyu, M.R. (ed): Handbook of Software Reliability Engineering. McGraw-Hill, New York (1996)

[3] Musa J.D., Okumoto, K.: A logarithmic {P}oisson execution time model for software reliability measurement. In: 7th International Conference Software Engineering, pp. 230-238. IEEE CS Press/ACM (1984)

[4] Langberg N., Singpurwalla, N.D.: Unification of some software reliability models. SIAM Journal on Scientific Computing. 6, 781-790 (1985)

[5] Dohi, T., Osaki, S., Trivedi, K.S.: An infinite server queueing approach for describing software reliability growth -- unified modeling and estimation framework. In: 11th Asia-Pacific Software Engineering Conference, pp. 110-119, IEEE CS Press (2004)

[6] Grottke M., Trivedi, K.S.: On a method for mending time to failure distributions. In: International Conference on Dependable Systems and Networks, pp. 560-569. IEEE CS Press (2005)

[7] Miller, D.R.: Exponential order statistic models of software reliability growth. IEEE Transactions on Software Engineering. SE-12, 12-24 (1986)

[8] Raftery, A.E.: Inference and prediction for a general order statistic model with unknown population size. Journal of the American Statistical Association. 82, 1163-1168 (1987)

[9] Wilson, S.P., Samaniego, F.J.: Nonparametric analysis of the order-statistic model in software reliability. IEEE Transactions on Software Engineering. 33, 198-208 (2007)

[10] Okamura, H., Watanabe, Y., Dohi, T.: An iterative scheme for maximum likelihood estimation in software reliability modeling. In: 14th International Symposium on Software Reliability Engineering, pp. 246-256, IEEE CS Press (2003)

[11] Goel, A.L., Okumoto, K.: Time-dependent error-detection rate model for software reliability and other performance measures. IEEE Transactions on Reliability. R-28, 206-211 (1979)

[12] Kuo, L., Yang, T.Y.: Bayesian computation for nonhomogeneous Poisson processes in software reliability. Journal of the American Statistical Association. 91, 763-773 (1996)

[13] Okamura, H, Watanabe, Y., Dohi, T.: Estimating mixed software reliability models based on the EM algorithms. In 2002 International Symposium on Empirical Software Engineering, pp. 69-78, IEEE CS Press (2002)

[14] Okamura, H, Murayama, A., Dohi, T.: EM algorithm for discrete software reliability models: a unified parameter estimation method. In: 8th IEEE International Symposium on High Assurance Systems Engineering, pp. 219-228, IEEE CS Press (2004)

[15] Okamura, H., Dohi, T.: Building phase-type software reliability models. In: 17th International Symposium on Software Reliability Engineering, pp. 289-298, IEEE CS Press (2006).

[16] Okamura H., Dohi, T.: Hyper-Erlang software reliability model. In: 14th Pacific Rim International Symposium on Dependable Computing, pp. 232-239, IEEE CS Press (2008)

[17] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. B-39, 1-38 (1977)

[18] Wu, C.F.J.: On the convergence properties of the {EM} algorithm. Annals of Statistics. 11, 95--103 (1983)

[19] McLachlan, G.J., Krishnan, T.: EM Algorithm and Extensions. John Wiley & Sons (1997)

Authors

Hiroyuki Okamura received the B.S.E., M.S. and Dr. of Engineering degrees from Hiroshima University, Japan, in 1995, 1997 and 2001, respectively. In 1998 he joined the Hiroshima University as an Assistant Professor, and is now working as an Associate Professor in the Department of Information Engineering, Graduate School of Engineering from 2003. His research areas include Performance Evaluation, Dependable Computing and Applied Statistics. He is a regular member of ORSJ, IPSJ and IEEE.

Tadashi Dohi received the B.Sc. (Engineering), M.Sc. (Engineering), and Ph.D. (Engineering) from Hiroshima University, Japan, in 1989, 1991, and 1995, respectively. In 1992, he joined the Department of Industrial and Systems Engineering, Hiroshima University, Japan, as an Assistant Professor. Now he is a Full Professor in the Department of Information Engineering, Faculty of Engineering, Hiroshima University, Japan, since 2002. In 1992, and 2000, he was a Visiting Research Scholar at University of British Columbia, Canada, and Duke University, USA, respectively, on leave of absence from Hiroshima University. His research areas include Software Reliability Engineering, Dependable Computing, and Performance Evaluation. Dr. Dohi has published over three hundred journal and peer-reviewed conference papers in the above research topics. He is a Regular Member of ORSJ, JSIAM, IEICE, REAJ, and IEEE. He also served as General Co-chair of AIWARM'04-08, WoSAR'08-10, MENS'10, APARM'10 and General Chair of ISSRE'11, Tokyo, Japan.