Simulation-based parameter estimation for complex models: a breast cancer natural history modelling illustration

Yen Lin Chia, Peter Salzman, Sylvia K Plevritis and Peter W Glynn

Stat Methods Med Res 2004 13: 507
DOI: 10.1191/0962280204sm380ra

Statistical Methods in Medical Research 2004; **13**: 507–524

Simulation-based parameter estimation for complex models: a breast cancer natural history modelling illustration

Yen Lin Chia, Peter Salzman Management Science and Engineering, Terman Engineering Center, Stanford University, Stanford, CA, USA, **Sylvia K Plevritis** Department of Radiology, Lucas Center for MR Spectroscopy and Imaging, Stanford, CA, USA and **Peter W Glynn** Management Science and Engineering, Terman Engineering Center, Stanford University, Stanford, CA, USA

Simulation-based parameter estimation offers a powerful means of estimating parameters in complex stochastic models. We illustrate the application of these ideas in the setting of a natural history model for breast cancer. Our model assumes that the tumor growth process follows a geometric Brownian motion; parameters are estimated from the SEER registry. Our discussion focuses on the use of simulation for computing the maximum likelihood estimator for this class of models. The analysis shows that simulation provides a straightforward means of computing such estimators for models of substantial complexity.

1 Introduction

Most mathematical models involve unknown parameters that must be estimated from observed data. As new models are created, estimation methods appropriate to those models must be developed. Among the general approaches available for construction of such estimators are the method of moments and the method of maximum likelihood. Method of moments requires the ability to compute population moments of the observables associated with the postulated model, whereas method of maximum likelihood requires the ability to compute the likelihood of the observed data under the given model. When the model is complex, such computations can prove challenging.

A powerful means of computing such model-based population moments or likelihoods is to use Monte Carlo simulation. For example, in the setting of the method of moments, one essentially performs Monte Carlo simulation at various points in the statistical parameter space and computes the population moments at these parameter points via computer-based Monte Carlo sampling. The point in the parameter space giving the best fit to the observed sample moments is then declared to be the moment based estimator. This approach has proved very effective in computing parameter estimators for complex stochastic models in a variety of applications areas; an illustration in the econometrics setting is presented in Duffie and Singleton.[1]

Address for correspondence: Peter W Glynn, Management Science and Engineering, Terman Engineering Center, Stanford University, Stanford, CA 94305-4026, USA. E-mail: glynn@stanford.edu

 10.1191/0962280204sm380ra

This paper illustrates the use of simulation-based estimation methods in the context of a natural history model for breast cancer. In Section 2, it is argued that this model is of interest in its own right. It is perhaps the simplest possible model of tumor growth based on simulation in which the growth process itself is modelled as a stochastic process with random variation occurring over the lifetime of the affected woman. Specifically, growth of the tumor is modelled as a geometric Brownian motion. This process is a stochastic version of deterministic exponential growth. We fit the model from data collected through the Surveillance, Epidemiology, and End Results (SEER) registry of the National Cancer Institute.

We describe three simulation-based approaches to maximum likelihood estimation for this model, starting from an easily implemented likelihood computation on a grid of points and progressing to more sophisticated iterative algorithms. Specifically, Section 6 describes an iterative algorithm known as the Kiefer–Wolfowitz (KW) algorithm that is based on simulating the likelihood itself, whereas the Robbins–Monro (RM) algorithm of Section 7 is a more sophisticated iterative variant based on simulating the gradient of the likelihood. In Section 8, it is found that via use of stochastic calculus, the likelihood can be computed in closed form in terms of modified Bessel functions of the first and second kind. Given the complexity of the likelihood and the mathematical background required to compute it, we view our geometric Brownian motion model as being right on the boundary of the models that are amenable to closed-form analysis.

The paper therefore provides an illustration of the various levels at which the estimation problem can be attacked, depending on the background and mathematical sophistication of the modeler. In Section 9, we provide a numerical comparison of the iterative methods of Sections 6 and 7, and study their convergence characteristics relative to the 'gold-standard' of the closed-form maximum likelihood estimator of Section 8. Concluding remarks are offered in Section 10.

2 Tumor growth model

A commonly proposed model for tumor growth assumes that the rate of growth is proportional to the number of malignant cells.[2] This is essentially equivalent to assuming that the volume $v(t)$ of the tumor at time t satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}v(t) = rv(t) \tag{1}$$

subject to the initial condition $v(0) = v_0$. Here, r is a positive (deterministic) constant, v_0 is the volume of a single malignant cell, and $t = 0$ is the time at which tumor growth is initiated.

Of course, there is individual variation in tumor growth across the population. If one assumes that each individual's tumor growth is controlled by their own individually determined growth rate (possibly governed by that individual's genetic inheritance), then the volume $v(t)$ of the tumor at time t satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}v(t) = Rv(t) \tag{2}$$

subject to $v(0) = v_0$, where R is an individually determined growth rate. To model heterogeneity across the population, one can assume that R is selected (randomly) from an appropriately chosen probability distribution. Note that once R is selected, the entire volume trajectory $(v(t): t \geq 0)$ is (conditionally on R) deterministic.

A very different approach to capturing individual variation across the population is to make the tumor growth process itself vary randomly across the population (as a stochastic process). Rather than assuming that the tumor grows predictably as a function of time [as in Equations (1) and (2)], we assume that for any $h > 0$,

$$\left(\frac{V((i+1)h)}{V(ih)} : i \geq 0\right)$$

is a sequence of independent and identically distributed (iid) random variables. This means that the proportion change in the tumor volume for a time interval of duration h is independent of that observed in other periods and follows a common probability distribution. If the volume process $V = (V(t): t \geq 0)$ is further assumed to be continuous, then it can be shown that V must necessarily take the form

$$V(t) = V(0) \exp(\mu t + \sigma B(t)) \tag{3}$$

for some (deterministic) constants μ and σ, where $B = (B(t): t \geq 0)$ is a standard Brownian motion process. A standard Brownian motion has stationary and independent increments, continuous paths and satisfies $B(t) \stackrel{D}{=} N(0, t)$ (where $\stackrel{D}{=}$ denotes equality in distribution and $N(\mu, \sigma^2)$ is a normally distributed random variable with mean μ and variance σ^2). This characterization of V relies on the fact that $(\log V(t): t \geq 0)$ is, under our assumptions, a continuous process with stationary increments, and hence must be a Brownian motion with (possibly nonzero) drift; detailed in p. 29 of Øksendal.[3] The process V described by Equation (3) is called a 'geometric Brownian motion' with parameters μ and σ^2.

If one assumes that μ and σ^2 are common parameters across the entire population, then statistical characteristics of the growth are identical across the population at the time of tumor initiation. The observed heterogeneity across the population can then be attributed to individual stochastic variation encountered during the development of the tumor. Roughly speaking, the model described by Equation (1) suggests that population heterogeneity has a purely genetic origin, whereas the model (3) suggests that such heterogeneity has an environmental explanation.

3 Intensity-based clinical detection model

Suppose that a woman develops a breast cancer tumor that is governed by the geometric Brownian model (3). We now describe an intensity-based model that

describes the time at which the tumor will be clinically detected. Specifically, if T is the instant at which clinical detection occurs, we assume

$$P(T \in [t, t+h) | T > t, V) = \gamma V(t)h + o(h) \tag{4}$$

as $h \downarrow 0$, where $o(h)$ denotes a function $f(h)$ for which $f(h)/h \to 0$ as $h \downarrow 0$ and γ denotes a positive (deterministic) constant. The model (4) asserts that the intensity of clinical detection is proportional to the tumor volume. Such volume based intensity models have been frequently suggested.[4]

Given Equations (3) and (4), we can (in principle) compute the probability distribution of the tumor volume $V(T)$ at the time of clinical detection, as a function of the model parameters μ, σ^2 and γ. Thus, if we are given a data set consisting of such tumor volumes, we can potentially estimate the parameter values of μ, σ^2 and γ that best fit the data. Note, however, that the distribution of tumor volume at detection depends on the assumed stochastic structure of the tumor growth process and its rate relative to that at which detection occurs. Thus, whereas the relative rates play a key role in determining the distribution of tumor volume, the absolute rates do not. Consequently, the probability distribution of $V(T)$ depends only on the relative magnitudes of μ, σ^2 and γ. In particular, we may take $\sigma^2 = 1$ if we so choose.

Proposition 1. *Let $P_{\mu,\sigma^2,\gamma}(\cdot)$ be the probability distribution under which V and T evolve according to the parameters μ, σ^2 and γ. Then,*

$$P_{\mu,\sigma^2,\gamma}(V(T) \in \cdot) = P_{\mu/\sigma^2,1,\gamma/\sigma^2}(V(T) \in \cdot)$$

Proposition 1 is proved in Appendix A.

Henceforth, we can and will assume that the unit of time has been chosen so that $\sigma^2 = 1$. To estimate μ and γ observed values of $V(T)$ are required. Such a data set is available through the SEER Program of the National Cancer Institute. Note that breast cancer screening was introduced in 1982. However, the SEER database does not record whether a woman diagnosed with breast cancer subsequent to 1982 was screen detected or clinically detected. As we are modelling only clinical detection, we therefore consider only pre-1982 data. Furthermore, we shall limit our study to those patients aged 40–80 years old (at the time of detection) who suffered from invasive breast cancer associated with only one primary tumor. The age restriction is enforced because it is widely believed that the cancers exhibited by very young and very old women are substantially different from those associated with the 40- to 80-year old cohort (e.g., the cancers associated with younger women are largely associated with the BrCa gene and appear to grow faster than those in the general population[5,6]). This leads to a data set of 35 504 women, for whom clinically detected tumor volumes are known.

For pre-1982 SEER data, the size (i.e., diameter) of the tumor is recorded by SEER as falling into one of the eight categories: less than 0.5, 0.5–0.9, 1.0–1.9, 2.0–2.9, 3.0–3.9, 4.0–4.9, 5.0–9.9 cm and tumors of 10.0 cm or more. Because of the small bin counts associated with the small and large size categories (and lack of reliability in tumor size

measurements at the extremes of the range), we collapse the first two categories and last two categories into single bins. The resulting SEER data set can be found in Table 1.

Recall that our model assumes observability of the tumor volume $V(T)$ at clinical detection. To convert size data into volume data, we assume that the tumor is spherical, so that the volume $V(T)$ is given by $(\pi/6)S(T)^3$, where $S(T)$ is the diameter (or size) at detection.

4 Parameter estimation via the method of maximum likelihood

The method of maximum likelihood is known to exhibit many favorable statistical properties, including statistical efficiency.[7] Because of the binned nature of the observed data, the likelihood takes the form

$$L(\mu, \gamma) = \prod_{i=1}^{6} P_\theta(V(T) \in [a_{i-1}, a_i))^{N_i} \tag{5}$$

where $\theta \triangleq (\mu, \gamma)$, $[a_{i-1}, a_i)$ is the ith volume category, and N_i is the observed bin count for $[a_{i-1}, a_i)$ (specific values are presented in Table 1).

The method of maximum likelihood suggests estimating the true value of θ underlying the population, call it $\theta^* \triangleq (\mu^*, \gamma^*)$, via the maximizer $\hat{\theta} \triangleq (\hat{\mu}, \hat{\gamma})$ that maximizes the likelihood L. The difficulty in applying maximum likelihood to this model is that the probabilities $P_\theta(V(T) \in [a_{i-1}, a_i))$ are challenging to compute. One powerful means of computing such probabilities for models of almost arbitrary complexity is to employ Monte Carlo simulation.

Perhaps the most straightforward means of applying simulation in the current setting is to select an appropriately dense grid of points $\theta_1, \theta_2, \ldots, \theta_m$ from the parameter space $\Lambda = \{(\mu, \gamma): \mu \geq 0, \gamma > 0\}$. At each selected point $\theta_i = (\mu_i, \gamma_i)$, one can then run n independent simulations of the process V up to time T, thereby yielding $V_1(T_1), \ldots, V_n(T_n)$. The likelihood can then be estimated via

$$\hat{L}_1(\theta_i) = \prod_{l=1}^{6} \hat{P}_{\theta_i, n}(V(T) \in [a_{l-1}, a_l))^{N_l} \tag{6}$$

Table 1 Distribution of breast cancer tumor sizes in SEER database

Size, s (cm)	Tumor size count
$s < 1$	2587
$1 \leq s < 2$	8189
$2 \leq s < 3$	9287
$3 \leq s < 4$	7203
$4 \leq s < 5$	4439
$s \geq 5$	3790

where

$$\hat{P}_{\theta_i,n}(V(T)) \in [a_{l-1}, a_l)) = \frac{1}{n}\sum_{j=1}^{n} I(V_j(T_j) \in [a_{l-1}, a_l))$$

The maximum likelihood estimator $\hat{\theta}_1$ is then given by

$$\hat{\theta}_1 = \arg\max_{\theta_i} \hat{L}_1(\theta_i) \tag{7}$$

Note that the simulated optimizer $\hat{\theta}_1$, in general, differs from the maximizer $\hat{\theta}$. However, as the sample size n and number of grid points m converge to infinity, one can establish $\hat{\theta}_1$ converges to $\hat{\theta}$. Some guidelines on how to choose m and n can be found in Ensor and Glynn.[8]

5 Simulation of tumor volume at detection

In Section 4, we discussed one simple means of employing simulation as a computational device for calculating the maximum likelihood estimators for μ and γ. Of course, any simulation-based algorithm requires the ability to efficiently generate the required random variates. In our setting, we need an algorithm for simulating the random variates $V(T)$.

The process V can easily be generated at multiples of the time increment h via the recursion

$$V((i+1)h) = V(ih)\exp(\mu h + \sqrt{h}N_{i+1}(0,1))$$

subject to $V(0) = v_0$, where $(N_i(0,1): i \geq 1)$ is a sequence of iid $N(0,1)$ random variables.

To simulate the detection time T, several different alternatives exist. The most straightforward approach to generating T is to simulate an exponential random variate W with unit mean, independent of V, and select T so that

$$\gamma \int_0^T V(s)\,\mathrm{d}s = W \tag{8}$$

It can easily be checked that T has the appropriate distribution. Unfortunately, we cannot simulate the exact value of the integral appearing in Equation (8) but only a discrete approximation to it, based on $(V(ih): i \geq 0)$. In particular, let T_h be defined as

$$T_h = h \min\left\{ n \geq 1: h\sum_{i=1}^{n} V(ih) \geq \gamma^{-1} W \right\}$$

If h is chosen small enough, T_h will be close to T. We assume throughout the remainder of this paper that the discretization error $|T - T_h|$ is small enough so as to be negligible.

The key to finding the maximizer of $L(\theta)$ is to accurately compute differences of the form $L(\theta') - L(\theta)$; if the difference is positive, the maximizer is likely to be in the direction of θ'. Thus, accurate computation of differences in the likelihood surface is a key to the successful computation of the maximizer $\hat{\theta}$ of $L(\cdot)$.

One means of efficiently computing differences in the Monte Carlo setting is to employ the method of common random numbers. Note that we can avoid using independent streams of random numbers in performing our simulations at the parameter points $\theta' = (\mu', \gamma')$ and $\theta = (\mu, \gamma)$. The idea is to use a common stream of random numbers to drive both sets of simulations, thereby (hopefully) inducing positive correlation. Positive correlation can significantly reduce variance (relative to independent simulations at θ and θ') in the context of computing differences.[9]

In the setting of our geometric Brownian motion model, there is a particularly natural means of implementing common random numbers. In particular, note that process $V_\mu = \{V_\mu(t): t \geq 0\}$ defined by

$$V_\mu(t) = \exp(\mu t + B(t))$$

has precisely the distribution of V associated with the parameter point $\theta = (\mu, \gamma)$. Thus, if $T(\mu, \gamma)$ satisfies

$$\int_0^{T(\mu,\gamma)} \exp(\mu s + B(s))\, ds = \frac{W}{\gamma}$$

the random variable $\exp(\mu T(\mu, \gamma) + B(T(\mu, \gamma)))$ will have the distribution of $V(T)$ associated with the parameter point $\theta = (\mu, \gamma)$. Hence, we can estimate $L(\theta)$ via the Monte Carlo method using

$$\hat{L}_2(\theta) = \prod_{i=1}^{6} \left(\frac{1}{n} \sum_{j=1}^{n} I(\exp(\mu T_j(\mu, \gamma) + B_j(T_j(\mu, \gamma))) \in [a_{i-1}, a_i)) \right)^{N_i}$$

where $(B_1, T_1(\mu, \gamma)), \ldots, (B_n, T_n(\mu, \gamma))$ are n independently simulated replications of $(B, T(\mu, \gamma))$. The simulated likelihood surface $\hat{L}_2(\cdot)$ is then a good approximation to the empirical likelihood $L(\cdot)$ when n is large. Note that the standard Brownian motion B and exponential variate W need only be generated n times in order to compute the entire approximating surface $\hat{L}_2(\cdot)$ [as opposed to the mn simulations that would be necessary if n independent simulations were performed independently at each of m parameter points $(\theta_1, \ldots, \theta_m)$]. Thus, the method of common random numbers improves on use of independent streams both by inducing positive correlation and by reducing the overall computer time necessary to simulate all the random variates.

Of course, the earlier discussion implicitly presumes that $\mu T(\mu, \gamma) + B(T(\mu, \gamma))$ can be exactly simulated. As noted earlier in Section 4, only a (very) close approximation to $\mu T(\mu, \gamma) + B(T(\mu, \gamma))$ can be generated (by setting h small enough so as to make the error negligible).

To optimize the simulated surface, one simple approach would involve evaluation of $\hat{L}_2(\cdot)$ at the grid points $\theta_1, \ldots, \theta_m$. A simulation-based estimator is then given by

$$\hat{\theta}_2 = \arg\min_{\theta_i} \hat{L}_2(\theta_i)$$

Note that $\hat{\theta}_2$ differs from $\hat{\theta}_1$ in its use of common random numbers rather than independent streams of random variables. A more sophisticated alternative is to optimize the simulated surface via an iterative numerical algorithm. Because $\hat{L}_2(\cdot)$ is not smooth in θ (due to the presence of the indicator variables), a nonsmooth iterative optimization scheme (such as the Nelder–Mead algorithm[10]) must be used. Use of such an iterative procedure permits the global maximizer

$$\hat{\theta}_3 = \arg\max_{\theta \in \Lambda} \hat{L}_2(\theta)$$

to be computed.

6 Kiefer–Wolfowitz algorithm

One problem with the simulation-based estimators $\hat{\theta}_2$ and $\hat{\theta}_3$ is that the surface $\hat{L}_2(\cdot)$ must be reoptimized every time additional simulations are added to the sample. In particular, recomputing $\hat{\theta}_3$ can involve significant effort, even if the previously computed estimator is used as a starting point for the iterative reoptimization. This creates difficulties if one needs to increase the simulated sample size n in order to reduce the Monte Carlo error to an acceptable level.

The KW algorithm offers a fundamentally different means of computing the maximizer $\hat{\theta}$ of $L(\cdot)$ via simulation. Note that we expect the likelihood function $L(\theta)$ to be 'smooth' in θ, in the sense that we expect it to be at least twice continuously differentiable (in fact, we establish this smoothness in Section 8). In the presence of such smoothness, $\hat{\theta}$ will be a root of

$$\nabla L(\hat{\theta}) = 0 \tag{9}$$

where $\nabla L(\theta)$ represents the gradient of $L(\cdot)$ evaluated at θ. The idea is now to develop a simulation-based algorithm for iteratively computing the solution $\hat{\theta}$ to Equation (9).

Suppose that at each $\theta \in \Lambda$, we are able to simulate an unbiased estimator of the gradient of L. In particular, we demand that

$$E[Z(\theta)] = \nabla L(\theta) \tag{10}$$

Starting at an initial guess θ_0, we can then define $\theta_1, \theta_2, \ldots$, via the iteration

$$\theta_{n+1} = \theta_n - a_{n+1} Z_{n+1}(\theta_n) \tag{11}$$

where $Z_{n+1}(\theta_n)$ is independently generated at time $n+1$ so that it has the distribution of the random vector $Z(\theta)$ evaluated at $\theta = \theta_n$. The sequence $\{a_n : n \geq 1\}$ appearing in Equation (11) must be chosen so that $a_n \geq 0$,

$$\sum_{n=1}^{\infty} a_n = \infty$$

$$\sum_{n=1}^{\infty} a_n^2 < \infty$$

The standard choice for $(a_n : n \geq 1)$ is $a_n = a/n$ for some value of $a > 0$. With this choice of $(a_n : n \geq 1)$, general conditions can be established under which $\theta_n \to \hat{\theta}$ with probability one as $n \to \infty$. Furthermore, $n^{1/2}(\theta_n - \hat{\theta})$ has an approximate (multivariate) normal distribution for n large; relevant theory is presented in Pflug.[11]

The KW algorithm is the special case in which $Z(\theta)$ is defined via a finite-difference approximation to the gradient. Specifically, if the finite-difference increments δ_1 and δ_2 are chosen small enough, then

$$Z_1(\mu, \gamma) \triangleq \frac{\hat{L}_2(\mu + \delta_1, \gamma) - \hat{L}_2(\mu - \delta_1, \gamma)}{2\delta_1}$$

$$Z_2(\mu, \gamma) \triangleq \frac{\hat{L}_2(\mu, \gamma + \delta_2) - \hat{L}_2(\mu, \gamma - \delta_2)}{2\delta_2}$$

produces a pair $\tilde{Z}(\mu, \gamma) \triangleq (Z_1(\mu, \gamma), Z_2(\mu, \gamma))$ such that

$$E[\tilde{Z}(\mu, \gamma)] \approx \nabla L(\mu, \gamma)$$

As discussed in Section 5, the n simulations at each of the four parameters points $(\mu + \delta_1, \gamma)$, $(\mu - \delta_1, \gamma)$, $(\mu, \gamma + \delta_2)$ and $(\mu, \gamma - \delta_2)$ can be done either independently (in which case a total of $4n$ simulations of B and W per iteration are needed) or via the common random numbers idea described there (leading to n simulations of B and W per iteration). Because the finite-difference approximation is defined in terms of a difference in the likelihood, use of common random numbers is strongly recommended in this KW setting.

This finite-difference approximation to the gradient induces a bias in the estimator $\tilde{Z}(\mu, \gamma)$ [as the estimator of $\nabla L(\mu, \gamma)$]. This can be controlled by choosing δ_1 and δ_2

sufficiently small. In the current setting, an additional (and more subtle) source of bias is also present. In particular, the estimator $\hat{L}_2(\theta)$ is biased as an estimator of $L(\theta)$, in the sense that

$$E[\hat{L}_2(\theta)] \neq L(\theta)$$

The bias $E[\hat{L}_2(\theta) - L(\theta)]$ can be reduced by making the number n of simulated replicates per iteration large.

Because of the persistent bias due to the presence of the finite differences, the KW algorithm does not enjoy the $n^{-1/2}$ convergence rate described earlier. Nevertheless, the associated simulation-based estimator $\hat{\theta}_4$ for $\hat{\theta}$ can still yield practically useful estimates of $\hat{\theta}$. (The asymptotic theory for the KW algorithm shows that the estimator's asymptotic behaviour can be enhanced by letting the difference increments δ_1 and δ_2 shrink to zero as the iteration count n tends to infinity.[12])

We implemented $\hat{\theta}_4$ for our model and SEER data set. The results are reported in Section 9.

7 Robbins–Monro algorithm

In Section 6, finite-difference approximations to the gradient are used to drive the iterative algorithm (11). Naturally, we expect improved algorithmic convergence when the gradient can be estimated directly (without resort to finite differences). Such algorithms require the ability to simulate an unbiased estimator $Z(\theta)$ of the gradient, so that $E[Z(\theta)] = \nabla L(\theta)$. In contrast, the algorithms of Section 6 require only the ability to simulate unbiased (or almost unbiased) estimators of $L(\theta)$.

If $\tilde{L}(\theta)$ is an unbiased (or almost unbiased) estimator of the likelihood $L(\theta)$, one might hope to put $Z(\theta) = \nabla\tilde{L}(\theta)$, in the belief that

$$E[Z(\theta)] = E[\nabla\tilde{L}(\theta)] = \nabla E[\tilde{L}(\theta)] = \nabla L(\theta)$$

Such an approach presumes that the gradient operator can universally be interchanged with the expectation. Though such an interchange is frequently valid, it is not universally so. In particular, in our setting, we noted earlier that $\hat{L}_2(\theta)$ is not smooth in θ due to the presence of indicator random variables in its definition. Hence, $\nabla\hat{L}_2(\theta)$ is not even well defined in our setting.

Observe that

$$\nabla L(\theta) = \sum_{i=1}^{6} N_i P_\theta(V(T) \in [a_{i-1}, a_i))^{N_i - 1} \nabla P_\theta(V(T) \in [a_{i-1}, a_i)) \prod_{j \neq i} P_\theta(V(T) \in [a_{j-1}, a_j))^{N_j}$$

As $P_\theta(V(T) \in [a_{i-1}, a_i))$ can be estimated easily via the proportion of simulations on which $V(T) \in [a_{i-1}, a_i)$, our focus is on $\nabla P_\theta(V(T) \in [a_{i-1}, a_i))$.

If $W(\gamma)$ is an exponential random variable with mean $1/\gamma$ and g is any nonnegative function,

$$
\begin{aligned}
&E[g(\mu h + \sqrt{h}N_1(0,1), \ldots, \mu h + \sqrt{h}N_n(0,1), W(\gamma))] \\
&\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_0^{\infty} g(x_1, \ldots, x_n, y)\gamma e^{-\gamma y}\, dy \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{(x_i - \mu h)^2}{2h}\right) dx_i \\
&\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_0^{\infty} g(x_1, \ldots, x_n, y)\left(\frac{\gamma}{\gamma_0}\right) e^{-(\gamma-\gamma_0)y} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi h}} \exp\left((\mu - \mu_0)x_i - \frac{(\mu^2 - \mu_0^2)h}{2}\right) \\
&\qquad \cdot \gamma_0 e^{-\gamma_0 y}\, dy \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{(x_i - \mu_0)^2}{2h}\right) dx_i \\
&\quad = E[g(\mu_0 h + \sqrt{h}N_1(0,1), \ldots, \mu_0 h + \sqrt{h}N_n(0,1), W(\gamma_0)) \cdot L_{nh}(\mu, \gamma; \mu_0, \gamma_0)]
\end{aligned}
$$

where

$$
L_t(\mu, \gamma; \mu_0, \gamma_0) = \exp\left((\mu - \mu_0)(\mu_0 t + B(t)) - \frac{(\mu^2 - \mu_0^2)t}{2}\right) \cdot \left(\frac{\gamma}{\gamma_0}\right) \exp(-(\gamma - \gamma_0)W(\gamma_0))
$$

Recognizing an indicator random variable as a special type of nonnegative function, we find that

$$
P_\theta(V(T_h) \in (a_{i-1}\, a_i]) = E_{\theta_0}[I(V(T_h) \in (a_{i-1}, a_i])L_{T_h}(\mu, \gamma; \mu_0, \gamma_0)]
$$

This idea permits the dependence on θ to be moved out of the nonsmooth indicator variable into the (smooth) function $L_t(\cdot)$. Differentiation is now possible:

$$
\begin{aligned}
\frac{\partial}{\partial \mu} P_\theta(V(T_h) \in (a_{i-1}, a_i])\bigg|_{\theta=\theta_0} &= E_{\theta_0}[I(V(T_h) \in (a_{i-1}, a_i])B(T_h)] \\
\frac{\partial}{\partial \gamma} P_\theta(V(T_h) \in (a_{i-1}, a_i])\bigg|_{\theta=\theta_0} &= E_{\theta_0}\left[I(V(T_h) \in (a_{i-1}, a_i])\left(\frac{1}{\gamma_0} - W(\gamma_0)\right)\right]
\end{aligned}
$$

This method yields unbiased estimators for the gradient of $P_\theta(V(T_h) \in (a_{i-1}, a_i])$ (thereby yielding gradient estimators for the likelihood). This approach to constructing simulation-based gradient estimators is called the likelihood ratio gradient estimators; Detailed in Glynn.[13]

When such unbiased gradient estimators are used with algorithm (11), we obtain what is known as the RM algorithm.[14] Note that in our setting, a small bias is incurred when multiplying the estimator of $P_\theta(V(T_h) \in (a_{i-1}, a_i])$ with that of $\nabla P_\theta(V(T_h) \in (a_{i-1}, a_i])$. (The product of two correlated unbiased estimators is biased.) This bias can be made arbitrarily small by letting the number of independent simulations per iteration of the RM algorithm to be large.

With regard to our tumor growth model, we implemented the RM algorithm both with the earlier (slightly biased) gradient estimator θ_5 and with a recently developed unbiased variant θ_5 (found in Glynn PW, Salzman P, personal communication; manuscript available from author). Our results are reported in Section 9.

8 Closed form for the likelihood

As discussed in Section 1, the geometric Brownian motion tumor growth model is a model that lies right at the boundary of what can be computed explicitly in 'closed-form'. In particular, through the use of stochastic calculus, one can establish that for each $\theta \in \Lambda$, the function

$$u(x) \stackrel{\triangle}{=} P_\theta(V(T) > v | V(0) = x) \tag{12}$$

satisfies an ordinary differential equation, with a solution that can be expressed as a (infinite series) special function. Because the likelihood $P_\theta(V(T) \in (a_{i-1}, a_i])$ can be expressed as the difference of two such solutions [one involving Equation (12) with $v = a_{i-1}$ and the other involving $v = a_i$], this allows us to compute the likelihood for this model in closed-form.

Theorem 1. *The function defined by Equation (12) satisfies the ordinary differential equation*

$$\frac{1}{2}x^2\frac{\mathrm{d}^2}{\mathrm{d}x^2}u(x) + \left(\mu + \frac{1}{2}\right)x\frac{\mathrm{d}}{\mathrm{d}x}u(x) - \gamma x u(x) = -\gamma x \quad \text{for } x > v \tag{13}$$

$$\frac{1}{2}x^2\frac{\mathrm{d}^2}{\mathrm{d}x^2}u(x) + \left(\mu + \frac{1}{2}\right)x\frac{\mathrm{d}}{\mathrm{d}x}u(x) - \gamma x u(x) = 0 \quad \text{for } 0 < x \le v \tag{14}$$

subject to $\lim_{x\to\infty} u(x) = 1$, $0 \le u(x) \le 1$ *for* $x > 0$, *and continuity of* $u(\cdot)$ *and* $u'(\cdot)$. *The unique solution is*

$$u(x) = \begin{cases} 2\sqrt{2\gamma}K_{2\mu+1}(2\sqrt{2\gamma v})v^{\mu+(1/2)}I_{2\mu}(2\sqrt{2\gamma x})x^{-\mu} & \text{if } x \le v \\ 2\sqrt{2\gamma x}(I_{2\mu}(2\sqrt{2\gamma x})K_{2\mu+1}(2\sqrt{2\gamma x}) + K_{2\mu}(2\sqrt{2\gamma x})I_{2\mu+1}(2\sqrt{2\gamma x})) & \\ -2\sqrt{2\gamma}x^{-\mu}K_{2\mu}(2\sqrt{2\gamma x})I_{2\mu+1}(2\sqrt{2\gamma v})v^{\mu+(1/2)} & \text{if } x > v \end{cases}$$

where $I_v(\cdot)$ *is the modified Bessel function of the first kind,*

$$I_v(x) = \sum_{n=0}^{\infty} \frac{1}{n!\Gamma(1+v+n)}\left(\frac{x}{2}\right)^{2n+v}$$

and $K_v(\cdot)$ *is the modified Bessel function of the second kind*,

$$K_v(x) = \frac{\pi}{2}\frac{I_v(x) - I_{-v}(x)}{\sin(v\pi)}$$

With this closed form, we can now analytically compute the (exact) likelihood $L(\theta)$. The (exact) likelihood surface can now be numerically optimized to compute the (exact) maximum likelihood estimator $\hat{\theta}$. Our choice of numerical optimizer was the Nelder–Mead Simplex package in Matlab.

9 Numerical comparison

As discussed in Section 8, the likelihood surface $L(\theta)$ can actually be computed in closed form for this model; Figure 1 shows a graph of the likelihood surface. The likelihood surface is quite flat in the μ-direction, suggesting both that the maximum likelihood estimator $\hat{\mu}$ will exhibit significant variability in estimating the true value μ^*

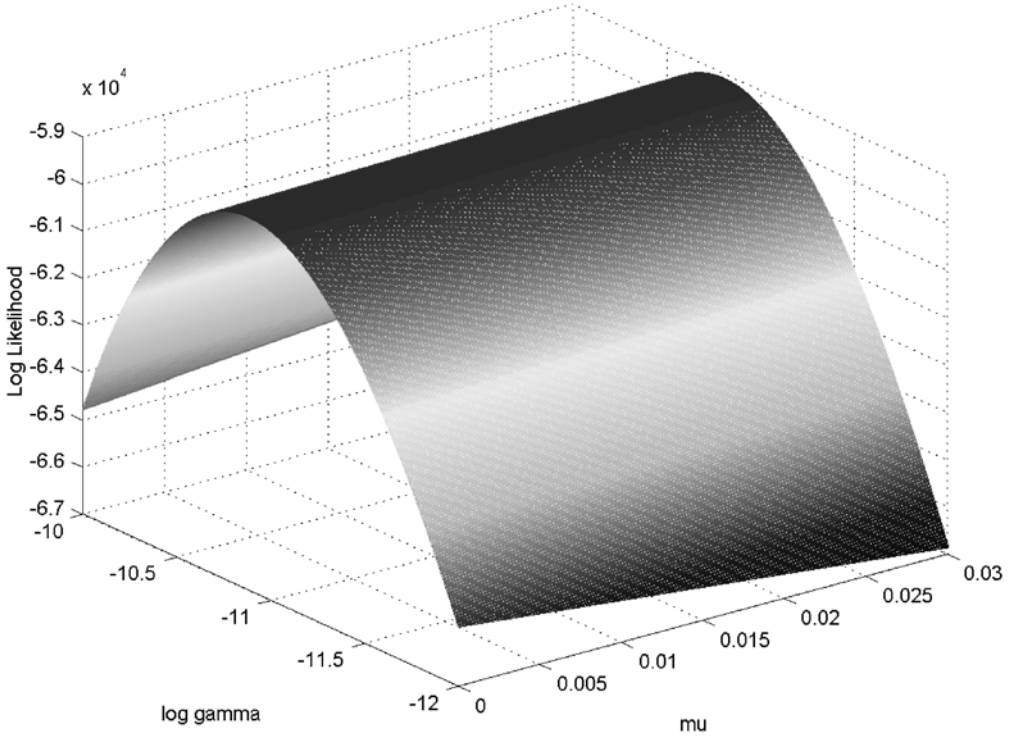

Figure 1. Likelihood surface plot obtained from the analytic solution.

and that numerical optimization in this variable will be challenging. The numerical optimization algorithm yielded maximum likelihood estimators

$$\hat{\mu} = 0.0154$$
$$\hat{\gamma} = 1.95 \times 10^{-5}$$

We used the bootstrap to compute a confidence region for the (true) parameter values μ^* and γ^*. We did this by simulating 1000 synthetic SEER data sets (each having sample size of 35 504 women), each independently simulated under the geometric Brownian motion model with parameters $\hat{\mu}$ and $\hat{\gamma}$. For each of the 1000 synthetic data sets, we computed the corresponding maximum likelihood estimators $(\hat{\mu}_i^{\text{boot}}, \hat{\gamma}_i^{\text{boot}})$ described in Section 8. The bootstrap of $(\hat{\mu}_i^{\text{boot}} - \hat{\mu}: 1 \leq i \leq 1000)$ and $(\log \hat{\gamma}_i^{\text{boot}} - \log \hat{\gamma}: 1 \leq i \leq 1000)$ are discussed further in Efron and Tibshirani.[15] These histograms (Figure 2) describe the sampling variability of $\hat{\mu}$ and $\log \hat{\gamma}$. The large probability mass in the histogram of $\hat{\mu}_i^{\text{boot}} - \hat{\mu}$ is due to the nonnegativity constraint on μ.

The 95% confidence intervals for μ^* and $\log \gamma^*$ are [0.00079, 0.02950] and [−10.89, −10.81], respectively. The estimated standard errors for $\hat{\mu}$ and $\log \hat{\gamma}$ are 0.007 and 0.02, respectively. As expected, the statistical sampling error in $\hat{\mu}$ as an estimator of μ^* is relatively larger than that for $\log \hat{\gamma}$.

We turn next to a discussion of our three simulation-based estimators $\hat{\theta}_4$, $\hat{\theta}_5$ and $\hat{\theta}_6$. To provide a fair comparison of the convergence characteristic of the three estimators, we stopped each of the three iterative algorithms after a total of 40 million women had been simulated (per algorithm). For the KW estimator, each iteration requires simulation at four different points in the parameter space Λ; 2000 women are simulated at each such parameter point (using the common random numbers approach described in Section 6). With 8000 women simulated per iteration, a total of 5000 iterations could be computed (under our 'budget constraint' of 40 million women in total). For the biased and unbiased RM estimators $\hat{\theta}_5$ and $\hat{\theta}_6$, we also allocated 8000 women per iteration, yielding 5000 iterations for each of these algorithms.

For all three algorithms, each iteration was started at the 'initial guess' for $(\hat{\mu}, \log \hat{\gamma})$ of $(0.1, -10)$. As described in Section 6, the KW algorithm was run with $\delta_1 = 0.05$ and $\delta_2 = 0.1$, respectively. The gain constants, a_1 and a_2, are 2.857×10^{-4} and 0.02857,

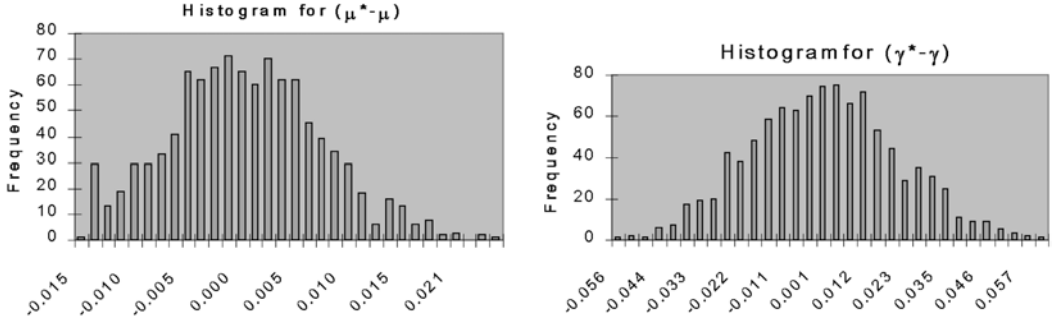

Figure 2. Histogram for $\hat{\mu}_i^{\text{boot}} - \hat{\mu}$ and $\log \hat{\gamma}_i^{\text{boot}} - \log \hat{\gamma}$.

respectively, for both biased and unbiased RM algorithms, and 5.633×10^{-6} and 2.817×10^{-4} for KW algorithm.

To evaluate the simulation error, we independently ran each algorithm five times from the same initial guess. This enables us to compute confidence regions for $\hat{\mu}$ and $\hat{\gamma}$ that are intended to capture the Monte Carlo variability of our three simulation-based estimators; The construction of confidence regions for stochastic approximations are discussed in Hsieh and Glynn.[16] The three confidence regions are displayed in Figure 3. Note that for 5000 iterations, the confidence region for the KW estimator does not cover $(\hat{\mu}, \log \hat{\gamma})$; the algorithm has not yet converged to the optimizer (although it is reasonably close).

10 Concluding remarks

We have described how simulation-based algorithms can be used to compute parameter estimators in the setting of a natural history model for breast cancer. All three

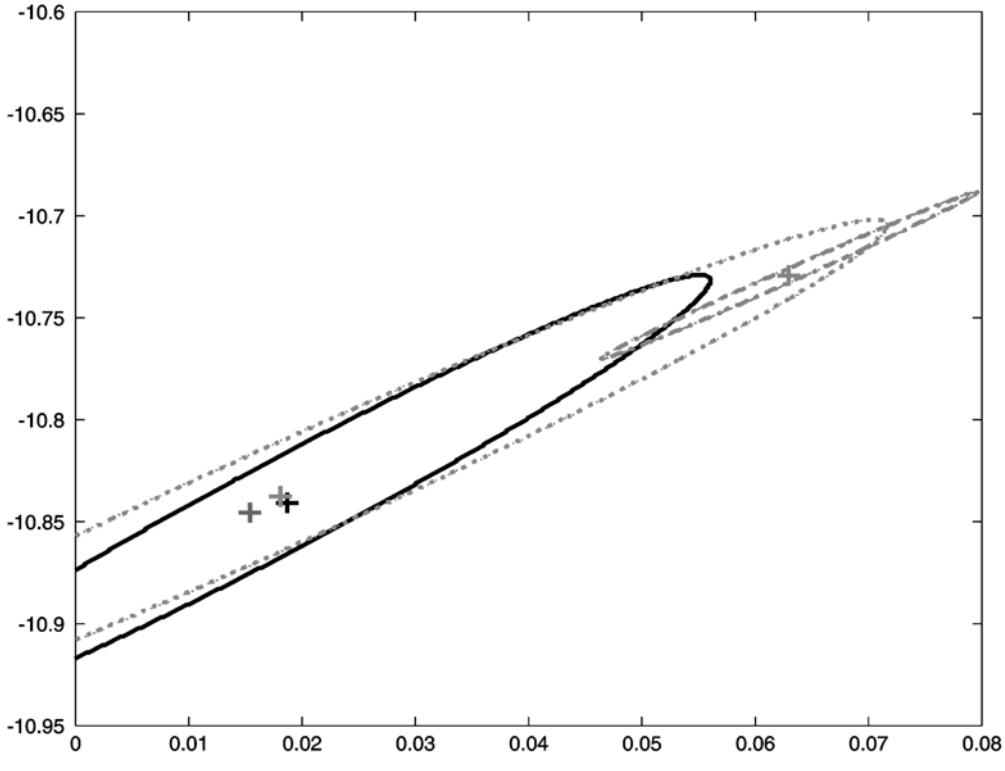

Figure 3. The confidence regions for the estimators from different simulation algorithms. RM unbiased – solid line, RM biased – dotted line, KW – dashed line, the analytical optimum solution – (0.0154, 1.95E-5).

implemented simulation methods give solutions that are close to the analytically computed maximum likelihood estimator. In future work, we intend to study the 'goodness of fit' of our geometric Brownian motion model relative to the deterministic growth model described in Section 2. This should shed some light on the question of environmental explanations for breast cancer versus genetic explanation for tumor growth.

Acknowledgement

We gratefully acknowledge funding from NIH grants 5R01 CA82904 and 1U01 CA097420.

References

1 Duffie D, Singleton K. Simulated moments estimation of Markov models of asset prices. *Econometrica* 1993; **61**: 929–52.
2 Steel G. *Growth kinetics of tumors*. Oxford: Clarendon Press, 1977.
3 Øksendal B. *Stochstic differential equations*. Springer, 1998.
4 Bartoszynski R, Edler L, Hanin L, Kopp-Schneider A, Pavlova L, Tsodikov A, Zorin A, Yakovlev A. Modelling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis. *Mathematical Biosciences* 2001; **171**: 113–42.
5 Chappuis PO, Nethercot V, Foulkes WD. Clinico-pathological characteristics of BRCA1- and BRCA2- related breast cancer. *Seminars in Surgical Oncology* 2000; **18**: 287–95.
6 Armes JE, Egan AJM, Southey MC, Dite GS, McCredie MRE, Giles GG, Hopper JL, Venter DJ. The histologic phenotypes of breast carcinoma occuring before age 40 years in women with and without BRCA1 or BRCA2 germline mutations. *Cancer* 1998; **83**: 2335–45.
7 Lehmann EL, Casella G. *Theory of point estimation*. Springer, 2001.
8 Ensor KB, Glynn PW. Grid-based simulation and the method of conditional least square. *Proceedings of the 1996 Winter Simulation Conference* New York: ACM Press. 1996, 325–31.
9 Rubinstein RY, Samorodnitsky G. Variance reduction by the use of common and antithetic random variables. *Journal of Statistical Computation and Simulation* 1985; **22**: 161–80.
10 Nelder JA, Mead R. A simplex method for function minimization. *Computer Journal* 1965; **7**: 308–13.
11 Pflug G. *Stochastic optimization*. Kluwer, 1997.
12 Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* 1952; **23**: 462–66.
13 Glynn PW. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM* 1990; **33**(10): 75–84.
14 Robbins H, Monro S. A stochastic approximation method. *Annals of Mathematical Statistics* 1951; **22**: 400–407.
15 Efron B, Tibshirani R. Bootstrap methods for standard erros, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986; **1**: 54–77.
16 Hsieh M-H, Glynn PW. Confidence regions for stochastic approximation algorithms, *Proceedings of the 2002 Winter Simulation* 2002, 370–76.
17 Relton FE. *Applied bessel functions*. New York: Dover Pub. Inc., 1965.
18 Steele JM. *Stochastic calculus and financial applications*. New York: Springer-Verlag, 2000.

Appendix A: Proof of Proposition 1

Let $E_{\mu,\sigma^2,\gamma}(\cdot)$ be the expectation operator associated with $P_{\mu,\sigma^2,\gamma}(\cdot)$. Note that the distribution of V under $P_{\mu,\sigma^2,\gamma}(\cdot)$ depends only on μ and σ^2; call it $P_{\mu,\sigma^2}(\cdot)$ and its corresponding expectation $E_{\mu,\sigma^2}(\cdot)$.

Observe that

$$P_{\mu,\sigma^2,\gamma}(T > t|V) = \exp\left(-\gamma \int_0^t V(s)\,\mathrm{d}s\right)$$

It follows that for any nonnegative function g,

$$\begin{aligned} E_{\mu,\sigma^2,\gamma}[g(V(T))] &= \gamma \int_0^\infty E_{\mu,\sigma^2,\gamma}\left[g(V(t))V(t)\exp\left(-\gamma\int_0^t V(s)\,\mathrm{d}s\right)\right]\mathrm{d}t \\ &= \gamma \int_0^\infty E_{\mu,\sigma^2}\left[g(V(t))V(t)\exp\left(-\gamma\int_0^t V(s)\,\mathrm{d}s\right)\right]\mathrm{d}t \end{aligned}$$

Substituting $u = t\sigma^2$ and $r = s\sigma^2$, we find that

$$E_{\mu,\sigma^2,\gamma}[g(V(T))] = \frac{\gamma}{\sigma^2}\int_0^\infty E_{\mu,\sigma^2}\left[g\left(V\left(\frac{u}{\sigma^2}\right)\right)V\left(\frac{u}{\sigma^2}\right)\exp\left(-\frac{\gamma}{\sigma^2}\int_0^u V\left(\frac{r}{\sigma^2}\right)\mathrm{d}r\right)\right]\mathrm{d}u$$

Let $e(t) = t$ for $t \geq 0$. For any set C, the scaling properties of Brownian motion imply that

$$\begin{aligned} P_{\mu,\sigma^2}\left(\log V\left(\frac{\cdot}{\sigma^2}\right) \in C\right) &= P\left(\frac{\mu}{\sigma^2}e(\cdot) + \sigma C\left(\frac{\cdot}{\sigma^2}\right) \in C\right) \\ &= P\left(\frac{\mu}{\sigma^2}e(\cdot) + C(\cdot) \in C\right) \\ &= P_{(\mu/\sigma^2),1}(\log V(\cdot) \in C) \end{aligned}$$

Consequently,

$$\begin{aligned} E_{\mu,\sigma^2,\gamma}[g(V(T))] &= \frac{\gamma}{\sigma^2}\int_0^\infty E_{(\mu/\sigma^2),1}\left[g(V(u))V(u)\exp\left(-\gamma\int_0^u V(s)\,\mathrm{d}s\right)\right]\mathrm{d}u \\ &= E_{(\mu/\sigma^2),1,(\gamma/\sigma^2)}[g(V(T))] \end{aligned}$$

proving the result. ■

Appendix B: Proof of Theorem 1

Standard properties of Bessel functions guarantee that the function $u(\cdot)$ defined in terms of the modified Bessel functions of the first and second kinds satisfies the stated ordinary differential equation subject to the given boundary conditions; basic properties of Bessel functions are presented in Relton.[17] It therefore remains only to show that this function $u(x)$ is indeed the probability $P(V(T) > v|V(0) = x)$.

Let $P_x(\cdot) \stackrel{D}{=} P(\cdot|V(0) = x)$ and $E_x(\cdot)$ be the associated expectation operator. Put

$$\chi(t) = u(V(t)) \exp\left(-\gamma \int_0^t V(s)\,\mathrm{d}s\right)$$

Because $u'(\cdot)$ is continuous at v, Itô's formula applies (see p. 111 of Steele[18]), yielding

$$\begin{aligned}\mathrm{d}\chi(t) = {} & [u'(V(t))V(t)\left(\mu + \frac{1}{2}\right) - u(V(t))\gamma V(t) + \frac{1}{2}u''(V(t))] \exp\left(-\gamma \int_0^t V(s)\,\mathrm{d}s\right) \mathrm{d}t \\ & + u'(V(t))V(t) \exp\left(-\gamma \int_0^t V(s)\,\mathrm{d}s\right) \mathrm{d}B(t)\end{aligned}$$

The function u satisfies the stated differential equation, so

$$\mathrm{d}\chi(t) = -\gamma I(V(t) > v)V(t) \exp\left(-\gamma \int_0^t V(s)\,\mathrm{d}s\right) \mathrm{d}t + u'(V(t))V(t) \exp\left(-\gamma \int_0^t V(s)\,\mathrm{d}s\right) \mathrm{d}B(t)$$

Because it can be verified that $u'(\cdot)$ is bounded. It follows that

$$\int_0^t u'(V(s))V(s) \exp\left(-\gamma \int_0^s V(u)\,\mathrm{d}u\right) \mathrm{d}B(s)$$

is a square-integrable martingale. Consequently,

$$E_x[\chi(t)] - E_x[\chi(0)] = -\gamma E_x\left[\int_0^t I(V(s) > v)V(s) \exp\left(-\gamma \int_0^s V(u)\,\mathrm{d}u\right) \mathrm{d}s\right]$$

By monotone convergence, the right hand side converges to

$$-\gamma E_x\left[\int_0^\infty I(V(s) > v)V(s) \exp\left(-\gamma \int_0^s V(u)\,\mathrm{d}u\right) \mathrm{d}s\right]$$

as $t \to \infty$. But the proof of Proposition 1 establishes that this expectation is precisely the quantity $-P_x(V(T) > v)$.

On the other hand, $\chi(t) \to 0$ almost surely as $t \to \infty$. As $|\chi(t)|$ is bounded by one, the Bounded Convergence Theorem proves that $E_x[\chi(t)] \to 0$ as $t \to \infty$. Observing that $E_x[\chi(0)] = u(x)$ completes the proof. ■