# Working Paper

## Parameter Estimation for Survival Model

*A. Asachenkov*
*B. Sobolev*

WP-93-1
January 1993

IIASA
International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

# Parameter Estimation for Survival Model

*A. Asachenkov*
*B. Sobolev*

WP-93-1
January 1993

*Working Papers* are interim reports on work of the International Institute for Applied Systems Analysis and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute or of its National Member Organizations.

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

## Foreword

This paper deals with clinical data processing for the patient after radical surgery to remove the solid tumor. The survival model with quadratic mortality rate, problem of parameter estimation and corresponding numerical algorithms are discussed. The respective problems arise from applied motivations that come from medical issues.

# Contents


**1 Background and Motivation** **1**

1.1 State equation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Mortality dynamics . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.3 Survival model with quadratic mortality rate . . . . . . . . . . . . . . . . . . 4

**2 Statement of the problem** **5**

**3 Consistency of Maximum Likelihood Estimates** **7**

3.1 Marginal density function . . . . . . . . . . . . . . . . . . . . . . . . 8

3.2 Joint probability density . . . . . . . . . . . . . . . . . . . . . . . . 10

**4 Numerical Algorithms** **13**

4.1 Iterative scheme . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

4.2 Local convergence analysis . . . . . . . . . . . . . . . . . . . . . . . 20

# Parameter Estimation for Survival Model

*A. Asachenkov*
*B. Sobolev*

# 1 Background and Motivation

The state of the organism during disease is assessed according to the *laboratory tests.* The disorders in normal functioning of main homeostatic systems, which are caused by the disease, lead to a deviation of these indices from the values which correspond to the state of the healthy body. A systematic disorder of the homeostasis which acquires a stable and uncompensated character considerably raises the probability of the patient's death.

This general definition permits us to construct a model of the oncological disease as it links two of its most important properties: the rate of growth of its functional disorders, i.e. the activity of the disease; and the intensity of decreasing the number of patients surviving for certain periods of time [10, 15].

Indeed, when assessing the activity of the disease using laboratory tests, the totality of individual trajectories of physiological indices can be put into correspondence with the survival function. Thereby, the dynamics of variation in the death probability can be determind by statistical characteristics of the distribution of those trajectories. Note that only the group of mortality characteristics can be acurately observed whereas the homeostatic disorders can be assessed only indirectly using the available clinical data. More details are presented in [20-22].

## 1.1 State equation

Let us consider the dynamics of observed indices from the patients after surgery. Let $t = 0$ be the instant of surgery. Denote $x(t) \in \mathbf{R}^n$ vector measured in clinic indices. Development of disease leads to deviations of these indices from the values corresponding to the healthy organism. Let the dynamics of the clinically measured indices, on the average, be described by the equation

$$\frac{dx(t)}{dt} = f(x;\alpha), \qquad x(0) = x_0 \geq 0, \qquad t \geq 0, \tag{1}$$

$$x \in \mathbf{R}^n, \qquad \alpha \in \mathbf{R}^m.$$

A solution of equation (1) $x(t, \alpha^*)$ which describes the average trajectory in the group of patients with "favorable clinical history" is called a *support solution* or a *reference trajectory,* and the vector $\alpha^*$ is called a *reference* or *support vector.*

To model the individual trajectories of these indices which have presumably stochastic character, the ODE with random perturbations of parameters can be used

$$\frac{d}{dt}x_t^\varepsilon = f(x_t^\varepsilon, \alpha^* + \xi_{t/\varepsilon}),$$

$$x_0^\varepsilon = x_0, \qquad t \in [0,T].$$

Here $x_t^\varepsilon$ is a perturbed solution, $\varepsilon > 0$ is a small parameter, $\xi_t$ is a stochastic process such that $E\xi_t = 0$ and $\text{cov}(\xi_t, \xi_{t+\tau}) \to 0$ as $\tau \to \infty$. These trajectories can be considered as a result of small perturbations of the dynamic system. The perturbed motion described by this model is the fast random fluctuation along the reference trajectory $x(t,\alpha^*)$. Let $Y_t^\varepsilon = x_t^\varepsilon - x(t,\alpha^*)$ be a deviation between the perturbed motion and reference trajectory. Then, the process $Y_t^\varepsilon$ can be approximated by the lineal stochastic differential

$$dY_t = \frac{\partial}{\partial x} f(\bar{x}_t, \alpha^*) Y_t dt + \frac{\partial}{\partial \alpha} f(\bar{x}_t, \alpha^*) dw_t, \tag{2}$$

$$Y(0) = 0, \qquad t \in [0,T],$$

where $\bar{x}_t = x(t,\alpha^*)$ is a support solution, and $w_t$ is a Gaussian process with independent increments, such that $Ew_t = 0$, $\text{cov}\,(w_t, w_t) = \Gamma t$, where $\Gamma$ is a matrix of intensities of $\xi_t$.

Rewrite the equation (2) in a more convenient form

$$dY_t = a(t) Y_t dt + D(t) dw_t \tag{3}$$

where $w_t$ is a Wiener process, therefore the matrix $D(t)$ satisfies the condition $DD^T = b\Gamma b^T$. The functions $a(t) = \partial f(x,\alpha)/\partial x$ and $b(t) = \partial f(x,\alpha)/\partial\alpha$ depend on the right-hand side of the model (1) and reflects our knowledge about the process in question.

## 1.2 Mortality dynamics

Assume, for simplicity, that we deal with just one cohort (a closed group of patients all of which have the same time entry to study) and there exist so-called failue or termination time $T$ for each trajectory. This failure is associated with death or specific health changes and can be described by a set of trajectories $(x_t(w_i), i = 1, \ldots, N)$ and the corresponding set of termination times $\Theta = \{T_1, \ldots, T_N\}$, where $N$ is a sample size. In principle, we can derive the frequency distribution at each instant of time from these $N$ individual trajectories as $N \to \infty$. Hence the trajectory set generates the probability density function of failure times, and we can study the relation between the trajectory set and probability distribution of random time of failure.

Let $T$ be a nonnegative continuous random variable (termination time) with a probability density $p(t)$. Then the distribution function $F(t)$ defines the probability that an individual trajectory will fail at or before time $t$. It is possible to define a continuous function of time

$$s(t) = 1 - F(t)$$

which represents the probability that the individual will survive to time $t$. The function $s(t)$ is called a *survival* function. If the probability density $p(t)$ exists, then

$$s(t) = \int_t^\infty p(v)dv, \qquad s(0) = 1.$$

It should be noted that $s(t_1) \geq s(t_2), t_2 \geq t_1$ hence $s(t)$ is a monotonic decreasing function of time.

Another important characteristic in survival analysis is the so-called mortality intensity $\lambda(t)$. It is simple to obtain the *failure rate form* for intensity as

$$\lambda(t) = -\frac{1}{s(t)}\frac{d}{dt}s(t) = -\frac{d}{dt}\log s(t).$$

Then mortality dynamics in a group is described by the survival function

$$\frac{d}{dt}s(t) = -\lambda(t)s(t), \qquad s(0) = 1,$$

$$s(t) = \exp\{-\int_0^t \lambda(u)du\}, \quad s(0) = 1.$$

The total mortality intensity in the interval [ 0 , T ] is

$$\Lambda(t) = \int_0^t \lambda(u)du = \log s(t), \quad \Lambda(0) = 0.$$

This may be interpreted as a pathological pressure upon the organism caused by the disease up to the instant of time $t$.

So, the rate of change in the survival function at $t$ is presented as the product of two independent factors: the failure rate and survival function. We can write the probability density function in the form

$$p(t) = \lambda(t)s(t).$$

Let us formulate some assumption that is usually used. It is known, from clinical practice, that the risk of failure (hazard of death in cancer) depends on the state of the organism (Manton, et al. 1984). Models which functionally relate the survival function to the random process were suggested by Woodbury and Manton (1977).

**Assumption 1.1** *A probability of occurrence of failure that is associated with mortality or morbidity is functionally related to the state of the body. Let the probability of occurrence of failure at time t for a given trajectory $x_t(w), w \in \Omega$ be* conditional *on the path of measurable indices over time*

$$s(t|x_t(w)) = P\{T > t|x_t(w)\}.$$

**Assumption 1.2** *For each trajectory $x_t(w_i), i = 1, \ldots, N$, there exists a piece-wise continuous function $\alpha_t^i$ such that*

$$x_t(w_i) = x(t, \alpha_t^i), \quad t \in [t_0, t_1, \ldots, T_i]$$

*where $x(t, \alpha^i)$ is a solution of (1), then $\alpha$ is replaced by $\alpha_t^i$.*

**Assumption 1.3** *The unconditional probability of failure, $s(t)$, is interpreted as*

$$s(t) = E\{s(t|x_t(w))\}$$

*where conditional probability is averaged over trajectories of the random process.*

In our model we consider the relationship between the deviations of measurable indices from reference trajectory and survival function $s(t)$. The deviations from the reference trajectory $x(t, \alpha^*)$ are caused by unmeasured endogenous and exogenous factors and can be considered as Gaussian-Markov process which satisfies the linear SDE

$$dY_t = a(t)Y_t dt + D(t)dw_t \qquad Y_0 = 0, \tag{4}$$

where $a(t)$ is a $n \times n$ matrix, $D(t)$ is a $n \times m$ matrix, $DD^T = b(t)\Gamma b^T(t)$, $w_t$ is a Wiener process. $Y_t$ is distributed as $N(m(t), \gamma(t))$ where a vector of means and variance matrix satisfy the equations

$$\frac{d}{dt}m(t) = a(t)m(t), \quad \frac{d}{dt}\gamma(t) = a(t)\gamma(t) + \gamma(t)a^T(t) + b(t)\Gamma b^T(t). \tag{5}$$

Woodbury and Manton suggested to consider the failure rate conditional on the path of physiological covariates over time

$$\mu(t, Y_t) = \lim_{\Delta t - 0} \frac{1}{\Delta t} Pr\{T \in (t, t + \Delta t) | T > t, (Y_v, 0 \leq v \leq t)\},$$

as a function which defines the survival chances for an individual with trajectory $(Y_v, 0 \leq v \leq t)$ and

$$s(t|Y_t) = \exp\{-\int_0^t \mu(u, Y_u) du\}.$$

This function is called the *individual* mortality rate in Yashin et al. (1986). However, this exponential formula does not necessarily hold without some conditions. Necessary and sufficient conditions for this expression were founded by Yashin and Arjas (1988). The relationship between conditional and unconditional failure rate was presented by Yashin (1985). Hence, the mortality dynamics observed in the group of patients is determined by

$$\lambda(t) = E\{\mu(Y_t) | T > t\}$$

where $E$ is averaging over all individual trajectories.

## 1.3 Survival model with quadratic mortality rate

Unfortunately, there are no general recommendations for the choice of the analytical form of $\lambda(t)$ as a function of measurable variables. The choice of corresponding analytical relation is determined by the analysis of experimental data and the ease of mathematical manipulations. On the other hand, the simplest approximation of an unknown function is the Taylor series. This yields a quadratic failure rate

$$\mu_0(t) + Y_t^T Q(t) Y_t,$$

where $\mu_0(t)$ is a function which determines the standard death hazard not related to a given disease, and $Q(t)$ is a $n \times n$-symmetrical positive definite matrix.

Thus, the conditional survival function in the group of patients can be presented in the form

$$s(t|Y_t) = \exp\{-\int_0^t [\mu_0(u) + Y_u^T Q Y_u] du\}.$$

The quadratic form of failure rate represents increased risk at both high and low physiological values. It specifies a range of values as optimal in the sense of survival. The statistical properties of the random process $\{Y_t, t \geq 0\}$ with quadratic failure selection and marginal distribution of the failure time is given in

**Proposition 1.1** (Yashin, 1985, [12]). *Let the random process*

$$\{Y_t(w) \in R^n, t \geq 0, w \in \Omega\}$$

*satisfy a linear SDE; the individual failure rate $\mu(t, Y_t) = \mu_0(t) + Y_t^T Q(t) Y_t$ satisfies quadratic hazard; and assume that $Y_0$ is distributed as $N(m_0, \gamma_0)$. Under these conditions*

$$\frac{d}{dt} s(t) = -(\mu_0(t) + m^T(t) Q(t) m(t) + tr[Q(t)\gamma(t)]) s(t), \quad s(0) = 1, \tag{6}$$

*and $Y_t$ is distributed as $N(m(t), \gamma(t) | T > t)$, where $m(t)$ and $\gamma(t)$ are the solutions of the following ODE*

$$\frac{d}{dt} m(t) = a(t) m(t) - 2\gamma(t) Q(t) m(t), \quad m(0) = m_0 \tag{7}$$

$$\frac{d}{dt}\gamma(t) = a(t)\gamma(t) + \gamma(t)a(t)^T + b(t)\Gamma b^T(t) - 2\gamma(t)Q(t)\gamma(t), \quad \gamma(0) = \gamma_0$$

*where*

$$m(t) = E\{Y_t|T \geq t\}$$

$$\gamma(t) = E\{(Y_t - m(t))(Y_t - m(t))^T|T \geq t\}.\square$$

Because $s(t)$ is the marginal probability of survival to time $t$, we will refer to $\lambda(t)$ as the marginal cohort failure rate. Similarly $N(m(t), \gamma(t)|T > t)$ will be called as the marginal distribution of $Y_t$ among the survivors at time $t$. Proposition 1.1 yields the mathematical relationship between the marginal failure rate for cohort and the parameters governing change in the means and covariance of the physiological process related to the failure. Now the marginal probability density for the random failure time in terms of statistical properties of the random process is

$$p(t) = \lambda(t)\exp[-\int_0^t \lambda(v)dv].$$

Thus, the mortality intensity observed in the group of patients is related to the dynamics of clinically controlled indicators by the formula

$$\lambda(t) = \mu_0(t) + m^T(t)Q(t)m(t) + tr[Q(t)\gamma(t)].$$

## 2 Statement of the problem

Consider the survival model

$$\begin{aligned} \frac{d}{dt}s(t) &= -[\mu_0(t) + m^T(t)Q(t)m(t) + tr(Q(t)\gamma(t))]s(t) \\ \frac{d}{dt}m(t) &= a(t)m(t) - 2\gamma(t)Q(t)m(t) \\ \frac{d}{dt}\gamma(t) &= a(t)\gamma(t) + \gamma(t)a^T(t) + b(t)\Gamma(t)b^T(t) - 2\gamma(t)Q(t)\gamma(t) \end{aligned} \tag{8}$$

$$s(0) = 1, \quad m(0) = m_0, \quad \gamma(0) = \gamma_0,$$

where

$$m(t) = E\{Y_t|T > t\}$$

$$\gamma(t) = E\{(Y_t - m(t))(Y_t - m(t))^T|T > t\},$$

and $Q$ and $\Gamma$ are unknown matrices.

Here $Y_t$ is a stochastic process such that

$$dY = a(t)Y_t dt + D(t)dw_t, \quad Y(0) = 0$$

$a(t) = \frac{\partial f(\bar{x},\alpha)}{\partial \alpha}$ , $b(t) = \frac{\partial f(\bar{x},\alpha)}{\partial \alpha}$, are a known function $DD^T = b(t)\Gamma b^T(t)$, and $\bar{x}(t)$ is a solution of the problem

$$\frac{d}{dt}\bar{x}(t) = f(\bar{x}(t), \alpha^*), \quad \bar{x}(0) = x_0,$$

where

- $\alpha^*$ is a known vector of parameters,
- $\bar{x}(t)$ is a reference trajectory.

**Remark**. As a rule, the model parameters have a practical interpretation. For example, the elements of matrix $Q$ define the degree of influence on the failure dynamics of each state variable. Parameter estimation over clinical data helps to study the role of various factors which are related to failure.

Assume, for simplicity, that matrices $Q$ and $\Gamma$ are diagonal. Introduce a new $2m \times 1$ vector, $\beta$, of unknown parameters

$$\beta = (\ \mathrm{diag}\ Q,\ \ \mathrm{diag}\ \Gamma)^T.$$

Let $\beta^*$ be the true value. For each patient we have a set of measurements

$$X^i = \{X^i_{t_k} \in \mathbf{R}^n, t_k < T_k, k = 1, \ldots, N_i\}.$$

Using a reference trajectory for a group of $N$ patients, we can obtain the set of trajectory deviations $Y_N = \{Y^i, i = 1, 2, \ldots, N\}$. Other available information is the failure times $\Theta = \{T_1, \ldots, T_N\}$. The solution of the problem (8) can be used to define the unconditional or marginal probability density for random failure times

$$p(t, \beta) = \lambda(t, \beta) \exp[-\Lambda(t, \beta)],$$

$$\Lambda(t, \beta) = -\log s(t, \beta)$$

where $\Lambda$ is the cumulative failure rate function. According to the likelihood principle, we have a log-likelihood function, for example

$$L(\beta) = \log p(\Theta, \beta),$$

where

$$p(\Theta, \beta) = \prod_{t \in \Theta} p(t, \beta).$$

An estimate of the unknown vector $\beta$ is given by

$$\hat{\beta} = \arg \max_{\beta \in B} l(\beta),$$

where

$$l(\beta) = \sum_{t \in \Theta} \log \lambda(t, \beta) - \Lambda(t, \beta).$$

The difficulty of this procedure is associated with the functions $m(t, \beta), \gamma(t, \beta), \Lambda(t, \beta)$ which are the functions of unknown parameter $\beta$.

There are two variants of this estimation problem with respect to the likelihood function

(i) we can use the failure times only

(ii) we can use the conditional form of the likelihood function. In this case we have to consider the random deviations of the process in question. The consistency of MLEs and a numerical algorithm for searching such estimates are given by Sobolev, 1989 [11].

## 3 Consistency of Maximum Likelihood Estimates

Denote the joint density function for $N$ independent and identically distributed (i.i.d.) random values with the probability density $p(\cdot,\beta)$ by $p_N(\cdot,\beta)$ or $p_N(\beta)$, and the likelihood function for a set of $N$ observations $x_1,\ldots,x_N$ by

$$p_N(x_1,\ldots,x_N,\beta)=\prod_{i=1}^{N}p(x_i,\beta). \tag{9}$$

Any $\hat{\beta}_N(x)\in B$, which maximizes $p_N$ over $\beta\in B$ is called a *maximum likelihood estimate* (MLE). Therefore the maximum likelihood estimate is a solution of the problem

$$\max\{p_N(x_1,\ldots,x_N,\beta);\beta\in B\}. \tag{10}$$

Often it is convenient to maximize $\log p_N$ in place of $p_N$.

It should be mentioned that MLEs do not always exist. Moreover, if MLEs exist, they are not necessarily unique. Having introduced the distance between any $p_1$ and$p_2$ on the real line as

$$\rho^2(p_1,p_2)=\int_{\mathbf{R}}(\sqrt{p_1}-\sqrt{p_2})^2dx. \tag{11}$$

Pitman formulated the sufficient conditions fo the existence, uniqueness and consistency of MLEs. For any set $H$ which intersects $B$, we can write

$$p^*(x,H)=\sup\{p(x,\beta),\beta\in H\cap B\}$$

$$p_N^*(x_1,\ldots,x_N,H)=\sup\{p_N(x_1,\ldots,x_N,\beta),\beta\in H\cap B\}$$

**Theorem 3.1** (Pitman, 1979 [7]). *Let $X_1,\ldots,X_N$ be i.i.d. random elements with the probability density $p(x,\beta_*)$. Assume that*

*(i) if $\beta\neq\beta^*,\rho(p,f^*)>0$;*

*(ii) for each $x$ the density $p$ is an upper semi-continuous function of $\beta$ in $B$, i.e. if $\beta'\in B$,*

$$\lim\sup[p(x,\beta);|\beta-\beta'|<h]=p(x,\beta')\ \textit{with}\ h\to 0.$$

*(iii) If $H$ is a compact subset of $B$ which contains $\beta^*$, and if for some $r$*

$$E_*\log\frac{p_r(\beta^*)}{f_r^*(H)}>-\infty,$$

*then $\hat{\beta}_N\in H$ exists such that*

$$p_N(x_1,\ldots,x_n,\hat{\beta}_N)=p_N^*(x_1,\ldots,x_N,H)$$

*and with probability one, $\hat{\beta}_N\to\beta^*$ as $N\to\infty$.*

*(iv) If in addition, with set $H^c$ as inferior to $\beta^*$,*

$$E_* \log \frac{p_r(\beta^*)}{p_r^*(H^c)} > -0,$$

*then with probability one the likelihood function has a global maximum at $\hat{\beta}_N$*

$$p_N(x_1, \ldots, x_n, \hat{\beta}_N) = p_N^*(x_1, \ldots, x_N, B),$$

*for the great $N$.* □

**Remark.** A completely different approach was given by Cramer [1]. His theorem is based on Taylor expansion of $\log p_N(\beta)$ and guarantees the convergency of MLE to $\beta^*$ in probability.

## 3.1 Marginal density function

**Proposition 3.1** *Let $T_1, \ldots, T_N$ be non-negative independent random variables with the probability density $p(t, \beta^*), t \geq 0$,*

$$p(t, \beta) = \lambda(t, \beta) \exp\{-\Lambda(t, \beta)\}$$

*where*

$$\lambda(t, \beta) = \mu_0 + m^T(t) Q(t) m(t) + tr[Q(t)\gamma(t)]$$

*and the functions $m(t, \beta), \gamma(t, \beta), \Lambda(t, \beta)$ are solution of problem (8).*

*If $H$ is a compact subset of $B$ which contains $\beta^*$, then there exists local MLE in $H, \hat{\beta}_N$, such that with the probability one $\hat{\beta}_N \to \beta^*$ as $N \to \infty$.*

**Proof** (Sobolev, 1989 [11]). For the one-dimensional case the conditions (i) and (ii) of Pitman's theorem are defined by the uniqueness and continuity of the solution of (8). Indeed, from (11) account $\rho(\beta, \beta^*)$ for any vector $\beta$ from $H$ we can take values from 0 to $\sqrt{2}$

$$\rho^2 = 2 - 2\sqrt{pp^*} dv.$$

It is zero if and only if $p$ is equal $p^*$. The uniqueness of the solution of (8) with respect to the parameter implies the density function for each $t$

$$p : \mathbf{R}^{2m} \to \mathbf{R}^1$$

and for each vector $\beta$ from $B \subset \mathbf{R}^{2m}$. Continuity of these functions does result in the supremum of $p_N(x_1, \ldots, x_N, \beta)$ in any compact set $H$ from $B$ being attained for some $\beta \in H$.

A sufficient condition for (iii) from Pitman's theorem is the following:

for some $r$ function $p_r(\cdot, \beta)$ is bounded for all $\beta$ in $H$ and all $(x_1, \ldots, x_r)$

$$E_* \log p_r(\beta^*) > -\infty.$$

Suppose that $p_r(\beta) \leq C$, then $p_r^*(H) \leq C$. Therefore

$$E_* \log \frac{p_r(\beta^*)}{p_r^*(H)} = E_* \log p_r(\beta_*) - E_* \log p_r^*(H) > -\infty - \log C$$

From the independence of random variables $T_1, \ldots, T_N$ we have

$$\begin{aligned} E_* \log p_r(\beta_*) &= \sum_{k=1}^{r} E_* \log p(T_k, \beta^*) \\ &= \sum_{k=1}^{r} E_*\{\log \lambda(T_k, \beta^*) - \Lambda(T_k, \beta^*)\}. \end{aligned} \tag{12}$$

The solutions $m(t,\beta)$ and $\gamma(t,\beta)$ of (8) are continuous, and function $\lambda(t,\beta) > 0$ is also continuous and bounded. Therefore in $[0,\infty)$ it has minimum $\lambda_{\min}$, such that $\forall t \geq 0$

$$\lambda(t,\beta) \geq \lambda_{\min} > 0,$$

$$\log \lambda(t,\beta) \geq \log \lambda_{\min} > -\infty$$

and maximum $\lambda_{\max}$, such that $\forall t \geq 0$

$$\lambda(t,\beta) \leq \lambda_{\max}.$$

Similarly, the following evaluations

$$\Lambda(t,\beta) = \int_0^t \lambda(u,\beta)du \geq \lambda_{\min} t,$$

$$\Lambda(t,\beta) \leq \lambda_{\max} t$$

are valid. So,

$$\begin{aligned} E_* \log p_r(\beta^*) &\leq \sum_{k=1}^{r} E_*\{\log \lambda_{\min} - \lambda_{\max} T_k\} \\ &= \sum_{k=1}^{r} \{\log \lambda_{\min} - \lambda_{\max} E_* T_k\} = r\{\log \lambda_{\min} - \lambda_{\max} \bar{T}\}. \end{aligned} \tag{13}$$

For each $k$ the mean $E_* T_k$ equals $\bar{T}$, which is a non-negatively bounded variable

$$\int_0^\infty S(t)dt.$$

Consider the integral

$$\int_0^\tau t dF(t).$$

Integrating it by parts twice, we have

$$\tau F(\tau) - \int_0^\tau t dF(t) = -\tau[1 - p(\tau)] + \int_0^\tau S(t)dt.$$

Here

$$\tau[1 - F(\tau)] = \tau \int_\tau^\infty dF(t) < \int_\tau^\infty t dF(t) \to 0 \text{ as } \tau \to \infty,$$

and

$$\int_0^\tau t dF(t) \to \int_0^\tau S(t)dt \text{ as } \tau \to \infty.$$

Therefore

$$\bar{T} = \int_0^\infty S(t)dt = \int_0^\infty \exp[-\Lambda(t)]dt \leq \int_0^\infty \exp[-\lambda_{\min} t]dt = 1/\lambda_{\min}$$

and the inequality

$$E_* \log p_r(\beta^*) \geq r \log \lambda_{\min} > -\infty,$$

is valid. □

It means that for a great $N$, with probability one, function $p_N(x_1, \ldots, x_N, \beta)$ will have a maximum $\hat{\beta}_N \in H$ in some neighborhood of point $\beta^*$ with arbitrary small radius.

## 3.2 Joint probability density

For the $i$-th patient we have a set of measurements

$$Y^i = \{Y^i_{t_k} \in \mathbf{R}^n, t_k < T_i, k = 1, \dots, n_i\}. \tag{14}$$

Denote the joint distribution function at $t_0, t_1, \dots, T_i$ as

$$p(Y) = p(Y_0, Y_1, \dots, Y_{n_i};\ t_0, t_1, \dots, t_{n_i}).$$

To simplify the notation we write $Y_k$ instead of $Y_{t_k}$. Then the joint probability density function for the random time and the values of the random process is equal to $p(t|Y)p(Y)$, where $p(t|Y)$ is a density function for random failure time at $t$, conditional on the observed process $Y$. The independence of the individual trajectories means that the likelihood function can be defined as

$$\mathcal{L}(\theta, \mathcal{Y}, \beta) = \prod_{i=1}^{N} p(T_i|Y^i)p(Y^i), \tag{15}$$

where $\mathcal{Y} = (Y^i, i = 1, \dots, N\}$. The functions $p(T|Y)$ and $p(Y)$ are defined by

**Proposition 3.1** (Yashin et al 1986 [13]). *Let the $m$-dimensinal process $\{Y_t, t \geq 0\}$ satisfy a linear stochastic differential equation (4). The failure rate is assumed to be a quadratic function*

$$\mu(t, Y_t) = \mu_0(t) + Y_t^T Q Y_t.$$

*Let the matrices $Q, \Gamma$ be a diagonal, conditional survival function*

$$s(t, Y_0^t) = \mathbf{P}\{T > t|Y_0^t\}$$

*and the failure rate conditional on trajectory $Y^i$ be*

$$\lambda_i(t) = -\frac{\partial}{\partial t} \log s(t, Y^i). \tag{16}$$

*Then the functions $\lambda_i(t)$ and $p(Y_t^i)$ can be represented as*

$$\lambda_i(t) = \mu_0(t) + \sum_{j=1}^{m} q_j(\gamma^{jj}(t) + m^j(t|Y_k^j)^2), \tag{17}$$

$$p(Y_t^i) = (2\pi)^{-\frac{m}{2}} \prod_{k=1}^{n_i} \left( \prod_{j=1}^{m} \gamma^{jj}(t_k^-) \right)^{-\frac{1}{2}} \exp\left( -\sum_{j=1}^{m} \frac{(Y_k^i - m^j(t_k^-|Y_{k-1}^i))^2}{\gamma^{jj}(t_k^-)} \right). \tag{18}$$

*where $m(t|Y_k^i)$ and $\gamma(t)$ are sectionally continuous functions on the intervals $t_k \leq t < t_{k+1}$ satisfies the equations*

$$\frac{d}{dt} m(t|Y_k^i) = a_1(t) m(t|Y_k^i) - 2\gamma(t) Q m(t|Y_k^i), \tag{19}$$

$$\frac{d}{dt}\gamma(t) = a_1(t)\gamma(t) + \gamma(t)a_1^T(t) + b(t)\Gamma b^T(t) - 2\gamma(t)Q\gamma(t)$$

*with the initial values for $t_k, k = 1, \dots, n_i$.*

$$m(t_k|Y_k^i) = Y_k^i, \gamma(t_k) = 0.$$

□

For discrete observations we can write

$$p(Y^i) = \prod_{\{k:t_k<T_i\}} \mathcal{N}(Y_k^i | m(t_k^- | Y_{k-1}^i); \gamma(t_k^-)), \tag{20}$$

where $\mathcal{N}(Y_k^i | m(t_k^- | Y_{k-1}^i); \gamma(t_k^-))$ is the conditional Gaussian density with the means $m(t_k^- | Y_{k-1}^i)$ and variances $\gamma(t_k^-)$ (see (19)).

Therefore the loglikelihood function has the form

$$\begin{aligned} L(\theta, \mathcal{Y}, \beta) &= \sum_{\{i:T_i \in \Theta\}} \Big( \log p(T_i | \mathcal{Y}^i; \beta) \\ &- \sum_{k=1}^{n_i} \sum_{j=1}^{m} \left( \log \gamma^{jj}(t_k^-; \beta) - \frac{(Y_k^j - m^j(t_k^- | Y_{k-1}^i; \beta))^2}{\gamma^{jj}(t_k^-; \beta)} \right) \Big). \end{aligned} \tag{21}$$

Assume that there exists vector $\beta^* \in B$ such that a.e.

$$p(t; \beta^*) = p(t) \text{ and } p(\mathcal{Y}; \beta^*) = p(\mathcal{Y}).$$

According to the maximum likelihood principle the desired estimate is a solution of the problem

$$\max\{L(\theta, \mathcal{Y}, \beta), \beta \in B\}.$$

**Proposition 3.2** (Sobolev, 1989 [11]. *The solution $\hat{\beta}_N$ of the system*

$$\frac{1}{N} \Delta_\beta L(\theta, \mathcal{Y}, \beta) = 0 \tag{22}$$

*with the probability one, converges to vector $\beta^*$ as $N \to \infty$.* □

**Proof.** For $r = 1, \ldots, l$.

$$\begin{aligned} \varphi^r(\beta, 1/N) &= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \beta^r} \log p(T_i | \mathcal{Y}^i; \beta) \\ &+ \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{n_i} \sum_{j=1}^{m} a^{rj}(t_k^- | Y_{k-1}^{ij}; \beta) \frac{Y_k^{ij} - m^j(t_k^- | Y_{k-1}^{ij}; \beta))}{\gamma^j(t_k^-; \beta)} \\ &+ \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{n_i} \sum_{j=1}^{m} \frac{c^{rj}(t_k^-; \beta)}{(\gamma^{jj}(t_k^-; \beta))^2} (\gamma^j(t_k^-; \beta) - (Y_k^{ij} - m^j(t_k^- | Y_{k-1}^{ij}; \beta))^2) = 0. \end{aligned} \tag{23}$$

Let $\theta_1$ be the set of observation times, which are identical for each $i$. For all $t_k \in \theta_1$ consider the random sums, for simplification the index $j$ is suppressed,

$$\frac{1}{N} \sum_{i=1}^{N} a^r(t_k^- | Y_{k-1}^i; \beta) \frac{(Y_k^i - m(t_k^- | Y_{k-1}^i; \beta))}{\gamma(t_k^-; \beta)}. \tag{24}$$

From Proposition 3.1, on interval $[t_{k-1}, t_k)$ the random values $m(t_k^- | Y_{k-1}^i; \beta)$ satisfy the linear equation

$$\frac{d}{dt} m(t | Y_{k-1}^i; \beta) = -A(t) m(t | Y_{k-1}^i; \beta), \tag{25}$$

with $m(t_{k-1} | Y_{k-1}^i) = Y_{k-1}^i$. Consequently, we can determine the fundamental matrix $R(t, \tau)$,

$$\frac{d}{dt} R(t, \tau) = -A(t) R(t, \tau), \quad R(\tau, \tau) = I. \tag{26}$$

The unique solution of (25) can be represented by means of the matrix $R$ as

$$m(t|Y^i_{k-1};\beta) = R(t,t_{k-1})Y^i_{k-1}. \tag{27}$$

Then, for each $t \in [t_{k-1}, t_k)$ and the random value

$$a^r(t|Y^i_{k-1};\beta) = \frac{\partial}{\partial\beta^r} m(t|Y^i_{k-1};\beta)$$

we have

$$\frac{d}{dt}a^r(t|Y^i_{k-1};\beta) = -A(t)a^r(t|Y^i_{k-1};\beta) + \frac{\partial}{\partial\beta^r}A(t)m(t|Y^i_{k-1};\beta), \tag{28}$$

with $a^r(t_{k-1};\beta) = 0$. The solution of the equation (28) on interval $[t_{k-1}, t_k)$ can be represented as

$$a^r(t|Y^i_{k-1};\beta) = \int_{t_{k-1}}^{t} R(t,\tau)B^r(\tau)m(\tau|Y^i_{k-1};\beta)d\tau = g(t,t_{k-1})Y^i_{k-1},$$

where $B^r(t) = \frac{\partial}{\partial\beta^r}A(t)$. Then for each $i$ the expression

$$a^r(t_k^-|Y^i_{k-1};\beta)\frac{(Y^i_k - m(t_k^-|Y^i_{k-1};\beta))}{\gamma(t_k^-;\beta)} =$$

$$\frac{g(t_k,t_{k-1})}{\gamma(t_k^-;\beta)}(R(t_k,t_{k-1};\beta)(Y^i_{k-1})^2 - Y^i_{k-1}Y^i_k).$$

is valid. According to strong law large numbers [1] with probability one the sum in (28) converges to

$$\frac{g(t_k,t_{k-1})}{\gamma(t_k^-;\beta)}(R(t_k,t_{k-1};\beta)E(Y_{k-1})^2 - E(Y_{k-1}Y_k)). \tag{29}$$

For the random process with linear dynamics of the means (25) we can write the differential equation

$$\frac{d}{dt}E(Y_{k-1}Y_k) = -A(t)E(Y_{k-1}Y_k). \tag{30}$$

with initial value $E(Y_{k-1})^2$. The solution of this equation can be represented as

$$R(t,t_{k-1};\beta_*)E(Y_{k-1})^2.$$

Consequently, the expression (29) can be transformed into

$$\frac{g(t_k,t_{k-1})}{\gamma(t_k^-;\beta)}(R(t_k,t_{k-1};\beta) - R(t_k,t_{k-1};\beta_*))E(Y_{k-1})^2. \tag{31}$$

Again for each $t_k$ from $\theta_1$ consider the random sums

$$\frac{1}{N}\sum_{i=1}^{N}\frac{c^r(t_k^-;\beta)}{(\gamma(t_k^-;\beta))^2}(\gamma(t_k^-;\beta) - (Y^i_k - m(t_k^-|Y^i_{k-1};\beta))^2). \tag{32}$$

The expression in parantheses can be transformed into

$$\gamma(t_k^-;\beta) - (Y^i_k - m(t_k^-|Y^i_{k-1};\beta^*) + m(t_k^-|Y^i_{k-1};\beta^*) - m(t_k^-|Y^i_{k-1};\beta))^2 =$$

$$= (\gamma(t_k^-;\beta) - (Y^i_k - m(t_k^-|Y^i_{k-1};\beta^*))^2)$$

$$-(R(t_k,t_{k-1};\beta)-R(t_k,t_{k-1};\beta^*))^2(Y^i_{k-1})^2+$$
$$+2(R(t_k,t_{k-1};\beta)-R(t_k,t_{k-1};\beta^*))(Y^i_kY^i_{k-1}-R(t_k,t_{k-1};\beta^*)(Y^i_{k-1})^2).$$

The same consideration leads to the proposition that with probability one the sum (32) has a limit

$$(R(t_k,t_{k-1};\beta)-R(t_k,t_{k-1};\beta^*))^2E(Y_{k-1})^2-\gamma(\bar{_k};\beta^*)+\gamma(t_k^-;\beta) \tag{33}$$

with $N\to\infty$ as an accuracy to some constant.

According to strong law large numbers [1] with probability one

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta^r}\log f(T_i|\mathcal{Y}^i;\beta)\to E_*[\frac{\partial}{\partial\beta^r}\log f(\beta)] \text{ as } N\to\infty. \tag{34}$$

From (31), (33), (34) it follows that with the probability one for each $r$ the equations

$$\frac{1}{N}\nabla_\beta L(\theta,\mathcal{Y},\beta)=0$$

has the limit

$$\varphi^r(\beta,0)=E_*[\frac{\partial}{\partial\beta^r}\log p(\beta)]$$
$$+\sum_k c_1(R(t_k,t_{k-1};\beta)-R(t_k,t_{k-1};\beta^*))$$
$$+\sum_k c_2(R(t_k,t_{k-1};\beta)-R(t_k,t_{k-1};\beta^*))^2$$
$$+\sum_k c(\gamma(t_k^-;\beta)-\gamma(t_k^-;\beta^*))=0.$$

In Section 3 for unconditional probability density of the failure time we have seen that

$$E_*[\frac{\partial}{\partial\beta^r}\log p(\beta)]=0.$$

because vector $\beta^*$ is a solution of the nonlinear system. Consequently, with probability one, the estimate $\hat{\beta}_N$ which satisfies the likelihood equation converges to $\beta^*$ as $N\to\infty$.

The sufficient condition for the uniqueness of this solution is non-singularity of the matrix $\phi(\beta,\frac{1}{N})=\Delta_\beta[\frac{1}{N}\Delta_\beta L(\theta,\mathcal{Y},\beta)]$ at $(\beta_*,0)$. This condition was given by Zuev [9].

# 4 Numerical Algorithms

Consider the problem

$$\max_{\beta\in B}L(\beta),\quad L:\mathbf{R}^1\to\mathbf{R}^1 \tag{35}$$

with the nonlinear equation according to $\beta$

$$U(\beta)=0, U:\mathbf{R}^l\to\mathbf{R}^l, \text{ where } U(\beta)\equiv\nabla L(\beta)=\partial L/\partial\beta^1,\ldots,\partial L/\partial\beta^l)^T. \tag{36}$$

Let $U:\mathbf{R}^l\to\mathbf{R}^l$ be a continuously differentiable function. The problem is to find a vector $\hat{\beta}\in B\subset\mathbf{R}^l$ such that $U(\hat{\beta})=0$ holds. For the solution of the nonlinear problem

(36) Newton's method and the approximation of the function $U$ in a neighborhood of an initial guess $\beta_0$ can be used [3]

$$U(\beta_0 + \delta\beta) = U(\beta_0) + \int_{\beta_0}^{\beta_0+\delta\beta} I(z)dz.$$

Here $I = \nabla U$ is a Jacobian of the system. Consider the linear approximation for $U(\beta_0+\delta\beta)$ with respect to increment $\delta\beta$

$$M(\beta_0 + \delta\beta) = U(\beta_0) + I(\beta_0)\delta\beta.$$

Letter $M$ denotes the approximation model. Then the correction $\delta\beta$ can be searched from the condition

$$M(\beta_0 + \delta\beta_0) = 0.$$

Consequently, for fixed $\beta_0$, $\delta\beta_0$ is the solution of the linear algebraic system

$$I(\beta_0)\delta\beta_0 = -U(\beta_0).$$

As a result we obtain the value $\delta\beta_0$ which is different from the actual $\delta\beta$. Therefore the vector

$$\beta_1 = \beta_0 + \delta\beta_0$$

can be selected as the next guess. Formally we have the following iterative

**Scheme 1.**

1. Set $\beta_0$ - an initial guess for $\hat{\beta}$.
2. For $n = 0, 1, 2, \ldots$ until convergence do:
   Solve the system

$$I(\beta_n)\delta\beta_n = -U(\beta_n) \tag{37}$$

   Set next guess $\beta_{n+1} = \beta_n + \delta\beta_n$.

It is known that Newton's iterative process has local quadratic convergence [3]. It means there exists scalar $\varepsilon > 0$ such that for all $\beta_0$ from open neighborhood of $\hat{\beta}$, with radius $\varepsilon$, the consequence $\{\beta_n, n = 0, 1, \ldots\}$ is defined correctly, converges to $\hat{\beta}$ as $n \to \infty$ and satisfies the inequality

$$\|\beta_{n+1} - \hat{\beta}\| \leq C\|\beta_n - \hat{\beta}\|^2, \quad n = 0, 1, 2, \ldots. \tag{38}$$

According to model (8) the likelihood function depends on parameter $\beta$ by an implicit manner

$$L(\beta) \equiv L(x(\beta)),$$

where $x(\beta)$ is a solution of ODE

$$\frac{d}{dt}x(t) = f(x(t), \beta), \quad x(0) = x_0, \tag{39}$$

$$t \geq 0, \quad \beta \in B \subset \mathbf{R}^l.$$

The problem

$$\max_{\beta \in B} L(x(\beta)) \tag{40}$$

is more sophisticated because it is necesary to calculate the functions $x(t,\beta), \partial x(t,\beta)/\partial\beta$ and $\partial^2 x(t,\beta)/\partial\beta^2, t \geq 0$ for each step of the iterative scheme. To overcome this obstacle and design an efective numerical algorithm the implicit function $\delta\beta(\beta) : \mathbf{R}^l \to \mathbf{R}^l$ in the neighborhood of $\beta$ can be used.

## 4.1 Iterative scheme

Consider the Newton-like scheme. Introduce a new fucntion close to $L(x(\beta))$ with the following properties:

(a) this function depends on $\delta\beta$ explicitly.

(b) it guarantees the quadratic convergence of scheme 1.

It should be mentioned that the $\delta\beta$ is not Newton's step for the problem (40). We have to modify the scheme so that the Newton-like iteration has the quadratic convergence to $\hat{\beta}$

$$\hat{\beta} = \arg\max_{\beta \in B} L(x(\beta)). \tag{41}$$

Assume that the solution of the problem (40), $\hat{\beta}$, can be represented as

$$\hat{\beta} = \beta_0 + \delta\beta, \tag{42}$$

where $\beta_0$ is known and $\delta\beta \in S_0 \subset \mathbf{R}^l$ is small in comparison with $\beta_0, S_0$ is the zero neighborhood to $\mathbf{R}^l$. Assumption about the smallness $\delta\beta$ can be described by

$$\delta\beta = \varepsilon\Delta\beta, \Delta\beta \in \mathbf{R}^l, \tag{43}$$

$\varepsilon > 0$ is a small parameter.

Supposing the required smoothness we can represent the solution of the system (8) in any neighborhood $\beta_0$ as a linear combination

$$\begin{aligned} \Lambda(t,\hat{\beta}) &= \Lambda(t,\beta_0) + b(t,\beta_0)\delta\beta, \\ m(t,\hat{\beta}) &= m(t,\beta_0) + a(t,\beta_0)\delta\beta, \\ \gamma(t,\hat{\beta}) &= \gamma(t,\beta_0) + c(t,\beta_0)\delta\beta, \end{aligned} \tag{44}$$

where

$$\begin{aligned} b &\equiv \nabla_\beta \Lambda, \\ a &\equiv \nabla_\beta m, \\ c &\equiv \nabla_\beta \gamma. \end{aligned} \tag{45}$$

The function $\lambda$ evaluated in $\hat{\beta}$ for all $t$ we can write in terms of a series inpower of a small scalar $\varepsilon \geq 0$

$$\lambda(t,\hat{\beta}) = \lambda(t,\beta_0) + \varepsilon\varphi(t,\beta_0)\Delta\beta + \ldots \tag{46}$$

where $\varphi$ is equal to

$$\varphi = \nabla_m \lambda^T a + \nabla_\gamma \lambda^T c + \nabla_\beta \lambda. \tag{47}$$

For residual terms for $t \geq 0$ the estimate $|R(t,\varepsilon)| \leq C\varepsilon^2$ holds. Constant $C$ does not depend on $t$ and $\varepsilon$ [18].

Note that maximization of function $L(\beta)$ equals the problem of the minimize searching for the function $-L(\beta)$. Now, consider the function

$$L(\beta) = \sum_{t\in\Theta}(\Lambda(t,\beta) - \log\lambda(t,\beta)) \tag{48}$$

and the corresponding nonlinear system

$$\nabla L(\beta) = 0 \tag{49}$$

For the fixed vector $\beta_0$ the problem is

$$\min l(\beta_0, \delta\beta), \tag{50}$$

where the function $g : \mathbf{R}^l$ to $\mathbf{R}$ has the form

$$g(\beta_0, \delta\beta) = \sum_{t\in\Theta}(\Lambda(t,\beta_0) + b(t,\beta_0)\delta\beta - \log[\lambda(t,\beta_0) + \varphi(t,\beta_0)\delta\beta]). \tag{51}$$

This function depends on $\delta\beta$ explicitly. Note that this process is not connected with the solving of the ODE. For fixed $\beta_0$ and $t$ the functions $a, b, c$ and $\lambda$ are known constants which were evaluated before the iterative process.

We can show that for given $\beta$ the function

$$g(x) \equiv g(\beta, x), \quad l : \mathbf{R}^l \to \mathbf{R}^l$$

is convex. Here for simplicity letter $x$ is used for $\delta\beta$. Indeed, let $s_t$ be a vector with elements

$$s_t^i = \frac{\varphi^t(t,\beta)}{\lambda(t,\beta) + \varphi(t,\beta)x}, \quad i = 1,\ldots,l, \quad t \in \Theta; \tag{52}$$

and $U(x)$ be a vector with elements

$$U^j(x) = \sum_{t\in\Theta}(b^j(t,\beta) - s_t^j(x)), \quad j = 1,\ldots,l. \tag{53}$$

Consider the nonlinear system

$$U(x) = 0. \tag{54}$$

It is simple to show that the equality

$$I = \sum_{t\in\Theta} s_t s_t^T \tag{55}$$

holds, where $I = \nabla U$ is a Jacobian. So the matrix $I$ is symmetric and nonnegative defined. Because the function $g(x)$ is convex and the necessary extreme condition (55) is a sufficient condition too.

Again, we can use the Newton method to solve the nonlinear system (54).

**Scheme 2.**

1. Set $x_0$ - initial values for given $\beta_0$

2. For $k = 0, 1, 2, \ldots$ until convergence do

$$\text{Compute } x_{k+1} = x_k - [I(\beta_0, x_k)]^{-1} U(\beta_0, x_k). \tag{56}$$

Set $\delta\beta_0 = x_\infty$.

The convexity of the function $l(x)$ implies the quadratic convergence of the sequence $\{x_k, k = 0, 1, 2, \ldots\}$ to the actual root of the minimization problem

$$x_\infty = \arg\min g(\beta_0, x). \tag{57}$$

The linearization of function $L$ gives us a value $\delta\beta_0$ different from the real vector $\delta\beta = \hat{\beta} - \beta_0$. We can replace a vector $\beta_0$ by the vector $\beta_1 = \beta_0 + \delta\beta_0$ and use it as the next approximation for the unknown vector $\hat{\beta}$.

Following the iterative scheme, we can propose to search the desired vector $\hat{\beta}$

**Scheme 3.**

1. Set $\beta_0$ - initial guess for parameter vector.

2. For $n = 0, 1, 2, \ldots$ do

   (a) Set $x_0$ - initial values for given $\beta_n$.

   (b) For $k = 0, 1, 2, \ldots$ until convergence do
   Calculate $x_{x+1} = x_k - [I(\beta_n, x_k)]^{-1} U(\beta_n, x_k)$.
   Set $\delta\beta_u = x_\infty$.

   (c) Calculate

$$\beta_{n+1} = \beta_n + \delta\beta_n. \tag{58}$$

Now we can formulate the following

**Proposition 4.1** *[11]. Let the iterative process (58) converge. Then its limit vector $\beta_\infty$ satisfies the solution of problem (40).* □

Examine Scheme 3. It is simple to show that this is a method of simple iteration. In fact, we can present the iteration in the form

$$\beta_{n+1} = G(\beta_n), \quad n = 0, 1, 2, \ldots \tag{59}$$

where function $G(\beta)$ is defined by the equality

$$G(\beta) = \beta + \delta\beta(\beta) \tag{60}$$

and the function $\delta\beta(\beta)$ is a solution of problem (50), i.e. it satisfies

$$\delta\beta(\beta) = \arg\min g(\beta, x) \tag{61}$$

The classical theorem of numerical mathematics for arbitrary iterative methods in terms of $\beta_{n+1} = G(\beta_n)$ determines the condition with respect to the function $G$ that implies the linear convergence of sequence

$$\{\beta_n, n = 0, 1, \ldots\}$$

to the desired vector $\hat{\beta}$.

**Proposition 4.2** *Let $G : B \to B$ where $B$ is a closed subset on $\mathbf{R}^1$. If for any norm there exists scalar $\alpha \in [0,1)$ the inequaity*

$$\|G(\beta\prime) - G(\beta)\| \leq \alpha\|\beta' - \beta\|$$

$$\forall \beta, \beta' \in B$$

*is fulfilled. Then*

*a) there exists the unique $\hat{\beta}$ such that $G(\hat{\beta}) = \hat{\beta}$,*

*b) for all $\beta_0 \in B$ the consequence*

$$\beta_{n+1} = G(\beta_n), \quad n = 0, 1, 2, \ldots$$

*linearly converges to $\hat{\beta}$,*

*c) for all $\nu \geq \|G(\beta_0) - \beta_0\|$ the evaluation*

$$\|\beta_n - \hat{\beta}\| \leq \nu \frac{\alpha^n}{1-\alpha}, \quad n = 0, 1, 2, \ldots$$

*is valid.* □

Zuev et al. [10] found that the inequality $\|\nabla G(\hat{\beta})\| < 1$ is the condition of convergence for (58). This theorem can be used to define only the condition of linear convergence. We are interested in the more speedy method. Due to this iteration scheme the non-newton step toward the desired $\hat{\beta}$ is obtained. The idea to use the implicit function to examine the convergence of Scheme 3 is found fruitful. Next we show how the iterative scheme with quadratic convergence can be obtained.

The nonlinear system which defines the extreme condition of the function $g(\beta, x)$

$$U(\beta, x) = 0, \text{ where } U(\beta, x) \equiv \nabla_x g(\beta, x), \tag{62}$$

sets the map

$$U : \mathbf{R}^l \times \mathbf{R}^l \to \mathbf{R}^l$$

which is continuous in a neighborhood $D_0 \subset \mathbf{R}^l \times \mathbf{R}^l$ of the point $(\hat{\beta}, 0)$. This point satisfies the system (62) $U(\hat{\beta}, 0) = 0$.

The solutions of the system (8) are continuous. Therefore the map $U$ is continuously differentiable in respect to $x$ in any neighborhood $D_1 \subset \mathbf{R}^l \times \mathbf{R}^l$ of the point $(\hat{\beta}, 0)$ and matrix $\nabla_x U(\beta, x)$ is non-singular in this point.

The implicit function theorem guarantees that (a) there exists a neighborhood $S_1 \subset \mathbf{R}^1$ of the point $\hat{\beta}$, in which (b) the system of continuous functions

$$x = F(\beta), F : \mathbf{R}^l \to \mathbf{R}^l, \tag{63}$$

is uniquely defined in $S_1$, (c) for all $\beta \in S_1$ the function $F(\beta)$ satisfies the system

$$U(\beta, F(\beta)) = 0$$

and it is the unique solution in $S_1$, (d) the function $F$ has the continuous derivative

$$\nabla F(\beta) = -[\nabla_x U(\beta, x)]^{-1} \nabla_\beta U(\beta, x). \tag{64}$$

Assume that in the desired point $(\hat{\beta}, 0)$ the equality

$$F(\hat{\beta}) = 0$$

holds, the problem can be formulated as a solution of the nonlinear system

$$F(\beta) = 0, F : \mathbf{R}^l \to \mathbf{R}^l, \tag{65}$$

where function $F$ is given in $S_1$ implicitly by system

$$U(\beta, F(\beta)) = 0, \quad U : \mathbf{R}^l \times \mathbf{R}^l \to \mathbf{R}^l. \tag{66}$$

Newton's method for (66) can be written in the form

1. Set $\beta_0$ - an initial guess for $\hat{\beta}$.
2. For $n = 0, 1, 2, \ldots$ until convergence do:
   Solve the system $J(\beta_n)\delta\beta_n = -F(\beta_n)$
   Set the next guess
   $$\beta_{n+1} = \beta_n + \delta\beta_n \tag{67}$$
   where $J(\beta_n) = \nabla F(\beta_n)$ is Jacobian of the system evaluated in $\beta_n$. According to the implicit function theorem it exists and is continuous in $S_1$.

We cannot calculate the values $F(\beta_n)$ immediately as a function of $\beta$. However, for all $\beta \in S_1$ we can use the values

$$\delta\beta_n \equiv F(\beta_n)$$

which is the minimizer of the function $l(\beta_n, x)$, i.e., for given $\beta_n$ as $F(\beta_n)$ we will use a solution of system $U(\beta_n, \delta\beta) = 0$.

The Newton-like iterative scheme is required.

**Scheme 4.**

1. Set $\beta_0$ - initial guess for parameter vector.
2. For $n = 0, 1, 2, \ldots$ do
   (a) Set $x_0$ - initial values for given $\beta_n$.
   (b) For $k = 0, 1, 2, \ldots$ until convergence do
   Calculate $x_{k+1} = x_k - [I(\beta_n, x_k)]^{-1} U(\beta_n, x_k)$.
   Set $F_n = x_\infty$
   (c) For $j = 1, \ldots, l$ calculate
   $x_{k+1} = x_k - [I(\beta_n + F_n e_j, x_k)]^{-1} U(\beta_n + F_n e_j, x_k)$.
   $e_j$ is $l \times 1$ vector with elements $e_j^k = \begin{cases} 0, & k \neq j \\ 1, & k = j. \end{cases}$
   Set $F_{jn} = x_\infty$
   (d) Calculate $A_n^{ij} = \frac{F_{jn}^i - F_n^i}{F_n^j}$ $\quad i, j = 1, \ldots, l$.
   (e) Set $A_n = \|A_n^{ij}\|_{l \times l}$
   (f) Solve $A_n \delta\beta_n = F_n$
   (g) Calculate $\beta_{n+1} = \beta_n + \delta\beta_n$.

## 4.2 Local convergence analysis

In this section the local quadratic converges for the iterative procedure 4 is proved, the truth of the inequality

$$\|\beta_{n+1} - \hat{\beta}\| \leq C\|\beta_n - \hat{\beta}\|^2, \quad n = 0, 1, 2, \ldots. \tag{68}$$

is shown. For the proof we use the same approach as for Newton's method [5].

Let $\|\cdot\|$ be the vector $l_1$-norm and the induced matrix norm on $\mathbf{R}^l$ [2]. Let $S_\varepsilon(x)$ be the open domain with radius $\varepsilon > 0$ and with center in $x$, i.e.

$$S_\varepsilon(x) = \{y \in \mathbf{R}^l : \|y - x\| < \varepsilon\}.$$

The main result is given in

**Proposition 4.3** (Sobolev, 1989 [11]). *Let the function*

$$F : \mathbf{R}^l \to \mathbf{R}^l \tag{69}$$

*be given in an open convex set $D$ implicitly by the system*

$$U(\beta, F(\beta)) = 0, \quad U : \mathbf{R}^l \times \mathbf{R}^l \to \mathbf{R}^l. \tag{70}$$

*Assume there exists the vector $\hat{\beta} \in B \subset \mathbf{R}^l$ such that equality $F(\hat{\beta}) = 0$ holds. Jacobian $J$ of function $F$ evaluated in $\hat{\beta}$ is nonsingular and the equality*

$$\|J(\hat{\beta})\|^{-1} < r < \infty \tag{71}$$

*holds.*

*Then there exists a small scalar $\varepsilon > 0$ such that for all $\beta_0 \in S_\varepsilon(\hat{\beta})$ the sequence derived by the iterative procedure 4 $\{\beta_n, n = 0, 1, 2, \ldots\}$ is defined correctly and converges to the desired $\hat{\beta}$ quadratically.*

**Proof.** First, let us show that the iterative process is defined correctly and the evaluation

$$\|\beta_{n+1} - \hat{\beta}\| \leq \frac{1}{2}\|\beta_n - \hat{\beta}\|$$

holds for all $n$, i.e., $\beta_{n+1} \in S_\varepsilon(\hat{\beta}) \quad n = 1, 2, \ldots$.

Consider the iterative scheme 4 for $k = 0$

$$\beta_1 = \beta_0 - [A_0]^{-1}F_0.$$

Vector $\beta_1$ is defined correctly if matrix $A_0$ is nonsingular. If any matrix $B$ is nonsingular and for any other matrix $A$ the inequality

$$\|B^{-1}(B - A)\| < 1$$

holds then the matrix $A$ is nonsingular too and the evaluation

$$\|A^{-1}\| \leq \frac{\|B^{-1}\|}{\|B^{-1}(B - A)\|}$$

is valid [5].

Consider matrix $A_0$ as $A$ and matrix $J_* = J(\hat{\beta})$ as $B$ as above. Then the following evaluation

$$\|J_*^{-1}(J_* - A_0)\| \leq \|J_*^{-1}\|\|(J_* - J_0 + J_0 - A_0)\| \leq r(\|J_* - J_0\| + \|J_0 - A_0\|)$$

holds. The theorem about high-order derivatives for implicit function guarantees that there exist continuous derivatives for function $F$ at least second order with appreciable smoothness of function $U$. Therefore function $F$ and its Jacobian $J$ are Lipschitz continuous in $D$, i.e. there are constants $p, q$ such that $\forall \beta, \beta' \in D$

$$\|J(\beta') - J(\beta\| \leq p\|\beta' - \beta\| \tag{72}$$

$$\|F(\beta') - F(\beta\| \leq q\|\beta' - \beta\| \tag{73}$$

are valid.

Consider the norm of difference

$$\begin{aligned} \|A_0 - J_0\| &= \left\| \left[ \frac{F_{j0}^i - F_0^i}{F_0^j} - J_0^{ij} \right]_{l\times l} \right\| \\ &= \left\| \left[ \frac{F^i(\beta_0 + F_0 e_j) - F^i(\beta_0) - J_0^{ij} F^j(\beta_0)}{F^j(\beta_0)} \right]_{l\times l} \right\|. \end{aligned} \tag{74}$$

The Jacobian definition and Newton's theorem implies immediately for all vectors $x$ and $x + y$ the equality

$$\begin{aligned} F(x+y) - F(x) - J(x)y &= \int_0^1 J(x+ty)y\,dt - J(x)y \\ &= \left( \int_0^1 (J(x+ty) - J(x))dt \right) y \end{aligned}$$

where it means the element-by-element integration of the matrix-valued function. Consequently, the elements of matrix $J_0 - A_0$ are evaluated as

$$= \int_0^1 (J^{ij}(\beta_0 + tF_0 e_j) - J_0^{ij})dt$$

Then (74) is transformed into the form

$$\int_0^1 \left\| \left[ J^{ij}(\beta_0 + tF_0 e_j) - J_0^{ij} \right]_{l\times l} \right\| dt.$$

Using (72) and (73) we have

$$\leq 1/2p\|F_0\| \leq 1/2p\|F_0 - F(\hat{\beta})\| \leq 1/2pq\|\beta_0 - \hat{\beta}\|. \tag{75}$$

Therefore for any

$$\varepsilon < \frac{1}{2rp(q+2)}$$

we obtain

$$\begin{aligned} \|J_*^{-1}(J_* - A_0)\| &\leq r(\|J_* - J_0\| + \|J_0 - A_0\|) \\ &\leq r(p\|\beta_0 - \hat{\beta}\| + 1/2pq\|\beta_0 - \hat{\beta}\|) \\ &< rp(1 + 1/2q)\varepsilon < 1/2 < 1. \end{aligned}$$

So the matrix $A_0$ is nonsingular, the inequality

$$\|A_0^{-1}\| \leq 2r$$

holds and $\beta_1$ is defined correctly. Now let us show that $\beta_1$ belongs to $S_\varepsilon(\hat{\beta})$. The equality

$$F(\hat{\beta}) = 0$$

implies the other equality

$$\beta_1 - \hat{\beta} = \beta_0 - \hat{\beta} - [A_0]^{-1}(F_0 - F(\hat{\beta})). \tag{76}$$

Let $\delta_n$ be the notation for the difference $\beta_n - \hat{\beta}$ and $F_*$ be the evaluation of the function $F$ at $\hat{\beta}$. Rewrite the last equality in the form

$$A_0\delta_1 = A_0\delta_0 - F_0 + F_*.$$

Adding the vector $J_0\delta_0$, where $J_0 = J(\beta_0)$, we receive

$$A_0\delta_1 = (A_0 - J_0)\delta_0 + (-F_0 + F_* + J_0\delta_0).$$

Then

$$\delta_1 = [A_0]^{-1}(A_0 - J_0)\delta_0 + [A_0]^{-1}(-F_0 + F_* + J_0\delta_0).$$

Using (75) we have

$$\|\delta_1\| < 2r(1/2pq\varepsilon\|\delta_0\| + \|(-F_0 + F_* + J_0\delta_0)\|).$$

Taking into account that Jacobian of the function $F$ is Lipschitz continuous in $D$ with (73) we have

$$\|(-F_0 + F_* + J_0\delta_0)\| < 1/2p\|\delta_0\|^2 < 1/2p\varepsilon\|\delta_0\|.$$

Then for the norm $\|\delta_1\|$ we can write

$$\|\delta_1\| < 2r(1/2pq\varepsilon + 1/2p\varepsilon)\|\delta_0\|$$

$$\leq pr(q+1)\varepsilon\|\delta_0\| \leq 1/2\frac{(q+1)}{(q+2)}\|\delta_0\| < 1/2\|\delta_0\|.$$

it means that $\beta_1$ belongs to $S_\varepsilon(\hat{\beta})$.

The inductive consideration implies that matrix $A_n$ is nonsingular, the evaluations

$$\|A_n^{-1}\| \leq 2r$$

$$\|\delta_{n+1}\| \leq 1/2\|\delta_n\|$$

for $n = 1, 2, \ldots$ are valid. It means the convergence of the sequence

$$\{\beta_n, n = 0, 1, 2, \ldots\} \tag{77}$$

to vector $\hat{\beta}$. Consequently the sequence

$$\{F(\beta_n), n = 0, 1, 2, \ldots\}$$

converges to zero.

To prove the quadratic convergence it is sufficient to show that the sequence (77) satisfies the inequality [6]

$$\|\delta_{n+1}\| \leq C\|\delta_n\|^2.$$

For $n > 0$ rewrite (76) in the form

$$A_n\delta_{n+1} = (A_n - J_n)\delta_n + (-F_n + F_* + J_n\delta_n).$$

As far as for $n$ the matrix $A_n$ is nonsingular and the expression

$$\delta_{n+1} = [A_n]^{-1}(A_n - J_n)\delta_n + [A_n]^{-1}(-F_n + F_* + J_n\delta_n)$$

is defined correctly. It implies the evaluation

$$\|\delta_{n+1}\| \leq 2r(\|A_n - J_n\|\|\delta_n\| + \|(-F_n + F_* + J_n\delta_n)\|).$$

With (75) the last inequality can be transformed to

$$\|\delta_{n+1}\| \leq 2r(1/2pq\|\delta_n\|^2 + \|(-F_n + F_* + J_n\delta_n)\|).$$

The implicit function $F$ is continuously differentiable. This implies the evaluation

$$\|(-F_n + F_* + J_n\delta_n)\| \leq 1/2p\|\delta_n\|^2.$$

As a result we receive the inequality

$$\|\delta_{n+1}\| \leq C\|\delta_n\|^2,$$

where

$$C = 1/2rp(q+1). \quad \square$$

# References


[1] Cramer, H. *Mathematical Methods of Statistics.* Princeton, 1946.

[2] Marchuk, G.I. *Methods of Numerical Mathematics.* 2nd edn., Springer-Verlag, New York, 1982.

[3] Ortega, J.M. and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.

[4] Dennis, J.E. and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, New Jersey, 1984.

[5] Brown, P.E. and Y. Saad. Hybrid Krylov method for nonlinear systems of equations, *SIAM J. Sci. Stat. Comput.*, **11**, 3 (1990).

[6] Schwartz, L. *Analysis*, V. 1, Mir Publishing House, Moscow, 1974.

[7] Pitman, E.J.G. *Some Basic Theory of Statistical Inference.* Chapman and Hall, London, 1979.

[8] Cox, D.R. and D. Oakes. *Analysis of Survival Data.* Chapman and Hall, London, 1984.

[9] Zuev, S.M. *Statistical Estimating Parameters of Mathematical Models of Disease.* Nauka, Moscow, 1988, (in Russian).

[10] Asachenkov, A., B. Sobolev and Ye. Smolianinov. Mathematical modeling and data analysis from immunological tests in oncological patients, IIASA Working Paper WP-89-32, 1989.

[11] Sobolev, B.G. Statistical estimating of coefficients in cancer disease model, *Dep. Numer. Math. Preprint*, 227 (1989), (in Russian)

[12] Yashin, A.I. Dynamics in survival analysis: Conditional Gaussian property versus the Cameron-Martin formula, *Steklov Seminar (Krylov, N.V., R.S. Liptser and A.A. Novicov, Eds.)*, pp. 467-485, Springer-Verlag, New York, 1985.

[13] Yashin, A.I., K.G. Manton and E. Stallard. Evaluating the Effects of Observed and Unobserved Diffusion Processes in Survival Analysis of Longitudinal Data, *Mathematical Modelling*, **7** (1986).

[14] Ventsel, A.D. and M.I. Freudlin. *Fluctuations in Dynamical Systems under Action of Small Random Perturbations.* Nauka, Moscow, 1979 (in Russian)

[15] Marchuk, G.I., A.L. Asachenkov, Ye.S. Smolianinov and B.G. Sobolev. On the problem of processing clinical and laboratory data on oncological patients, *Sov. J. Numer. Anal. Math. Modelling*, **4**, 5(1989).

[16] Manton, K.G. and E. Stallard. *Recent Trends in Mortality Analysis.* Academic Press, Orlando, 1984.

[17] Woodbury, M.A. and K.G. Manton. A Random-Walk Model of Human Mortality and Aging, *Theor. Popul. Biol.*, **11**(1977).

[18] Asachenkov, A.L. On one algorithm for solving inverse problems using the theory of adjoint equations, *Sov. J. Numer. Anal. Math. Modelling*, **2**, 3(1987).

[19] Yashin, A.I. and E. Arjas. A note on random intensities and conditional survival function, *J. Appl. Prob.*, **25**(1988).

[20] Asachenkov, A.L., G.I. Marchuk, R.R.Mohler, S.M. Zuev. Disease Dynamics, Birkhäuser, Boston-Basel-Berlin, (1992) forthcoming.

[21] Asachenkov, A.L. and B.G. Sobolev. Models and Data Analysis of Cancer Patients. In *Selected Topics on Mathematical Models in Immunology and Medicine* (R. Mohler and A. Asachenkov eds.), IIASA Paper, CP-90-007, 1990, pp. 61-69.

[22] Asachenkov, A.L. and R.R. Mohler. A system overview of immunology, disease and related data processing, IIASA Working Paper, WP-89-031, 1989, 29pp.