1

Application of Inverse Problems in Epidemiology and Demography

A. Michalski

Institute of control sciences, Moscow

Abstract: Different problems in epidemiology and demography can be considered as solution of inverse problem, when using observed data one estimates the process caused the data. Examples are estimation of infection rate on dynamics of the disease, estimation of mortality rate on sample of survival times, estimation of survival in wild on survival in laboratory. Specific property of inverse problem - instability of solution is discussed, procedure for stabilization is presented. Examples of morbidity estimation on incomplete data, HIV infection rate estimation on dynamics of AIDS cases and estimation of survival function in wild population on survival of captured animals are presented.

Keywords and phrases: Inverse problem, Epidemiology, Demography, Incomplete follow-up, HIV infection rate, AIDS cases dynamics, Survival in wild

1.1 Introduction

Interpretation of observations in different disciplines of life science can be considered as a solution of mathematical inverse problem. Examples are epidemiology, demography and biodemography. The important in epidemiology indicators such as prevalence of a disease and incidence of it are related by cause-effect relationship. This means that the process of a disease occurrence in formal way causes process of accumulation of the disease cases in population. The other example is relationship between rate of infection and the number of corresponding diagnosed cases. In demography cause-effect relationship exists between mortality and survival processes. Mortality process 'make influence' on survival in population.

In all these examples the value of the 'effect' can be estimated on population observations while the direct estimation of the 'cause' is impossible or

needs great funds. On the other hand information about the 'cause' often is important for better understanding of the phenomenon investigated and mathematical methods for estimation of 'cause' on 'effect' data are needed. The report describes mathematical formulations of the 'cause-effect' problem, difficulties of estimation of the 'cause process' on population data and a procedure elaborated to overwhelm them. Three examples with results of calculations are presented: estimation of morbidity on the results of incomplete follow-up, estimation of HIV infection rate on the dynamics of AIDS cases, estimation of survival in wild population on survival of captured animals in laboratory.

1.2 Mathematical formulation

Many problems from epidemiology and demography can be written as a relationship between unobserved process $\Psi(x)$ and observed process $U(\mathrm{x})$ in form

$$U(x) = A\Psi = \int_a^b K(x,t)\Psi(t)dt, \tag{1.1}$$

where A is integral operator given by a kernel function $K(x,t)$, which is defined by the nature of the problem investigated. More detail consideration for this function is given bellow. The main property of the equation (1.1) with continues kernel is that exact solution is unstable in respect to small variations in the observed function $U(\mathrm{x})$. In mathematical terms this means that there exists a sequence of functions δU such that the sequence of corresponding solutions of equation (1.1) $\delta\Psi = A^{-1}\delta U$ do not tend to zero while the sequence δU tends to zero. Such problems are called ill-posed problems by Tikhonov and Arsenin (1977). In practical applications this means that a small disturbance in observations U can lead to big disturbance in the exact solution $A^{-1}U$. Such property is well known in numerical solution of large sets of linear equations. In this case A is a matrix such that matrix A^TA has small eigen value, which means that the inverse matrix $\left(A^TA\right)^{-1}$ has large eigen value and disturbance in the solution $\Psi = \left(A^TA\right)^{-1}A^TU$ is high. Often the sensitivity of the system is so high that even machine arithmetic errors are enough to change the solution Ψ dramatically.

To obtain the stable solution for the equation (1.1) one is to put additional restrictions to the solution. Tikhonov and Arsenin (1977) proposed to put such restrictions by minimizing on Ψ a functional

$$\|U - A\Psi\|^2 + \alpha\Omega\left(\Psi\right), \tag{1.2}$$

where $\Omega(.) > 0$ is a stabilization functional, defined such that for any constant C the set $\{\Psi : \Omega(\Psi) \leq C\}$ is a compact set, α is a positive stabilization parameter. Optimal value for α depends on the level of disturbance δ in observed data U. It is proved that if $\delta^2/\alpha \to 0$ while $\delta \to 0$ and $\alpha \to 0$, then minimizer of (1.2) Ψ_α tends to the exact solution of equation (1.1). The problem of proper selection value for stabilization parameter α if level of disturbance δ does not tend to zero is still a challenging task. Different approaches and methods described in Evans and Stark (2002) including cross-validation and Bayesian approaches.

The different approach to stabilization parameter selection is based on estimate for mathematical expectation for quadratic functional value minimized on finite sample, which is described in Michalski (1987). For solution Ψ_α, which minimizes functional $\|U - A\Psi\|^2 + \alpha \|B\Psi\|^2$ for α such that $\mathrm{m} > 2TrA_\alpha$, with probability no less than 1-η the inequality is valid

$$E_{Y,U} \|Y - A\Psi_\alpha\|^2 < \frac{\|U - A\Psi_\alpha\|^2}{1 - 2TrA_\alpha/\mathrm{m}} + const + \sqrt{\frac{const}{\eta}}. \qquad (1.3)$$

Here U, Y – independent realizations of size m, generated from the same distribution, A, B – matrixes, $A_\alpha = A\left(A^TA + \alpha B^TB\right)^{-1} A^T$. The left side of (1.3) is the mean value of disagreement between possible vectors of experimental and predicted data. To get it small one can use for stabilization parameter α value, which minimizes expression $I_\alpha = \|U - A\Psi_\alpha\|^2 / (1 - 2TrA_\alpha/\mathrm{m})$. The quantity $\|U - A\Psi_\alpha\|^2$ is a square residual for empirical data. It is interesting to note that cross-validation criterion takes form $I_\alpha^{cv} = \|U - A\Psi_\alpha\|^2 / (1 - TrA_\alpha/\mathrm{m})^2$. For small amount of data it is demonstrated in Michalski (1987), that criterion I_α produces better results than criterion cross-validation I_α^{cv}.

1.3 Estimation of morbidity on the results of incomplete follow-up

In Michalski *et al.* (1996) considered a problem of estimation morbidity on data of irregular health examinations. This problem leads to solution of a matrix equation

$$A\Psi = U$$

with U - proportion of diagnosed cases among observed people by years of investigation, Ψ - probability for healthy person to become sick by years. A is a triangular matrix with 1 at the main diagonal and elements a_{ij} equal to proportion of people, examined in the year i and been healthy before, among those, who skipped the examination in the year j after the last examination.

In the case of a complete follow-up study the matrix A is the identity matrix and the morbidity estimate for different years are just the ratio between the number of cases and the number of people, examined in the same year.

Stabilization of the matrix equation was made by minimization (1.2) with stabilization functional $\Omega(p) = \|B\Psi\|^2 = \Psi^T B^T B\Psi$, B is matrix with two non zero diagonals. It holds -1 at the main diagonal and 1 at the second. This structure of stabilization functional reflects the hypothesis, that the morbidity will not change significantly in consequent years.

The described approach was applied in Michalski *et al.* (1996) for estimation of malignant neoplasm (ICD9 140-208) morbidity among participants in the clean-up operations after the accident on the Chernobyl Nuclear Power Station in 1986. The value for stabilization parameter $\alpha$was selected using described above criterion I_α. Estimates show, that observed morbidity increases in time with higher rate than the real, unobserved one, because of 'morbidity accumulation' effect among people skipping regular examinations. The described approach adjusts estimates for this effect.

1.4 Estimation of HIV infection rate on the dynamics of AIDS cases

Large latent period between HIV infection an AIDS manifestation makes it difficult to judge about the amount of HIV infected people in population. Specific expensive surveys of risk groups are needed to get reliable information about HIV prevalence. Implementation of inverse problems approach can help to estimate the number of HIV infected people from dynamics of AIDS cases, which is reported for the health care system needs. The number of people infected by HIV in year t at age x $\Psi(t,x)$ is related with the number of AIDS diagnoses in year t at age x $U(t,x)$ by integral equation

$$U(t,x) = \int_0^x L(x,s) \exp\left(-\int_s^x \mu_c(t-x+\tau,\tau)\,d\tau\right) \Psi(t-x+s,s)\,ds \quad (1.4)$$

where $\mu_c(t,x)$ – mortality in year t at age x, $L(x,s)$ – probability density function for distribution of AIDS diagnoses age x if at age s a person was infected with HIV. Age specific mortality supposed to be known from national data, function $L(x,s)$ can be estimated from the clinical data and data about AIDS cases among patients which were infected with HIV during blood transfusion. The most common models for $L(x,s)$ are exponential, Weibull, Markov chain model. Write equation (1.4) in matrix form

$$U = A\Psi,$$

where U and Ψ are vectors composed by values of functions $U(.)$ and $\Psi(.)$ for corresponding birth cohorts, A - block-diagonal matrix composed by triangular matrixes with elements for k-th cohort

$$a_{ij}^k = \begin{cases} 0 & s_j > x_i^k \\ \beta(t_i^k, x_i^k) L(x_i^k, s_j) \exp\left(- \int\limits_{s_j}^{x_i^k} \mu_c \left(d_k + \tau, \tau\right) d\tau \right) & s_j \leq x_i^k \end{cases} .$$

To stabilize solution of (1.4) the stabilization functional was used in form

$$\|Y - A\Psi\|^2 + \alpha\Omega\left(\Psi\right)$$

with $\Omega\left(\Psi\right) = \sum\limits_k \frac{1}{m_k} \sum\limits_{j=2}^{m_k} \left(\Psi_j^k - \Psi_{j-1}^k\right)^2$. The stabilized solution takes form

$$\Psi_\alpha = \left(A^T A + \alpha D\right)^{-1} A^T Y,$$

matrix D - block-diagonal matrix composed by three diagonal matrixes. For *k-th* cohort the matrix holds $2/m_k$ at the main diagonal, $-1/m_k$ at the other two diagonals and $1/m_k$ as the first and the last elements of the matrix.

Results of HIV infection rate from AIDS diagnoses dynamics estimation on simulated data are presented. Stabilization parameter value was selected using described criterion I_α.

1.5 Estimation of survival in wild population on survival of captured animals in laboratory

A specific problem arises in connection with investigation of life span in wild populations of different spices. A problem how to estimate survival curve in wild population of flies is considered in Muller *et al.* (2004). A portion of wild flies were cached and kept in laboratory in conditions similar to conditions in wild nature. Survival curve was calculated for cached cohort and some mathematical technique is to be applied to produce survival curve for wild population. This is typical inverse problem. If laboratory conditions do not change survival of fly then survival in laboratory $S_{lab}(.)$ is related with survival in wild stable population $S_{wild}(.)$ by integral equation

$$S_{lab}(x) = \frac{1}{e_0} \int_x^\omega S_{wild}(y)dy \tag{1.5}$$

where ω is maximum life span, e_0 is life expectancy in wild population. By differentiating the last equation on x one obtains equation

$$S_{wild}(x) = \frac{\frac{d}{dx}S_{lab}(x)}{\frac{d}{dx}S_{lab}(0)}.$$

One can estimate numerically the derivative from survival function in laboratory and calculate fron it $S_{wild}(x)$. This is done in Muller *et al.* (2004).

The other possibility is to solve numerically equation (1.5) itself. The corresponding matrix equation is

$$AX = S_l \tag{1.6}$$

where $X = \frac{1}{e_0}S_w$, S_w and S_l are vectors of values of survival functions observed daily in wild and in laboratory populations respectively, A– triangular matrix with 0 bellow the main diagonal and 1 at the other places. Solution of the system (1.6) was stabilized as described above. Results of estimation with simulated and real data are presented. Stabilization parameter value was selected using described criterion I_α.

References

1. Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*, Wiley, New York.

2. Evans, S. N. and Stark, P. N. (2002). Inverse problems as statistics, *Inverse Problems*, **18**, R55-R97.

3. Michalski A. I. (1987). Choosing an algorithm of estimation based on samples of limited size, *Automatization and Remote Control*, **48**, 909-918.

4. Michalski, A.I., Morgenstern, W., Ivanov, V.K. and Maksyitov, M.A. (1996). Estimation of morbidity dynamics from incomplete follow-up studies, *Journal Epidemiology and Biostatistics*, **1**, 151-157.

5. Muller, H.-G., Wang, J.-L., Carey, J. R., Caswell-Chen, E. P., Chen, C., Papadopoulos, N. and Yao, F. (2004). Demographic window to aging in the wild: constructing life tables and estimating survival functions from marked individuals of unknown age, *Aging Cell*, **3**, 125-131.