PRIOR INFORMATION IN BAYESIAN IDENTIFICATION OF A LINEAR REGRESSION MODEL

Ladislav Jirsa[1], **Ferdinand Varga**[2], **Miroslav Kárný**[1]

[1]*Institute of Information Theory and Automation*
Academy of Sciences of the Czech Republic
[2]*2nd Medical School, Charles University*
Prague, Czech Republic
E-mail: jirsa@utia.cas.cz

Abstract: We present a construction of prior information for Bayes identification of linear regression model with normal noise. We apply this methodology for modelling of time-activity function $A(t)$ of thyroid after administration of radioactive iodine ^{131}I in nuclear medicine. The model is tested on 2 355 data sequences, containing 4–9 pairs of (t_i, A_{t_i}). 3 pairs are used for identificaion, activity of the 4$^{\text{th}}$ one is predicted. Excluding 0.81 % of outlying sequences, the mean of relative prediction error is -0.0004, median -0.0544 and standard deviation 0.42. Distribution of integral of $A(t)$, proportional to absorbed dose, is numerically simulated using MCMC and approximated by log-normal *pdf*.

Keywords: fictitious data, information matrix, Gauss-inverse-Wishart, MCMC

1. INTRODUCTION

Linear regression model is a widely used tool for probabilistic modelling. Its identification using Bayes methodology (Peterka, 1981) enables to formulate prior information improving precision of estimated parameters or, as shown, allowing the estimation if there is not enough data to match some necessary conditions.

In nuclear medicine and radiation protection, for determining absorbed dose of radiation caused by a radioactive source (particularly ^{131}I in thyroid gland) with the standard MIRD methodology (Loevinger *et al.*, 1988), it is necessary to know integral of the source activity $A(t)$ as a function of time. Logarithm of $A(t)$ is described by a static linear regression model with normal noise.

The task is specific by a low amount (usually 3–5) of measured data pairs (t_i, A_{t_i}). For prediction of activity, the model must be identified even with the 2–3 initial measurements, ideally with specified uncertainty of the prediction. Bayes methodology can be successfully applied for estimation tasks with a few noisy data, e.g. (Fonseca, 1991; Heřmanská and Kárný, 1997). Theory of prior information in linear regression models exists, e.g. (Kárný *et al.*, 2001; Kracík and Kárný, 2005). But Bayes methodology leaves subjective space for its construction, therefore the prior knowledge and a limited information in a few noisy data must be carefully balanced.

The bi-phasic model of $A(t)$ dominates over the classical mono-exponential model in the study using long sequences $\{(t_i, A_{t_i})\}$ measured twice as more often as usually (Heřmanská *et al.*, 2001). However, its identification using data of usual amount and frequency leads to physically meaningless estimates in about 40 % of cases. Therefore, a robust approach has to be developed to utilise all the data available in clinical practice.

2. MATERIALS AND METHODS

2.1 Aim of the work

The aim is to estimate probaility density function $f(\xi|\text{data, prior})$, where

$$\xi = \int_0^{+\infty} A(t)\ \mathrm{d}t. \tag{1}$$

Data are represented by a measured sequence $\{(t, A_t)\} \equiv \{(t_i, A_{t_i})\}_{i=1}^n$, where $2 \leq n \lesssim 9$.

2.2 Model description

The bi-phasic model of $A(t)$ is a linear regression model

$$\ln A(t) = k_1 + k_2 \ln t + k_3 t^{\frac{2}{3}} \ln t - \frac{t}{T_\mathrm{p}} \ln 2, \tag{2}$$

$\vartheta \equiv (k_1, k_2, k_3)'$, where $'$ means transposition, is a vector of regression coefficients and T_p is a physical half-life of ^{131}I (8.04 days). Unit of activity is MBq, unit of time is day, $t > 0$.

The equation (2) can be formally rewritten as

$$d_t = \psi_t' \vartheta + e_t, \tag{3}$$

$d_t = \ln A_t + t/T_\mathrm{p} \ln 2$ and $\psi_t = (1,\ \ln t,\ t^{2/3} \ln t)'$. The term $e_t \sim \mathcal{N}(0, r)$ represents normal noise with variance r unknown but constant. Let us denote *data vector* $\Psi_t = (d_t,\ \psi_t')'$. Vector of *unknown parameters* is $\Theta = (\vartheta', r)'$.

2.3 Conjugated system and posterior pdf

The conjugated posterior *pdf* is Gauss-inverse-Wishart (or Normal-inverse-Gamma)

$$\begin{aligned} f(\vartheta, r|L, D, \nu) &= \mathcal{I}(L, D, \nu)^{-1} \times \\ &\times r^{-\frac{\nu}{2}} \exp\left\{-\frac{1}{2r}\left[\left({}^{\lfloor\psi}L\vartheta - {}^{\lfloor d\psi}L\right)' {}^{\lfloor\psi}D \left({}^{\lfloor\psi}L\vartheta - {}^{\lfloor d\psi}L\right) + {}^{\lfloor d}D\right]\right\} \end{aligned} \tag{4}$$

with a normalising constant $\mathcal{I}(L, D, \nu)$ and $r > 0$. Data are expressed by finite sufficient statistics: extended information matrix V (decomposed into $V \equiv L'DL$, where L is a lower triangular matrix with unit diagonal and D is a diagonal matrix) and a count statistics ν

$$V_t = V_{t-1} + \Psi_t \Psi_t' \qquad \nu_t = \nu_{t-1} + 1. \tag{5}$$

Without a prior information, V_0 is a zero matrix 4×4 and $\nu_0 = 0$ (so called *improper prior*). We introduce partitioning of V (and also L and D) into submatrices ${}^{\lfloor d}V$ of size 1×1, ${}^{\lfloor\psi}V$ of size $\mathring{\psi} \times \mathring{\psi}$ and ${}^{\lfloor d\psi}V$ of size $\mathring{\psi} \times 1$, where $\mathring{\psi} = 3$ is a length of the regression vector, like

$$V = \begin{pmatrix} {}^{\lfloor d}V & {}^{\lfloor d\psi}V' \\ {}^{\lfloor d\psi}V & {}^{\lfloor\psi}V \end{pmatrix}. \tag{6}$$

Then, (4) is obtained by algebraic operations on the multidimensional normal *pdf*.

If $\hat{\vartheta} = {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor d\psi}L$, the marginal *pdf* of (4) on ϑ is ($\propto$ means proportional up to a constant)

$$f(\vartheta|L,D,\nu) \propto \left[1 + \left({}^{\lfloor d}D\right)^{-1}\left(\vartheta - \hat{\vartheta}\right)'\,{}^{\lfloor\psi}L'\,{}^{\lfloor\psi}D\,{}^{\lfloor\psi}L\left(\vartheta - \hat{\vartheta}\right)\right]^{-\frac{1}{2}(\nu-2)}. \qquad (7)$$

First and second central moment of r and ϑ is

$$\begin{aligned} \mathcal{E}(r) &= \frac{{}^{\lfloor d}D}{\nu - \mathring{\psi} - 4} \equiv \hat{r}, & \mathrm{var}(r) &= \frac{2\hat{r}^2}{\nu - \mathring{\psi} - 6}, \\ \mathcal{E}(\vartheta) &= {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor d\psi}L \equiv \hat{\vartheta}, & \mathrm{cov}(\vartheta) &= \hat{r}\;{}^{\lfloor\psi}L^{-1}\,{}^{\lfloor\psi}D^{-1}\left({}^{\lfloor\psi}L'\right)^{-1}. \end{aligned} \qquad (8)$$

If $\mathring{\psi} = 3$, then for existence of $\mathcal{I}(L,D,\nu)$, $\mathrm{cov}(\vartheta)$ or $\mathrm{var}(r)$, must $\nu > 5$, 7 or 9 respectively.

2.4 Prior information

Prior information is expressed by two means: the prior restriction of ϑ domain (support) and construction of the prior statistics V_0 (resp. L_0 and D_0) and ν_0.

Domain restriction. The function $A(t)$ must meet the following requirements (see Figure 1):

1. $A(t) = 0$ for $t = 0$ and $t \to +\infty$,
2. $A(t)$ has a single global maximum $A(t_{\mathrm{max}})$ in t_{max},
3. $t_{\mathrm{d}} < t_{\mathrm{max}} < t_{\mathrm{u}}$, where, according to medical experience (Heřmanská, 1993), $t_{\mathrm{d}} = 4$ hours (0.167 days) and $t_{\mathrm{u}} = 72$ hours (3 days),
4. given $t_1 > t_{\mathrm{max}}$, $A(t)$ decreases for $t > t_1$ faster than decrease caused by a simple physical decay, i.e. by the term $-\frac{t}{T_{\mathrm{p}}}\ln 2$.

The first two requirements are fulfilled if $k_2 > 0$ and $k_3 < 0$. Because the form of (2) disables the analytical solution, the approximate procedure for $g(t) = \ln A(t) + t/T_{\mathrm{p}} \ln 2$ is shown. The first derivative $\dot{g}(t) = 0$ gives

$$k_2 + k_3 t^{\frac{2}{3}}\left(\frac{2}{3}\ln t + 1\right) = 0 \qquad (9)$$

with solution denoted $t_{1\mathrm{b}}$. Requiring $t_{\mathrm{d}} < t_{1\mathrm{b}} < t_{\mathrm{u}}$ and considering conditions above, we get

$$-k_3 t_{\mathrm{d}}^{\frac{2}{3}}\left(\frac{2}{3}\ln t_{\mathrm{d}} + 1\right) < k_2 < -k_3 t_{\mathrm{u}}^{\frac{2}{3}}\left(\frac{2}{3}\ln t_{\mathrm{u}} + 1\right). \qquad (10)$$

For $k_3 < 0 < k_2$ and $t < t_{\mathrm{m}} \equiv \exp(-3/2)$ days $\approx$ 5 hours 21 mins, $g(t)$ is always increasing, therefore, t_{d} is replaced by t_{m} in (10). The decay term is included by adding $t/T_{\mathrm{p}} \ln 2$ with corresponding values of t_{m} and t_{u} to the leftmost and rightmost side of (10), i.e.

$$0.019 < k_2 < -3.6\,k_3 + 0.26.$$

The requirements 1.–4. are then summarised in the linear form

$$M\,\vartheta < b, \qquad M = \begin{pmatrix} 0 & 1 & 3.6 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0.26 \\ -0.019 \\ 0 \end{pmatrix}. \qquad (11)$$

This constraint provides support for (7), i.e. the characteristic function $\chi(\vartheta)$ equal 1 iff (11) holds. The modified normalising constant neednot be considered in the numerical solution.

Prior statistics. As shown in (8), the statistics ν have sharp lower bounds so that the posterior *pdf* or its momets exist. With zero priors V_0 and ν_0 in (5), at least 6 data pairs must be processed if the posterior *pdf* exists, even with infinite mean noise, and at least 8 if the noise should be finite. According to the Bayes rule, we split the statistics $\nu_t = \nu_0 + n_t$, where n_t is number of processed data, and similarly $V_t = V_0 + \sum_{t=1}^{n_t} \Psi_t \Psi_t'$, where V_0 and ν_0 are constructed so that the prior *pdf* exist. The form of (4) allows the corresponding separation

$$f(\vartheta, r|L_t, D_t, \nu_t) \equiv f(\vartheta, r|V_t, \nu_t) \;\propto\; \mathcal{L}(\Psi(t); \vartheta, r)\; f(\vartheta, r|V_0, \nu_0), \qquad (12)$$

where $\Psi(t) \equiv (\Psi_1, \Psi_2, ..., \Psi_{n_t})$, $\mathcal{L}(\Psi(t); \vartheta, r)$ is the likelihood (must be finite) and $f(\vartheta, r|V_0, \nu_0)$ is the prior *pdf* (must be *pdf* with finite appropriate moments).

Theory of merging data based knowledge of multiple participants can be used for construction of V_0 (Kracík and Kárný, 2005). One participant ("expert") yields *fictitious data* (Kárný *et al.*, 2001; Kárný *et al.*, 2005), expressing the requested typical properties of a $A(t)$ and the model noise, to another participant ("estimator") assigns his belief (weight) to these data and processes them with the weights like real measured data. Specifically applied to this particular task,

$$V_0 = \sum_{i=1}^{m} \Psi_i^0\, \Psi_i^{0\prime}, \qquad\qquad \Psi_i^0 = \lambda_i \Psi_i^{\text{fict}} + \rho_i, \qquad (13)$$

where $\lambda_i \in \langle 0, 1\rangle$ is a weight of the i-th fictitious data vector Ψ_i^{fict} and $\rho_i = (r_i^{\text{fict}}, 0, 0, 0)'$ where r_i^{fict} is a noise of the i-th log-activity (zeros in the $\mathring{\psi}$-part of ρ express "exact" measurement of time) and m is the number of ficititious data vectors. For V_0 regular, $m \geq \mathring{\psi} + 1 \equiv 4$. As the fictitious data, some representative data pairs from averaged historical measurements were chosen to describe the initial, maximum and terminal stages of accumulation. r^{fict} is a usual uncertainty of one measurement observed from the same data.

For existence of finite prior mean noise (8), ν_0 was chosen 7.05. After processing 2 pairs of real data, finite posterior noise covariance exists. The part of subjectivity is the performance-optimum choice of Ψ_i^{fict}, $\lambda_i = 0.01$ and $r_i^{\text{fict}} = 0.0015$.

2.5 Algorithmic solution

All the computations were done by *square-root algorithms* operating on the matrices L and D, directly constructed from the data vector Ψ (Kárný *et al.*, 2005) because of stability. Measured activities were divided by administered activities for a similar scaling. The space of ϑ was transformed in ϑ^* for zero mean and unit covariance of (7). With entry-wise $\sqrt{\lfloor\psi D}$,

$$T \;=\; \sqrt{\tfrac{\nu-2}{\lfloor d D}\, \lfloor\psi D}\;\; \lfloor\psi L \qquad\qquad \vartheta^* \;=\; T(\vartheta - \hat{\vartheta}). \qquad (14)$$

Similarly (11), $M^* = MT^{-1}$, $b^* = b - M\hat{\vartheta}$ so that $M^*\vartheta^* < b^*$. The transformed *pdf* (7) with the support restriction (14) was sampled using Langevin diffusion algorithm (Roberts and Tweedie, 1996) which does not require normalising constant. The optimum Markov Chain (MC) step size could not be determined analytically for the posterior *pdf*, it was estimated by a heuristic rule obtained from multiple runs of the MC with different step sizees on 3 876 data sequences. MCs perform close to their optimum. Initial point of MC was chosen by optimization of the quadratic form in the denominator of (7) with the constraints (11). 5 000 samples were found sufficient after 500 of burn-in. Each parameter sample ϑ_j^* was, after the inverse transformation, substituted into (2) and integral ξ_j (1) was computed from 0 to 70 days using the adaptive step-size algorithm. For each data sequence, samples ξ_j and $\ln \xi_j$ created two histograms, distribution of which was tested by Kolmogorov-Smirnov test, Bayes-based test and skewness of both histograms was computed.

3. RESULTS

First, prediction of the model was tested both with nontrivial (presented) and trivial ($V_0 = \mathrm{diag}(10^{-12})$, $\nu_0 = 5$) prior statistics on 2 355 data sequences of at least 4 data pairs. 3 pairs were used for identification and the 4th one, usually following after 1–3 days, was predicted. This choice is justified by usually not more than 3 measurements after a diagnostic administration.

Without the prior constraints (11) and with trivial prior statistics, 40 % of data sequences were excluded for leading to estimates violating the requirements for physical behaviour of $A(t)$. Despite the best predictions, number of outlied predictions (relative error >3) is high. Then, the prior constraints were considered, either with trivial or nontrivial prior statistics. All the data sequences led to meaningful estimates. The case with nontrivial prior statistics performs lower both relative prediction error and its standard deviations (see Table 1) and decrease standard deviation of $f(\ln \xi)$ by 64 % in average compared to the trivial ones.

Table 1: Relative prediction errors in cases: 1) no prior constraints, 2) prior constraints and trivial prior statistics, 3) prior constraints and nontrivial prior statistics

#	mean	median	st.dev.	data	outliers
1)	0.0576	−0.0066	0.475	1 403	2.28 %
2)	−0.0968	−0.1456	0.431	2 355	0.85 %
3)	−0.0004	−0.0544	0.416	2 355	0.81 %

Next, distribution $f(\xi)$ in (1) was tested. Kolmogorov-Smirnov test did not prove normality of either ξ or $\ln\ \xi$. Bayes test preferred log-normal *pdf* of ξ but on a narrow space normal vs. log-normal. Then, a skewness comparison of $f(\xi)$ and $f(\ln\ \xi)$ after excluding samples out of $\hat{\xi} \pm 3\sigma$ was done. For $f(\xi)$, mean, median and standard deviation of skewness were 1.66, 0.84 and 3.53 respectively, whereas for $f(\ln \xi)$ 0.29, 0.24 and 0.61 respectively. Although the distribution was not classified, normal *pdf* (zero skewness) might correspond better with $f(\ln\ \xi)$. Practical experience shows this approximation sufficient with respect to existing uncertainty.

Figure 1 shows an example of $A(t)$ identified on 2 initial data pairs, other data are predicted.

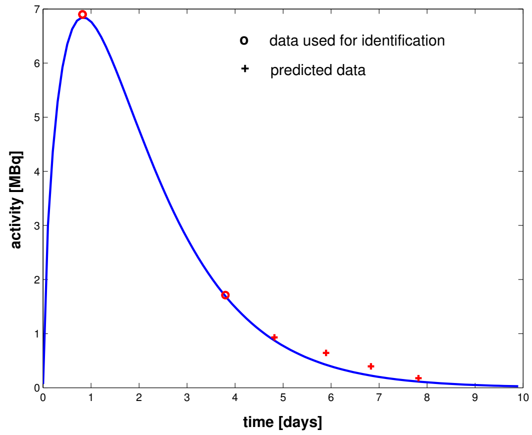

Figure 1: Example of $A(t)$ (one sample) identified on 2 data pairs

4. CONCLUSIONS

Robust and stable probabilistic identification of bi-phasic model of thyroid activity $A(t)$ after ^{131}I administration was presented. The model was tested by prediction of future data. Prior information guarantees physical meaningfullness of $A(t)$ and the prior statistics improve predictive abilities of the bi-phasic model (2) and variance of $f(\xi)$ (1). Although standard deviations of relative prediction errors seem high (above 40 % of activity magnitude), we must take into accout limited quality and relatively high natural uncertainty of measured data. It was observed that reliable predictions are given even after 2 measurements.

Algorithmic solution appears robust and stable. Impossibility of analytic $f(\vartheta) \to f(\xi)$ requires numerical transformation outlined in the paper. On contemporary PCs, one determining of $f(\vartheta)$ takes 1–2 seconds in MATLAB and fractions of seconds in C++. To estimate distribution of absorbed dose by MIRD, $f(\xi)$ is directly applicable for its linear dependence on the dose.

Use of the model can contribute to treatment planning and quality, radiation protection and quality of future data. Further work would focus on improvement of prior information, better analysis of convergence using a stopping rule and more exact analytical approximation of $f(\xi)$.

ACKNOWLEDGEMENTS

This research was supported by the grants AV ČR 1ET 1007 50404 and MŠMT ČR 1M0572.

REFERENCES

Fonseca, C. M. (1991), ‘Bayesian estimation of the intensity of low-level radiation sources’, *Jaderná energie* **37**, 83–97.

Heřmanská, J. (1993), *Bayesian Approach to Dosimetric Data Evaluation for Medical Use of ^{131}I*, Clinic of Nuclear Medicine, 2nd Medical Faculty, Charles University, Prague. Associated Professor Thesis, 103 pp. In Czech.

Heřmanská, J., Kárný, M., Zimák, J., Jirsa, L., Šámal, M. and Vlček, P. (2001), ‘Improved prediction of therapeutic absorbed doses of radioiodine in the treatment of thyroid carcinoma’, *Journal of Nuclear Medicine* **42**(7), 1084–1090.

Heřmanská, J. and Kárný, M. (1997), ‘Bayesian estimation of effective half-life in dosimetric applications’, *Computational Statistics and Data Analysis* **24**(5), 467–482.

Kárný, M., Böhm, J., Guy, T. V., Jirsa, L., Nagy, I., Nedoma, P. and Tesař, L. (2005), *Optimized Bayesian Dynamic Advising: Theory and Algorithms*, Springer, London.

Kárný, M., Khailova, N., Nedoma, P. and Böhm, J. (2001), ‘Quantification of prior information revised’, *International Journal of Adaptive Control and Signal Processing* **15**(1), 65–84.

Kracík, J. and Kárný, M. (2005), Merging of data knowledge in Bayesian estimation, *in* J. Filipe, J. A. Cetto and J. L. Ferrier, eds, ‘Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics’, INSTICC, Barcelona, pp. 229–232.

Loevinger, R., Budinger, T. F. and Watson, E. E. (1988), *MIRD Primer for absorbed dose calculations*, The Society of Nuclear Medicine, New York.

Peterka, V. (1981), Bayesian system identification, *in* P. Eykhoff, ed., ‘Trends and Progress in System Identification’, Pergamon Press, Oxford, pp. 239–304.

Roberts, G. O. and Tweedie, R. L. (1996), ‘Exponential convergence of Langevin distributions and their discrete approximations’, *Bernoulli* **2**(4), 341–363.