Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006
Draft of the paper for the proceedings

Identification of Thyroid Gland Activity in Radiotherapy

Ladislav Jirsa
Academy of Sciences, Prague, Czech Republic
Anthony Quinn
Trinity College, Dublin, Ireland
aquinn@tcd.ie
Ferdinand Varga
Charles University, Prague, Czech Republic

Summary

The Bayesian identification of a linear regression model for time dependence of thyroid gland activity in ^{131}I radiotherapy is presented. Prior knowledge is elicited via hard parameter constraints and via the merging of information from an archive of patient records. This prior regularization is shown to be crucial in the reported context, where data typically comprise only 2-3 high-noise measurements. The posterior distribution is simulated via a Langevin diffusion algorithm, whose optimization for the thyroid activity application is explained. Improved patient-specific predictions of thyroid activity are reported. The posterior inference of the total patient radiation dose is presented, allowing the uncertainty of the dose to be quantified in a consistent manner.

Keywords and Phrases: External information; Information matrix; Prior constraints; Normal-inverse-Gamma; Langevin diffusion; Nonparametric stopping rule.

1. INTRODUCTION

Radioactive ^{131}I is used for radiotherapy in thyroid gland carcinoma, since anorganic iodine is selectively accumulated in the thyroid gland. The activity of the organ, A_t (MBq), at time t (days) is related to the total *dose* (energy absorbed per unit mass), $\mathcal{D} = \mathcal{S}\xi$, via $\xi = \int_0^{+\infty} A_t \;\mathrm{d}t$, where $\mathcal{S}$ is a known organ- and isotope-specific constant. ξ is therefore a key quantity in the design of an effective and safe administration of ^{131}I. For practical reasons, measurement pairs, $\{(t_i, d_{t_i})\}_{i=1}^n \equiv D$, are available at only a small number, $2 \leq n \lesssim 9$, of non-uniform sampling times, t_i, and are subject to considerable measurement noise. The principal task of this paper is to infer A_t, and hence ξ, in this difficult measurement regime.

2. A REGRESSION MODEL FOR THYROID ACTIVITY

A_t should observe the following prior constraints: **(i)** $A_t \geq 0, = 0$ at $t = 0$ and as $t \to +\infty$; **(ii)** A_t is unimodal, achieving a maximum at t_m, where $t_m \in (t_l, t_u)$; **(iii)** from medical experience $t_l = 4$ hours (0.167 days) and $t_u = 72$ hours (3 days); and **(iv)** for some $t_h > t_m$, the decrease of A_t, $t > t_h$, is faster than the physical decay of ^{131}I. The following 3-parameter model was proposed in (Heřmanská, 2001):

$$\ln A_t = a_1 + a_2 \ln (c\,t) + a_3\,(c\,t)^{\frac{2}{3}} \ln (c\,t) - \frac{t}{T_p} \ln 2, \tag{1}$$

Here, $a = [a_1, a_2, a_3]'$ is the *patient-specific* vector of unknown linear regression coefficients, T_p is the physical half-life of ^{131}I (8.04 days) and c is a known time-scale coefficient, the value of which will be discussed shortly. The positive measurement process, d_t, is taken as conditionally log-normal. Since $A_t \gg 0$, it follows that

$$x_t \equiv \ln d_t + \frac{t}{T_p} \ln 2 = \psi_t' a + e_t, \tag{2}$$

where $e_t \sim \mathcal{N}(0, r)$ is uncorrelated (white) noise, with variance r. In (2), $\psi_t = [1,\ \ln (c\,t),\ (c\,t)^{2/3} \ln (c\,t)]'$ is the regressor at time t. A sample activity curve, A_t, is illustrated in Fig. 1.

Grants AV ČR 1ET 1007 50404 and MŠMT ČR 1M0572 partially supported this work.

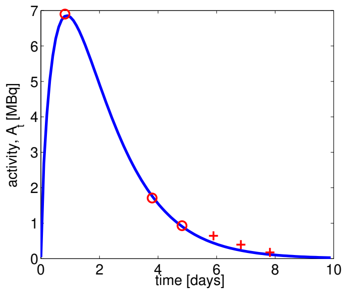

Figure 1: *Sample activity curve, A_t, identified on 3 data pairs.*

3. BAYESIAN CONJUGATE INFERENCE OF THYROID ACTIVITY

The conjugate parameter prior for $f(x_t|a,r)$ in (2) is $\mathcal{N}i\mathcal{G}(V_0,\nu_0)$, *i.e. Normal-inverse-Gamma*, with prior parameters V_0 and ν_0. However, a hard constraint, $a \in \mathbb{A} \subset \mathbb{R}^3$, must be respected in order to satisfy **(i)**–**(iv)** (Section 2). This is expressed via the uniform prior, $\chi_{\mathbb{A}}(a)$, with support $\mathbb{A}$. The marginal posterior distribution of a is then (Kárný, 2005)

$$f(a|L,\Lambda,\nu) \propto \tilde{f}(a|L,\Lambda,\nu)\chi_{\mathbb{A}}(a), \tag{3}$$

$$\tilde{f}(a|L,\Lambda,\nu) = \mathcal{I}^{-1}(L,\Lambda,\nu)\left[1+\lambda_{11}^{-1}\left(a-\tilde{\mathsf{E}}a\right)' L'_{aa}\Lambda_{aa}L_{aa}\left(a-\tilde{\mathsf{E}}a\right)\right]^{-\frac{1}{2}(\nu-2)} \tag{4}$$

Here, the n measurements, D (Section 1), have entered the posterior inference via the LD-decomposition (Kárný, 2005) $V = L'\Lambda L$, of the *extended information matrix*,

$$V = V_0 + \sum_{i=1}^{n} \Psi_{t_i}\Psi'_{t_i}, \tag{5}$$

where $\Psi_{t_i} = (x_{t_i},\ \psi'_{t_i})'$, from (2), is the *extended regressor* at sampling time t_i. The lower-diagonal matrix, L, has been partitioned as

$$L = \begin{pmatrix} l_{11} & 0 \\ l_{a1} & L_{aa} \end{pmatrix},$$

where $l_{11} = 1$ is the $(1,1)$ element. Similarly, λ_{11} is the $(1,1)$ element of diagonal matrix, Λ. In (4), $\nu = \nu_0 + n$ is the *degrees-of-freedom* parameter. The normalizing constant, $\mathcal{I}(L,\Lambda,\nu)$, of the unconstrained distribution (4) is available in closed form, but that of the complete inference (3) must be evaluated numerically, if needed. Finally, the posterior moments of a with respect to the *unconstrained* inference, $\tilde{f}(a|L,\Lambda,\nu)$ (4), are, respectively,

$$\tilde{\mathsf{E}}[a|L,\Lambda,\nu] \equiv \tilde{\mathsf{E}}a = L_{aa}^{-1}L_{a1}, \qquad \tilde{\mathsf{cov}}[a|L,\Lambda,\nu] = \frac{\lambda_{11}}{\nu-7}\, L_{aa}^{-1}\Lambda_{aa}^{-1}\left(L'_{aa}\right)^{-1}. \tag{6}$$

3.1. *Prior Support,* $\mathbb{A}$

Constraints **(i)** and **(ii)** in Section 2 are satisfied if $a_3 < 0 < a_2$. Examining the function $g_t = \psi'_t a$ (2), its unique maximizer, t_{mg}, solves $g'_t = 0$. Then, the unique maximizer, t_m, of A_t (see **(ii)**) exists and $t_{mg} > t_m$ since $\ln 2/T_p > 0$. For $t > t_{mg}$, $g'_t < 0$, satisfying **(iv)** for $t_h = t_{mg}$. Imposing $t_l < t_m < t_u$ (see **(iii)**) via the derivative $(\ln A_t)'$, we get

$$-\,a_3(c\,t_l)^{\frac{2}{3}}\left(\frac{2}{3}\ln(c\,t_l)+1\right)+\frac{t_l}{T_p}\ln 2 < a_2 < -\,a_3(c\,t_u)^{\frac{2}{3}}\left(\frac{2}{3}\ln(c\,t_u)+1\right)+\frac{t_u}{T_p}\ln 2.$$

The first inequality here can be written as $a_2 + ka_3 > q$, where $q > 0$. For $t_m = t_l$, it becomes $a_2 + ka_3 = q$. If $k < 0$, then it is necessary to have $a_2 < 0$ for some values of a_3, in order that t_m can reach its lower limit. This violates the condition $a_2 > 0$. Hence, to guarantee $a_2 > 0$, c can be chosen to ensure that $k \geq 0$. In particular, $k = 0$ if $c = \frac{1}{t_l}\exp\left(-\frac{2}{3}\right) \equiv 1.3388$. The first inequality above becomes $a_2 > t_l \ln 2/T_p$, $\forall a_3$. Then, $a_2 > 0$ is redundant.

These bounds may be combined into the following linear matrix element-wise inequality, whose solution space, $\mathbb{A}$, provides the support for (3):

$$M\ a\ <\ b, \qquad M = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 4.8687 \\ 0 & -1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0.2586 \\ -0.0144 \end{pmatrix}. \tag{7}$$

3.2. *Prior Parameters, V_0 and ν_0*

External information is merged from a database of 3 876 sampled patient activity curves, collected at the Clinic of Nuclear Medicine at Motol Hospital in Prague. V_0 is constructed using these data sequences $D_j, j = 1, \ldots, 3\,876$, each scaled to $(0, 1]$. They form m clusters, one for each of the standard measurement times in the Hospital, and so can be described by a mixture, $M(x_t, t)$, of components in (x_t, t)-space (2). The full-rank prior information matrix is then $V_0 = w \int M(x_t, t)\Psi\Psi' \,\mathrm{d}\Psi$, using the information merging technique in (Kracík, 2005). The weight, w, is assigned the value 5.3×10^{-5}, which optimizes the prediction of the 4^{th} measurement (Section 4). This value of V_0 in (5) ensures that $\tilde{\mathsf{E}}[a|V_0, \nu_0] \in \mathbb{A}$ (Jirsa, 2007).

ν_0 was chosen as 7.05 so that $\tilde{\mathsf{cov}}[a|L_0, \Lambda_0, \nu_0]$ exists. In the radiotherapy context, the minimal number of observations is $n = 2$, giving $\nu = 9.05$, ensuring finiteness of the posterior variance of r.

3.3. *Posterior Inference Methodology*

The hard constraint, $\chi_{\mathbb{A}}(a)$ (3), obviates analytical evaluation of the posterior moments of a. Moreover, the posterior inference of ξ (Section 1) is unavailable since the mapping $a \to \xi(a)$ cannot be expressed in closed form. Hence, a simulation-based method is required.

The distribution of the transformed variables, $a^* = T(a - \tilde{\mathsf{E}}a)$, which uses (6) with $T = \sqrt{\frac{\nu-7}{\lambda_{11}}\Lambda_{aa}}L_{aa}$, was sampled. Here, $\sqrt{\Lambda_{aa}}$ denotes the element-wise square-root. The transformed support, $\mathbb{A}^*$ (7), is the solution space of $M^*\ a^* < b^*$, with $M^* = MT^{-1}$ and $b^* = b - M\tilde{\mathsf{E}}a$. From (6), the unconstrained posterior distribution of a^* is therefore zero-mean with identity covariance matrix. Markov chain realizations were generated efficiently in these transformed variables, via the Langevin diffusion algorithm (Roberts, 1998).

The chain was initialized at the MAP estimate, found by constrained optimization of the quadratic denominator (4). Within a burn-in period (200–400 samples), the initial step-size of the Markov chain was trimmed to reach its near-optimum value (Roberts, 1998), ensuring good mixing of the chain.

After the burn-in, the samples were each inverted via the linear mapping above, generating approximate realizations from (3). Finally, each realization of a was substituted into (1), and the equivalent realizations from $f(\xi|L, \Lambda, \nu)$ were obtained by numerical integration (Section 1).

Using a nonparametric Bayesian stopping rule (Quinn, 2007), sampling from $f(\xi|L, \Lambda, \nu)$ in a specific patient case is stopped when the uncertainty in the 95% confidence interval bounds and width falls below 3 %. The average number of samples at stopping, for 700 archive data sequences, is 4 529. Bayesian confidence intervals for ξ can be evaluated under this procedure in fractions of a second, when implemented in C++ on a standard PC. Hence, it is suitable for on-line use in clinical practice.

4. PREDICTION OF THYROID ACTIVITY AND INFERENCE OF DOSE

The influence of the priors above was examined in the context of a database of 2 355 records, each containing at least 4 measurement pairs, D. For each patient case, A_t (1) was inferred from the first 3 measurements, and used to predict the 4^{th}, which typically followed after 1–3 days (see Fig. 1). This reflects the usual practice of taking no more than $n = 3$ measurements per patient. Predictions were generated for each patient as follows: (a) via (4), with $V_0 = \epsilon I_4$, $\epsilon = 0^+$, and $\nu_0 = 7.05$ (in which case about 40 % had to be rejected, as the inferred A_t was physically impossible); (b) via (3), initialized as in (a) above; and (c) via (3), with V_0 assigned as in Section 3.2. The major prediction advantage in case (c) is evident in Table 1. The statistics describe the distribution of relative prediction errors over the set of 2 355 records. The 'data' column gives the number of data sequences yielding valid predictions. The percentage of outliers in relative prediction error (being more than 3 standard deviations from the mean), for valid predictions, is in the last column.

Table 1: *Relative prediction error for three prior knowledge structures: (a) without hard parameter constraints or externally informed V_0; (b) with hard parameter constraints but without externally informed V_0; and (c) with both hard parameter constraints and side-informed V_0 (see Section 3.2).*

Prior	mean	median	st.dev.	data	outliers
(a)	0.0576	−0.0066	0.475	1 403	2.28 %
(b)	−0.0968	−0.1456	0.431	2 355	0.85 %
(c)	−0.0004	−0.0544	0.416	2 355	0.81 %

The influence of externally-informed V_0 (Section 3.2) on the inference of a_2 is illustrated in Fig. 2 for a specific patient case.

The empirical distribution of ξ, corresponding to the data in Fig. 1, is shown in Fig. 3. For data from 700 patients, the mean skewness of $f(\xi|L, \Lambda, \nu)$ is 1.66, whereas that of $f(\ln \xi|L, \Lambda, \nu)$ is 0.29, supporting the hypothesis of a nonsymmetric pdf such as the log-normal.

5. DISCUSSION AND CONCLUSION

The hard constraint (7) has been effective in ensuring physically realizable inferences of thyroid activity in ^{131}I radiotherapy (Table 1). This, and the merging of external information from a large database of patient data, via prior parameter V_0 (Section 3.2), provides reliable activity predictions in this data-parsimonious, high-noise context. The influence of this prior information on identification of parameter a_2 is illustrated in Fig. 2. Although the parametric form of $f(\xi|L, \Lambda, \nu)$ (see Fig. 3) is not known, we found that it satisfied several log-normal properties, notably positive support and skewness.

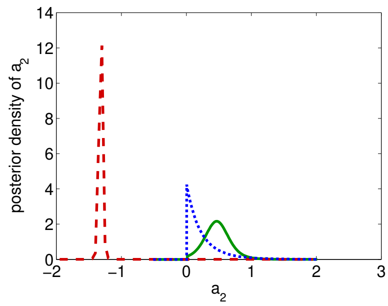

Figure 2: *Marginal posterior inference of a_2 for a specific patient. The* dashed *curve corresponds to case (a) in Table 1, the* dotted *curve to case (b), and the* solid *curve to case (c).*

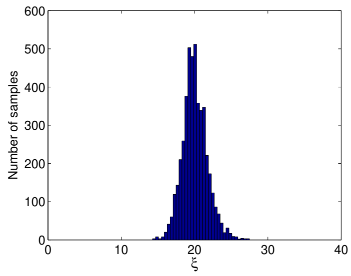

Figure 3: *Empirical distribution of ξ estimated using the data in Fig. 1.*

This is in accordance with (Hamby, 1999), where $f(\xi)$ is found to be analytically log-normal, although their inference is not patient-specific. In future work, the fixed time-scale coefficient, c (1), could be introduced via another hard-constrained patient-specific parameter, resulting in a model with 4 regression coefficients. Other methods of exploiting data-based prior information can be tested, although the one presented in Section 3.2 provides significant improvement in identification, as reported in Section 4, as well as improved chain convergence and computational speed. The identified model will be used clinically in the design of patient-specific optimized administrations of ^{131}I, contributing to treatment quality and radiation protection.

REFERENCES

Hamby, D. M. and Benke, R. R. (1999). Uncertainty of the iodine-131 ingestion dose conversion factor. *Radiation Protection Dosimetry*, **82(4)**, 245–256.

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.

Jirsa, L., Quinn, A. and Varga, F. (2007). Identification of Thyroid Gland Activity in Radiotherapy. *Bayesian Analysis*, under review.

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

Kracík, J. and Kárný, M. (2005). Merging of data knowledge in Bayesian estimation. *Proc. of the 2nd International Conference on Informatics in Control, Automation and Robotics* (J. Filipe, J. A. Cetto, J. L. Ferrier, eds.), Barcelona, 229–232.

Quinn, A. and Kárný, M. (2007). Learning for nonstationary Dirichlet processes. *International Journal of Adaptive Control and Signal Processing*, **21(8)**.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximation to Langevin diffusions. *J. Roy. Statist. Soc. B* **60(1)**, 255–268.