Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

M. Kárný, A. Halousková, L. Zörnigová

Bayesian Pooling of Expert Opinions

No. 1799 February 1994

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (+42)(2)6641 4903
E-mail: utia@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the Institute.

Abstract. The quantification and exploitation of prior knowledge in control design becomes important for improving control efficiency. The paper presents results gained in extracting the common knowledge contained in different experts' opinions. A special case of the standard Bayesian formulation is given which is believed to reflect specific features of technical systems, namely,
– availability of a parametrized system model;
– possibility to quantify particular information items;
– availability of objective data complementing the subjective opinions of experts;
– preference of conservative opinion pooling (overconfidence is more dangerous).

The general formulation is applied to the technically important specific classes of ARX models and Markov chains.

Keywords: prior information, Bayes statistics, expert opinions, ARX model.

1 Introduction

Adaptive systems relying on recursive estimation of an unknown (multivariate) parameter of the system model constitute a wide and successful research as well as application area. A broad range of strong results have been gained. Quantification of prior information, i.e. sensible choice of initial estimation conditions and/or of tracking reference model [9] has been, however, underestimated. Such information may be very helpful, especially if there is a lack of real informative data.

The quantification problem has been naturally addressed in various application areas but according to our best knowledge almost nothing was done in the algorithmic control design. Just some particular results, oriented to the special important case of recursive least squares (RLS), arose in our group [5, 6, 2, 7]. They are quite useful, but too tightly related to the studied case giving no hint how to extend them to other situations. This incomplete understanding was observed during unsuccessful attempts to apply the methodology to the learning of qualitative models [3]. For this reason, a general approach to the expert opinion pooling, as described for instance in [1], has been elaborated adopting Berger's dictum: *even an approximate model of the relations among experts is* *(usually) better than a mechanically specified pooling rule.* This paper can be taken as status report on our effort to follow this recommendation.

2 Problem and its conceptual solution

Let several expert opinions be given, concerning probabilistic description of a system. A well grounded compromise is looked for, representing the different convictions as a whole.

The problem solution is based on the Bayesian approach; quantified expert knowledge is used as data for transforming a prior flat probability distribution into the posterior one, respecting all the opinions. Such a treatment is legitimate if there exists some common knowledge behind the different experts' opinions.

For computational reasons, the point estimate of the compromise is searched for, namely, maximum a posteriori probability (MAP) one.

3 Main result

Let $f_j(x)$, $j = 1, \ldots, n$ be probability (density) functions (p.(d).f.s) expressing particular opinions of n experts on an unknown quantity x. Then the recommended pooling of opinions into the p.(d.)f. $\hat{f}_0(x)$ is described by:

Main result [*MAP pooling*] Let particular opinions on an unknown quantity x be expressed in terms of:
• positive p.(d.)f.s $f_j(x)$, $j = 1, \ldots, n < \infty$;
• a pre-prior (flat) p.(d.)f. $\bar{f}(x)$ specifying a common experts' agreement on probable values of x.

Then, the recommended pooling is given by the p.(d.)f.

$$\hat{f}_0(x) = \prod_{j=1}^{n} [f_j(x)]^{\nu_j} \tag{1}$$

where $\nu = [\nu_1, \ldots, \nu_n]'$ is the eigenvector corresponding to the maximum eigenvalue of the symmetric positive semidefinite (n, n) matrix F

$$F_{kl} = \int \ln(f_k(x)) \ln(f_l(x)) \bar{f}(x)\, dx \quad k, l = 1, \ldots, n$$

(the symbol $\int\, d\bullet$ means multivariate integration over whole domain of the integrand).

The norm of ν is specified by the requirement $\int \hat{f}_0(x)\, dx = 1$.

If the entries of F are positive then the eigenvector is unique and has positive entries.

$\square$

This result is explained and justified in the following. It is applied to summarizing prior knowledge on regression coefficients and to creating a complex Markov-chain based predictor using several simpler models.

4 Modelling of expert-supplied information

The approach can be summarized as follows:

Pooling of experts' opinions is reduced to simple use of Bayesian paradigm at a "second" layer. The opinions, quantified by a matrix p, are taken as noisy data which are related to an unknown finite-dimensional "hyper"parameter, say q, w, Ω. The q part of the hyperparameter describes the ideal opinion compromise, the parameters w, Ω are exploited for modelling of relationships of experts' opinions.

A plausible probabilistic model $f(p|q, w, \Omega)$ relating the data p to the hyperparameter q, w, Ω is intuitively chosen and a flat prior p.d.f. $f(q, w, \Omega)$ specified. The pooling then consists of point estimation of q. The maximum a posteriori probability estimate $\hat{q}$ is chosen

$$\hat{q} = \arg\max_q f(q|p)$$

as feasible one with known reasonable properties.

The maximized p.d.f. $f(q|p)$ is the marginal p.d.f. of $f(q, w, \Omega|p)$ which results from the Bayes formula

$$f(q, w, \Omega|p) \propto f(p|q, w, \Omega) f(q, w, \Omega)$$

where $\propto$ means equality up to a normalization.

Particular elements for applying this pooling methodology are described in this section.

4.1 Problem elements and basic assumptions

Several expert opinions, labelled by $j = 1, \ldots, n < \infty$, concern an unknown quantity x. They are expressed by the (m, n) matrix p,

$$p = [p_{i|j} = Probability_j(x \in X_i)], \tag{2}$$

where $\{X_i\}_{i=1}^m$ is a disjunct covering of x-range.

For a continuous x, sets X_i are chosen so that $p_{i|j} \approx f_j(x_i)$ with $f_j(x_i)$ equal to the corresponding p.d.f. and x_i being a point from X_i. The asymptotic case with the maximum volume of X_i, say δ converging to zero, is of primary interest. Thus, m much larger than n is assumed. Even if x is discrete, this situation is of practical interest for us: the pooling is significant when the number of discrete x values is very large.

The model construction relies upon the following assumptions:

- All probabilities involved are assumed to be positive. This mutual absolute continuity precludes strict contradictions among experts (somebody assigns zero probability to an event probable in another's opinion).

- There exists a common "pre-prior" probability

  $$\bar{p}_i = \overline{Probability}(x \in X_i) \approx \bar{f}(x_i)\delta$$

  with $\bar{f}(\cdot)$ denoting corresponding p.d.f. in continuous case. This p.(d.)f. expresses a vague prior knowledge which should be modified by the opinions pooled.

- The "ideal" pool of the opinions is assumed to exist. It represents all relevant pieces of information about the variable x which are behind particular opinions. Due to the probabilistic nature of x, the ideal pool is given by an m-vector q with entries

  $$q_i = Probability_0(x \in X_i) \approx f_0(x_i)\delta \propto f_0(x_i) \qquad (3)$$

  where $f_0(\cdot)$ denotes the relevant unknown p.d.f. for continuous case.

The pooling problem consists of estimation of q using p and $\bar{p}$.

The pooling problem is (still) an incorrectly formulated task: no model of (always existing) dependences among opinions is given and their relationship to the ideal pooling is unspecified, too.

The modelling task is based on the presumption:

- The expert opinions are not exact, they contain "expert-induced uncertainty" (noise), but they are consistent with the ideal pool, i.e. for all j, $p_{i|j}$ is high (low) iff q_i is high (low).

Consequently, the related model is probabilistic and can be expressed by the p.d.f. $f(p|q, \text{nuisance parameter})$. This, together with a prior p.d.f. on the hyperparameter= (q,nuisance parameter) provides a Bayesian estimate (at the second level), i.e. solution of the pooling problem.

4.2 Model $f(p|q,$ nuisance parameter$)$

Log-normal model form is chosen as widely applicable model relating expert opinions p (2) to the ideal pool q (3). The following arguments support the choice:

- The "observed data" p are positive.
- The constructed p.d.f. should be heavy-tailed as occurrence of outlying observations is always high due to the subjectivity of their origin.
- An invariant level of *relative deviations* of data from the unobserved ideal seems to be a relevant description of uncertainty (the key property of log-normal model).

Covariance structure of the constructed model will reflect the following conditions:

- The possibility to make an error in guessing q_i is the higher the lower is the pre-prior probability $\bar{p}_i$.
- The various opinions $p_{i|j},\ j = 1,\ldots,n$ on a particular x_i are generally correlated (up to a repetition).

By combining both the above lists which outline the intuitive background of the model choice, the following log-normal model is selected

$$f(p|q,w,\Omega) = \mathcal{LN}_p(Ln(q)w', \Psi^{-1} \otimes \Omega^{-1}) \tag{4}$$

where

$Ln(\cdot)$ is the vector (matrix) of the logarithms of the argument entries,

$'$ means transposition,

$\Psi = diag(\bar{p}_1,\ldots,\bar{p}_m)$,

$\otimes$ denotes Kronecker product.

The unknown nuisance parameter is formed by the n-vector weight w and the positive definite matrix Ω.

The above log-normal p.d.f. $\mathcal{LN}_p(\cdot,\cdot)$ has the following explicit form

$$\begin{aligned} &\mathcal{LN}_p(Ln(q)w', \Psi^{-1} \otimes \Omega^{-1}) = \\ &= \left|\frac{\Psi}{2\pi}\right|^{0.5n} \left|\frac{\Omega}{2\pi}\right|^{0.5m} \exp\{-0.5tr([P-Qw']\Omega[P-Qw']')\} \prod_{i=1}^{m}\prod_{j=1}^{n} p_{i|j}^{-1} \end{aligned} \tag{5}$$

where tr denotes the matrix trace and

$$P = \Psi^{0.5}Ln(p),\ Q = \Psi^{0.5}Ln(q). \tag{6}$$

Remarks

1. The factorization unicity of the position-parameter matrix Qw' will be enforced by the obvious normalization requirement $\sum_{i=1}^{m} q_i = 1$.

2. The introduced unknown precision matrix Ω models dependences among particular opinions and Ψ relates the possibility of erroneous guess to the pre-prior prevision $\bar{p}$.

4.3 Prior p.d.f. of q, w, Ω

The commonly acceptable prior knowledge has been reflected in the hyper-parametrized model. Consequently, the constructed prior p.d.f. has to be very flat.

The key q element is assigned the uniform prior distribution restricted on the simplex

$$\sum_{i=1}^{m} q_i = 1, \;\; q_i > 0. \tag{7}$$

For the w constituent, the uniform (improper) prior p.d.f. on n-dimensional real space can be and will be chosen.

In the "Ω-direction", the p.d.f. $f(\Omega) \propto \exp[-0.5\varepsilon tr(\Omega)]$ with a small positive ε is selected. It defines a flat distribution with a slightly higher expectation of independent opinions than any other type dependence.

5 Pooling result

For constructing the required pooling rule, the nuisance parameters Ω, w have to be excluded. It is done in two steps in the subsequent section. Then, maximum a posteriori probability estimate is found and taken as optimal pooling result.

5.1 The parametrized model $f(p|q)$

Proposition 1 [*The reduced hyperparametrized model* $f(p|q,w)$] The model (5) and prior p.d.f. $p(\Omega) \propto \exp[-0.5\varepsilon tr(\Omega)]$ gives

$$f(p|q,w) \propto |\varepsilon I + [P - Qw']'[P - Qw']|^{-0.5(n+m+1)} \prod_{i=1}^{m} \prod_{j=1}^{n} p_{i|j}^{-1} \tag{8}$$

where P, Q are related to p, q through (6).

Proof. It is implied by the relation between the joint p.d.f. $f(p,\Omega|q,w) \propto$

$f(p|q,w,\Omega)f(\Omega)$ and the marginal p.d.f. $f(p|q,w) = \int f(p,\Omega|q,w)\,d\Omega$. The integration is based on a straightforward substitution for Wishart distribution, e.g. [11].

$\square$

Proposition 2 [*The parametrized model $f(p|q)$*] The model (8) and improper uniform prior of w on n-dimensional real space lead to

$$f(p|q) \propto |U|^{-0.5(m+n)} \left(1 - \frac{Q'PU^{-1}P'Q}{Q'Q}\right)^{-0.5(m+n)} (Q'Q)^{-0.5n} \prod_{i=1}^{m}\prod_{j=1}^{n} p_{i|j}^{-1} \quad (9)$$

with q in the set (7) and with

$$U = \varepsilon I + P'P. \quad (10)$$

Proof. It is implied by the relation between the joint p.d.f. $f(p,w|q) \propto f(p|q,w)f(w)$ and the marginal p.d.f. $f(p|q) = \int f(p,w|q)\,dw$. The integration relies on a simple substitution found after completing squares of the determinant argument in (8) with respect to w.

$\square$

5.2 MAP pooling

The posterior expectation $\mathcal{E}(q|p)$ is the best estimate of q, the best pooling, for a wide range of optimality criteria. The complexity of the $f(p|q)$ form calls for a feasible approximation: the maximum a posteriori probability estimate is selected.

Proposition 3 [*The MAP estimate of q*] The normalized posterior negative log-likelihood $\mathcal{L}(q|p)$ of q restricted to the set (7) takes the form

$$\mathcal{L}(q|p) = -2m^{-1}\log(f(p|q)) + c_1 = log(1 - \frac{Q'PU^{-1}P'Q}{Q'Q}) + c_2 \quad (11)$$

where c_1, c_2 do not depend on q. The MAP estimate $\hat{q}$ of q is determined by the vector $\hat{Q} = \Psi^{0.5}Ln(\hat{q})$ of the form

$$\hat{Q} = P\nu \quad (12)$$

where ν is the eigenvector of $P'P$ corresponding to its largest eigenvalue and normalized so that $\hat{q}$ sums to unity.

The eigenvalue is simple and entries of the eigenvector are positive.

Proof. Taking into account that the condition (7) determines the value $Q'Q$ uniquely, the form of the likelihood follows directly from (9). The needed maximum of $\mathcal{L}(q|p)$ has to lie obviously in the space spanned over P columns, thus it has to have the form (12). The optimality is implied by extremal properties of eigenvectors.

The inspected matrix is obviously symmetric positive definite. Due to normalization of $p_{i|j} \leq 1$ it has positive entries. Perron-Frobenius's theorem [10] then implies rest of the properties.

$\square$

Note that the result is independent of the specific value of ε. It, however, has helped to distinguish among eigenvectors of $P'P$: for $\varepsilon = 0$ the maximum is reached for any non-zero eigenvector of $P'P$.

For reference purposes, the gained result is "translated" into the pooling rule written in an explicit continuous-space form.

Proposition 4 [*The continuous MAP pooling*] For the continuous case characterized by the positive p.d.f.s $f_j(x)$, the MAP pooling takes the form

$$\hat{f}_0(x) = \prod_{j=1}^{n} [f_j(x)]^{\nu_j} \tag{13}$$

where $\nu = [\nu_1, \ldots, \nu_n]'$ is the eigenvector corresponding to the maximum eigenvalue of the symmetric positive semidefinite (n, n) matrix F with entries

$$F_{kl} = \int \ln(f_k(x)) \ln(f_l(x)) \bar{f}(x)\, dx\,, \;\; k, l = 1, \ldots, n. \tag{14}$$

The norm of ν has to be chosen so that $\int f_0(x)\, dx = 1$.

If the matrix F has positive elements the eigenvalue is simple and entries of the eigenvector are positive.

Proof. Essentially, the previous proposition is translated using the original notations and limiting the upper bound on X_i volumes (δ) to zero and consequently m to infinity.

$\square$

Remarks

1. The found pooling rule coincides with the so called weighted logarithmic pooling rule. The construction of the weights is the practical novelty of the approach. The weights are derived through the analysis of mutual relations between the particular opinions.

2. Any positive multiple of F leads to the same q. This enables us to use any representant in this equivalence class.

3. The symbol $\hat{}$ stresses that the pooling rule provides an *estimate* of the ideal pool.

4. Through the transition to the limiting continuous case the restriction $p_{i|j} \leq 1$ and consequently positivity of F entries may be lost. Then unicity and positivity of ν is not guaranteed without assuming $F_{kl} > 0$. This assumption is often fulfilled, as the particular opinions are assumed "positively correlated" with the ideal pool and, consequently, they are mutually positively correlated, too.

6 Application to ARX model

A large class of adaptive systems is based on estimation of the Gaussian ARX model. The estimation reduces to recursive least squares (RLS) and the solved pooling problem may be used for the construction of their initial conditions, whenever valuable expert knowledge is available. This problem was solved in previous papers [4, 6] by using quite different technique. It is worthwhile to compare the qualitative properties of the proposed method with the former plausible results.

6.1 ARX normal model

The ARX normal model with a single output $y(t)$ [8], possibly excited by an input $u(t)$ fulfilling natural conditions of control [11], specifies the joint distribution of data $D = (y(1), \ldots, y(T), u(1), \ldots, u(T))$, $T < \infty$ as follows

$$f(D|\theta, r) \propto \prod_{t=1}^{T} \mathcal{N}_{y(t)}(\theta' \phi(t), r) \tag{15}$$

where $\mathcal{N}_y(\hat{y}, r)$ is the Gaussian p.d.f. of y with the mean $\hat{y}$ and the dispersion r; θ is an i_θ-dimensional vector of unknown regression coefficients; $\phi(t)$ is an i_θ-dimensional regression vector (a known function of $y(1), y(2), \ldots, y(t-1), u(1), u(2), \ldots, u(t)$). The regression coefficients θ and the dispersion r parametrize the model.

6.2 Pooling problem

Using self-reproducing Gauss-inverse-Wishart (GiW) prior p.d.f. [11] for describing the unknown parameter θ, r, Bayesian estimation reduces to evaluation of its finite-dimensional statistics which formally coincide with the quantities evaluated in recursive least squares. The prior p.d.f. just introduces non-trivial initial conditions to them. Their careful choice has been advocated repeatedly [7]. The same prior p.d.f. can be and should be used as the reference point for the advanced technique of restricted forgetting [9].

The choice can be based on various informational sources, various expert opinions. They can be gained by using measured, simulated or fictitious data for correcting a pre-prior flat distribution [2]. The sources, however, need not be fully compatible and/or may repeat the same piece of information. Thus, their pooling should be done according to the general theory presented.

Analytic evaluation of the key matrix F (14) is, however, impossible for GiW p.d.f., i.e. only a combination of information pieces on regression coefficients ($x = \theta$) is analytically feasible. Luckily enough, it does not restrict applicability, as the detailed expert opinions concern mostly the unknown regression coefficients θ. Just a rough guess $\hat{r}$ of r is provided.

6.3 Form of expert opinions

For a given r, self-reproducing p.d.f.s are normal and their statistics coincide with the least squares estimate of θ and corresponding precision matrix (inversion of covariance matrix). Pooling should combine the opinions

$$f_j(\theta|r) = \mathcal{N}_\theta(\theta(j), \omega^{-1}(j)) \tag{16}$$

parametrized by the mean $\theta(j)$ and precision matrix $\omega(j)$.

The following flat pre-prior p.d.f. is chosen

$$\bar{f}(\theta|r) = \mathcal{N}_\theta(0, \epsilon^{-1} I), \; I = \text{unit matrix}, \; \epsilon > 0 \text{ small.} \tag{17}$$

6.4 Evaluation of weights of the pooling rule

Proposition 5 [*The matrix F for the ARX model*] For the opinions (16) and the pre-prior p.d.f. (17) the matrix F (14) has the entries

$$F_{kl} = M_0(k)M_0(l) - M_0(k)M_1(l) - M_0(l)M_1(k) + M_2(k,l) \tag{18}$$

with

$$M_0(k) = \ln\left(\left|\frac{\omega(k)}{2\pi}\right|\right) \tag{19}$$

$$M_1(k) = \theta'(k)\omega(k)\theta(k) + \frac{tr(\omega(k))}{\epsilon} \tag{20}$$

and

$$
\begin{aligned}
& M_2(k,l) = \epsilon^{-2}\left\{tr[\omega(k)]tr[\omega(l)] + 2tr[\omega(k)\omega(l)]\right\} + \qquad (21)\\
+ \quad & \epsilon^{-1}[\theta'(l)\omega(l)\theta(l)tr(\omega(k)) + \theta'(k)\omega(k)\theta(k)tr(\omega(l)) + 4\theta'(l)\omega(l)\omega(k)\theta(k)] + \\
+ \quad & \theta'(k)\omega(k)\theta(k)\theta'(l)\omega(l)\theta(l).
\end{aligned}
$$

For ϵ approaching zero, the eigenvectors of F will coincide with those of the "dominant" ${}^0F$ part of F, namely,

$$
{}^0F_{kl} = tr[\omega(k)]tr[\omega(l)] + 2tr[\omega(k)\omega(l)], \quad k,l = 1,\ldots,n. \qquad (22)
$$

Proof. By the definition (14) $F_{kl} = \bar{\mathcal{E}}[\ln(f_k(\theta))\ln(f_l(\theta))]$ where the expectation $\bar{\mathcal{E}}$ is taken with respect to the pre-prior p.d.f. $\bar{f}(\theta)$ (17). For the opinions (16) it reads

$$
F_{kl} = M_0(k)M_0(l) - M_0(k)M_1(l) - M_0(l)M_1(k) + M_2(k,l)
$$

(a constant factor is omitted, see Remark 2 in 5.2) where $M_0(k)$ is given by (19) and

$$
M_1(k) = \bar{\mathcal{E}}[(\theta-\theta(k))'\omega(k)(\theta-\theta(k))] = \theta'(k)\omega(k)\theta(k) + \frac{tr(\omega(k))}{\epsilon}
$$

$$
M_2(k,l) = \bar{\mathcal{E}}[(\theta-\theta(k))'\omega(k)(\theta-\theta(k))(\theta-\theta(l))'\omega(l)(\theta-\theta(l))].
$$

Using symmetry of the precision matrices $\omega(\cdot)$, the fact that odd moments of $\bar{f}(\theta)$ are zero and the form $\epsilon^{-1}I$ of the corresponding covariance, the following identity is found

$$
\begin{aligned}
M_2(k,l) &= \bar{\mathcal{E}}[\theta'\omega(k)\theta\theta'\omega(l)\theta] + \epsilon^{-1}[\theta'(l)\omega(l)\theta(l)tr(\omega(k)) + \\
&+ \theta'(k)\omega(k)\theta(k)tr(\omega(l)) + 4\theta'(l)\omega(l)\omega(k)\theta(k)] + \\
&+ \theta'(k)\omega(k)\theta(k)\theta'(l)\omega(l)\theta(l).
\end{aligned}
$$

Lengthy but straightforward computations give the "dominating" term

$$
\bar{\mathcal{E}}[\theta'\omega(k)\theta\theta'\omega(l)\theta] = \epsilon^{-2}\left\{tr[\omega(k)]tr[\omega(l)] + 2tr[\omega(k)\omega(l)]\right\}
$$

and conclude the evaluations. The limit behaviour is obvious.

$\square$

The evaluation of the eigenvector ν corresponding to the maximum eigenvalue of F is a standard task. The iterative procedure generating for $k = 1,2,\ldots$ the approximations $\nu^{(k)}$ of ν

$$
\nu^{(k)} = \frac{F\nu^{(k-1)}}{\sum_{i=1}^n \nu_i^{(j)}}, \quad \nu^{(0)} > 0 \qquad (23)
$$

seems to be well adapted to the known properties of the matrix F. Thus, it remains to determine its length in order to get normalized p.d.f. $f_0(\theta)$.

Proposition 6 [*Pooling the opinions on regression coefficients*] The eigenvector ν, corresponding to the (unique) maximum eigenvalue of ${}^0F$ (22) can be always chosen to have positive entries.

Let the eigenvector ν lie on the line with the direction vector ν^1 which sums to unity, i.e. $\nu = \alpha\nu^1$ for some positive α.

Let $\hat{\theta}$ be i_θ-vector, $\hat{\omega}$ positive semidefinite matrix and real λ such that

$$\sum_{j=1}^{n}[(\theta-\theta(j))'\omega(j)(\theta-\theta(j)) - \ln(|\omega(j)/2\pi|)]\nu_j^1 = (\theta-\hat{\theta})'\hat{\omega}(\theta-\hat{\theta}) + \lambda, \quad (24)$$

i.e. left hand side is completed to squares with respect to θ.

Then, the appropriate normalization is given by $\hat{\alpha} \in (0,1]$ (it always exists) which solves the equation

$$\ln(\alpha) + \alpha\lambda/i_\theta = \ln(2\pi) - \ln(|\hat{\omega}|)/i_\theta \quad (25)$$

and the optimal opinion pooling is the $\mathcal{N}_\theta(\hat{\theta}, [\hat{\alpha}\hat{\omega}]^{-1})$.

Proof. The unicity of the eigenvalue ν and positivity of its entries is implied by the general theory relying upon the positivity of F-entries [10].

Application of pooling rule (1) and completion of squares with respect to θ gives the form of the exponent of the resulting p.d.f. The normalization condition (Proposition 4) leads to the equation for α. Existence of the solution of (25) is implied by the continuous dependence of the normalizing integral on α and by the asymptotic values for $\alpha = 0$ and $\alpha = 1$:

– For $\alpha = 0$ the pooling leads to uniform distribution with infinite integral.

– For $\alpha = 1$, the inequality between geometric and arithmetic mean implies $\prod_{j=1}^{n}[f_j(\theta)]^{\nu_j^1} \leq \sum_{j=1}^{n}\nu_j^1 f_j(\theta)$, i.e. the inspected integral does not exceed unity.

Thus, for some $\alpha \in (0,1]$ the normalization condition can be met.

□

The following simple example indicates that this solution has the same intuitively appealing properties as the former solution [7] has. The current solution is more generally applicable and flexible: unlike previous solutions it assigns individual weights to particular opinions.

Example [*Pooling of position parameter guesses*] Let $n = 2$, $i_\theta = 1$, $\omega(1) = \omega(2) = 1$. Then, ${}^0F = const. \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\nu^1 = [0.5, 0.5]'$, $\hat{\theta} = 0.5(\theta(1)+\theta(2))$ and $\hat{\omega} = 1$. The corresponding $\hat{\alpha}$ solves the equation

$$\ln(\alpha) + \alpha[0.25(\theta(1)-\theta(2))^2 + \ln(2\pi)] = \ln(2\pi).$$

This equation gives $\hat{\alpha} = 1$ for $\theta(1) = \theta(2)$, i.e. the identical opinions were taken as a single one. For $\theta(1) \neq \theta(2)$, $\hat{\alpha}$ decreases, i.e. resulting dispersion increases, with increasing distance of $(\theta(1) - \theta(2))^2$. Several results are presented in the following table.

opinion difference $0.5(\theta(1) - \theta(2))$	**norm of** ν $\hat{\alpha}$	**standard deviation** $(\hat{\alpha}\hat{\omega})^{-0.5}$	$\frac{\textbf{opinion difference}}{\textbf{standard deviation}}$
0	1.000000	1.000	0.000
5	0.141100	2.660	1.888
10	0.047800	4.570	2.188
50	0.003050	18.107	2.761
500	0.000047	145.562	3.435
2 000	0.000004	527.519	3.791

6.5 Completion to $\Theta = (\theta, r)$

Let us assume that all precision matrices in the previous paragraph take the form

$$\omega(j) = \tilde{\omega}(j)/r$$

for some r-independent $\tilde{\omega}(j)$. Then, the vector ν^1 found from ${}^0F$ approximation of F is independent of r (cf. (22)). The resulting point estimate of the regression coefficients θ is independent of r, too. Let us solve the equation (25) for determining the norm α with $\omega(j)$ replaced by $\tilde{\omega}(j)$ and $\hat{\omega}$ by the corresponding $\hat{\tilde{\omega}}$. It can be found that the resulting pooling has the form

$$\hat{f}_0(\theta|r) = \mathcal{N}_\theta(\hat{\theta}, r[\hat{\alpha}\hat{\tilde{\omega}}]^{-1})\tilde{f}(r)$$

where the factor $\tilde{f}(r)$ arising from the imposibility to fulfil (25) for all r has inverse Wishart form. Thus, it can be associated the inverse Wishart part of the GiW distribution in such a way that the prior knowledge on sole r is respected. In other words: for unknown r we proceed as if r would be known and $\omega(j)$ scaled for $r = 1$. Then, the prior estimate $\hat{r}$ of r, and the number of degrees of freedom which determines confidence in it is simply added to the initial conditions gained by pooling.

7 Application to Markov chains

This section serves for

- illustrating that the proposed pooling extends applicability area comparing to previous attempts,

- addressing other very important class of models which are practically useful,
- demonstrating that the methodology can be exploited to unexpected purposes; here, to a reduction of dimensionality.

7.1 Markov chain

Markov chain is a probabilistic model describing the relationships of discrete valued output $y(t) \in \{i\}_{i=1}^{m_y}$ to the discrete valued input $u(t) \in \{i\}_{i=1}^{m_u}$ and to the finite past input-output history stored in a discrete valued "regression vector" (observable vector state) $\phi(t)$. A more precise label controlled Markov chain of a higher order is dropped for simplicity.

The joint distribution of the data $D = (y(1), \ldots, y(T), u(1), \ldots, u(T))$, $T < \infty$ observed within time range $[1, T]$ is given by this model as follows

$$f(D|\Theta) \propto \prod_{t=1}^{T} f(y(t)|\phi(t), \Theta) = \prod_{i=1}^{m_y} \prod_{\zeta} \Theta_{i|\zeta}^{s_{i|\zeta}(T)-1} \tag{26}$$

where

$\Theta = [\Theta_{i|\zeta} = p(y(t) = i|\phi(t) = \zeta, \Theta)]$ is an unknown parameter consisting of transition probabilities, restricted by obvious conditions

$$\Theta_{i|\zeta} \geq 0, \quad \sum_{i=1}^{m_y} \Theta_{i|\zeta} = 1 \quad \forall \zeta;$$

ζ is a multi-index from a discrete set of possible ϕ-values;

$s_{i|\zeta}(t)$ is the number of the observed situations $y(\tau) = i, \phi(\tau) = \zeta$ for $\tau \in \{1, 2, \ldots, t\}$ increased by 1.

7.2 Pooling problem

Markov chains describe a very general class of dynamic systems which are naturally discrete or treated by a digital technology. The related estimation is very easy: the sufficient statistic $s(t) = [s_{i|\zeta}(t)]$ is finite-dimensional, thus there is a reproducing prior p.d.f. (of Dirichlet type) which adds just non-trivial initial conditions to the statistic and guarantees that the posterior p.d.f. is again of Dirichlet type

$$p(\Theta|y(1), \ldots, y(t), u(1), \ldots, u(t)) \propto \prod_{i=1}^{m_y} \prod_{\zeta} \Theta_{i|\zeta}^{s_{i|\zeta}(t)-1}$$

where $s_{i|\zeta}(t)$ is number of the observed situations $y(\tau) = i, \phi(\tau) = \zeta,\ \tau = 1, 2, \ldots, t$ increased by a number $s_{i|\zeta}(0) > 0$. Thus, the estimation consists just of this simple counting.

Dimensionality is the main drawback of the model. It restricts its applicability, as the dimension of the sufficient statistic $s(t)$ is extremely large even for medium dimension of ϕ and small cardinality of the data-value sets. Consequently, huge amount of data is needed for obtaining reliable estimates and, moreover, large computer memory has to be available and managed.

The first drawback can be weakened by using prior information in the same way as in ARX case. The need to cure both of them, however, leads to necessity to approximate the model. Just this aspect will be addressed (not precluding the former one). The following technique is proposed here:

- several simplified Markov chain type models are identified on the same data set;
- predictors gained from the simplified models are taken as particular opinions and the ideal predictor is estimated using the general pooling theory.

Notice, that the predicted output value is to play the role of the quantity x: the theory cannot be applied to model parameters as they are of a very different nature and have no straightforward relation to the parameters of the full model (26).

7.3 Form of expert opinions

The original parametrized model is supposed to be reduced to n simpler models

$$p(y(t)|\phi_j(t), \Theta_j), \quad j = 1, \ldots, n \tag{27}$$

where $\phi_j(t)$ are simplified regression vectors (typically containing a selection of $\phi(t)$ entries) and unknown parameters Θ_j are the corresponding transition probabilities $\Theta_{i|\zeta_j}$ from $\phi_j(t) = \zeta_j$ to $y(t) = i$.

Use of reproducing Dirichlet priors for estimating each Θ_j leads to the posterior Dirichlet distribution which provides the following prediction for the subsequent output

$$\begin{aligned} p(y(t) = i|\phi_j(t) = \zeta_j(y(1), \ldots, y(t-1), u(1), \ldots, u(t-1)) = \\ = \frac{s_{i|\zeta_j}(t-1)}{\sum_{i=1}^{m_y} s_{i|\zeta_j}(t-1)} \end{aligned} \tag{28}$$

where $s_{i|\zeta_j}(t)$ is the number of the observed situations $y(\tau) = i, \phi_j(\tau) = \zeta_j,\ \tau = 1, 2, \ldots, t$ increased by a number $s_{i|\zeta_j}(0) > 0$ for $j = 1, \ldots, n$.

For a fixed t, fixed past history $y(1),\ldots,y(t-1), u(1),\ldots,u(t-1)$ and given $\phi(t)=\zeta$ which determines the simplified regressors $\phi_j(t)=\zeta_j$, the predicted value $y(t)$ plays the role of the unknown quantity x; the predictors (28) become opinion dependent models for $y(t)\equiv x$, i.e.

$$f_j(y(t)=i) = f_j(y(t)=i|\phi_j=\zeta_j) = \frac{s_{i|\zeta_j}(t-1)}{\sum_{i=1}^{m_y} s_{i|\zeta_j}(t-1)}. \qquad (29)$$

The pre-prior distribution describes the discrete variable $y(t)=x$, typically $\bar{p}_i = 1/m_y$. The searched for ideal pool is then the predictor

$$q_i = f(y(t)=i|\phi(t)=\zeta, y(1),\ldots,y(t-1),u(1),\ldots,u(t-1))$$

.

7.4 Evaluation of weights of the pooling rule

Proposition 7 [*The matrix F for the Markov model based predictors*] For a fixed t, fixed past history $y(1),\ldots,y(t-1),u(1),\ldots,u(t-1)$ and given $\phi(t)$ the opinions are assumed to be given by (29) with arbitrary pre-prior p.f. $\bar{p}$. Then the matrix F (14) has the entries

$$F_{kl} \quad = \quad \sum_{i=1}^{m_y} \ln\Big(\frac{s_{i|\zeta_k}(t-1))}{\sum_{i=1}^{m_y} s_{i|\zeta_k}(t-1))}\Big)\ln\Big(\frac{s_{i|\zeta_l}(t-1))}{\sum_{i=1}^{m_y} s_{i|\zeta_l}(t-1))}\Big)\bar{p}_i \qquad (30)$$

Proof. Implied by definitions of the involved elements and by (29).

□

Proposition 8 [*Prediction based approximation of Markov chain*] Let ν^1 be the (unique) eigenvector corresponding to maximum eigenvalue of the matrix (30) (with positive entries and unit sum). Then, the pooling based approximant of the ideal predictor at time t takes the form

$$\hat{f}_0(y(t)=i|\phi(t)=\zeta) = \prod_{j=1}^{n}\left(\frac{s_{i|\zeta_j}(t-1)}{\sum_{i=1}^{m_y} s_{i|\zeta_j}(t-1)}\right)^{\alpha\nu_j^1} \qquad (31)$$

where α solves the normalizing condition

$$\sum_{i=1}^{m_y}\prod_{j=1}^{n}\left(\frac{s_{i|\zeta_j}(t-1)}{\sum_{i=1}^{m_y} s_{i|\zeta_j}(t-1)}\right)^{\alpha\nu_j^1} = 1. \qquad (32)$$

Proof. A direct application of the general theory and of the problem formulation.

□

Example [*Approximation of a Markov chain*] Let $y(t) \in \{0,1\}$, no input is present, $\phi'(t) = [y(t-1), y(t-2)]$, i.e. "autoregression" of the order 2 is assumed as a complete model to be approximated. Two approximate Markov models are used, specified by $\phi_1(t) = y(t-1)$ (first order autoregression) and $\phi_2(t) = y(t-2)$ (autoregression with respect to the output delayed by two periods).

Full model is simulated with transition probabilities:

(ζ_1, ζ_2)	$\Theta_{1\mid\zeta_1,\zeta_2}$
$(0,0)$	0.60
$(0,1)$	0.50
$(1,0)$	0.40
$(1,1)$	0.40

The following tables contain results obtained over the time span 10000. The behaviours of three predictors

M0 given by simulated parameters (the "true" one);
M1 resulting from Bayesian estimation (the optimal one);
M2 approximate predictor (the proposed one)
are compared.

The following performance index measuring prediction capabilities is used for their comparison

$$S_\beta(\zeta) = S_{M\beta}(\zeta_1, \zeta_2) = \frac{1}{T} \sum_{1 \le \tau \le T, \phi(\tau) = \zeta} |y(\tau) - \mathcal{E}_{M\beta}[y(\tau)|\zeta]| =$$

$$= \frac{1}{T} \sum_{1 \le \tau \le T, \phi(\tau) = \zeta} |y(\tau) - f_{M\beta}(y(\tau) = 1|\zeta)|, \quad M\beta = \mathbf{M0}, \mathbf{M1}, \mathbf{M2}$$

Model M0: predictive abilities of the true model

T	$S_0(0,0)$	$S_0(0,1)$	$S_0(1,0)$	$S_0(1,1)$
10	0.1400	0.1500	0.1400	0.0300
50	0.1760	0.1300	0.1200	0.0560
100	0.1420	0.1400	0.1360	0.0580
500	0.1220	0.1470	0.1408	0.0728
1000	0.1020	0.1370	0.1352	0.1138
5000	0.1127	0.1416	0.1360	0.0970
10000	0.1130	0.1445	0.1386	0.0888

Model M1: predictive abilities of the Bayesian predictor

T	$S_1(0,0)$	$S_1(0,1)$	$S_1(1,0)$	$S_1(1,1)$
10	0.1368	0.1498	0.1422	0.0390
50	0.1747	0.1297	0.1222	0.0580
100	0.1410	0.1400	0.1380	0.0634
500	0.1207	0.1470	0.1424	0.0722
1000	0.0999	0.1370	0.1357	0.1152
5000	0.1113	0.1416	0.1372	0.0968
10000	0.1115	0.1445	0.1399	0.0883

Model M2: predictive abilities of the approximate model

T	$S_2(0,0)$	$S_2(0,1)$	$S_2(1,0)$	$S_2(1,1)$
10	0.1497	0.1517	0.1417	0.0440
50	0.1798	0.1324	0.1217	0.0576
100	0.1449	0.1403	0.1375	0.0664
500	0.1259	0.1469	0.1420	0.0753
1000	0.1083	0.1371	0.1356	0.1160
5000	0.1169	0.1415	0.1370	0.0981
10000	0.1178	0.1444	0.1395	0.0905

The obtained results are encouraging: approximation errors are sometimes even hidden under random fluctuations.

□

8 Conclusions

In the paper, pooling prior information in technical systems is addressed within modelling-Bayesian-estimation framework. The method, which may be used also in the artificial intelligence area, has been developed for initializing parameter estimates in model-based adaptive control (the expert knowledge implementation improves substantially its start). *Wherever there is a lack of data, the expert knowledge can substitute it* efficiently and purposefully.

The proposed methodology is a clear step in quite promising direction which re-formulates and generalizes previous attempts. The application to ARX model indicates it clearly. In addition to obvious applicability to other types of models (like Markov ones), the proposed technique suits to construction of a complex model from a set of simpler ones. This was demonstrated on the important task of approximating high-dimensional Markov chain. As the presented examples indicate, this "by-product" might be quite useful.

The further development will be oriented to:

- Elaboration of the test cases up to all algorithmic details;
- Inspection of alternative modelling of expert mutual correlations which would suit to model classes where mutual absolute continuity is violated (models in which unknown parameters specify the model support);
- Search for connections of particular knowledge integrations (e.g. various probabilistic/fuzzy-type rules) to the particular hyperparametrized models (e.g. uniform type).

References

[1] Berger J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

[2] Böhm J., M. Kárný (1992). Merging of user's knowledge into self-tuning controllers. *4th IFAC Symposium Adaptive Control and Signal Processing*, Grenoble, 427-432.

[3] Csaki Z. (1992). Identification as learning in reasoning about incomplete knowledge of physical systems. *Proceedings ECAI 92*, John Wiley & Sons, 715-718.

[4] Kárný M. (1984). Quantification of prior knowledge about global characteristics of linear normal model. *Kybernetika*, **20**, 5, 376-385.

[5] Kárný M., J. Böhm (1990). Probabilistic modelling of imprecisely known systems for robust LQ design. *Preprints 1st ECC*, Grenoble, 426-431.

[6] Kárný M. (1991). A note on feeding uncertain knowledge into recursive least squares. *Preprints 30th CDC*, Brighton, U.K., **1**, 975-976.

[7] Kárný M. (1992). Soft prior information for recursive least squares. Submitted to *IEEE Tran. SMC.*

[8] Kárný M. (1993). Parametrization of multi-output multi-input autoregressive-regressive models for self-tuning control. *Kybernetika*, **28**.

[9] Kulhavý R. (1987). Restricted exponential forgetting in real-time identification. *Automatica*, **23**, 5, 598-600.

[10] Fiedler M. (1981). Special matrices and their use in numerical mathematics (in Czech). SNTL, Prague.

[11] Peterka V. (1981). Bayesian system identification. In: Eykhoff P. (ed.) *Trends and Progress in System Identification*, Pergamon Press Oxford, 239-304.