Recursive Estimation of Hidden Markov Models

László Gerencsér
Gábor Molnár-Sáska
György Michaletzky

Stochastic Systems Research Group
MTA SZTAKI, Budapest

Morgan Stanley Workshop on Quantitative and Mathematical Finance

20-21 October 2005
Budapest

OUTLINE

- Simple estimation problems

- HMM's

- Exponential stability of the filter

- Invariant measures

- Recursive estimation

- Strong approximation

- Continuous-time systems

ESTIMATION, IDENTIFICATION, CALIBRATION

Example 1: AR(1) processes:

$$y_n + ay_{n-1} = e_n, \qquad |a| < 1,$$

where (e_n) is a w.s.s. orthogonal process.

Example 2: A short rate model (Vasicek):

$$dr_t = (a - br_t)dt + \sigma dw_t, \qquad b > 0.$$

The problem: estimate, identify, calibrate the unknown parameters of the model.

Both examples are linear stochastic systems (LSS).

FROM LINEAR TO NON-LINEAR

Example 1.: Quantized-AR-processes:

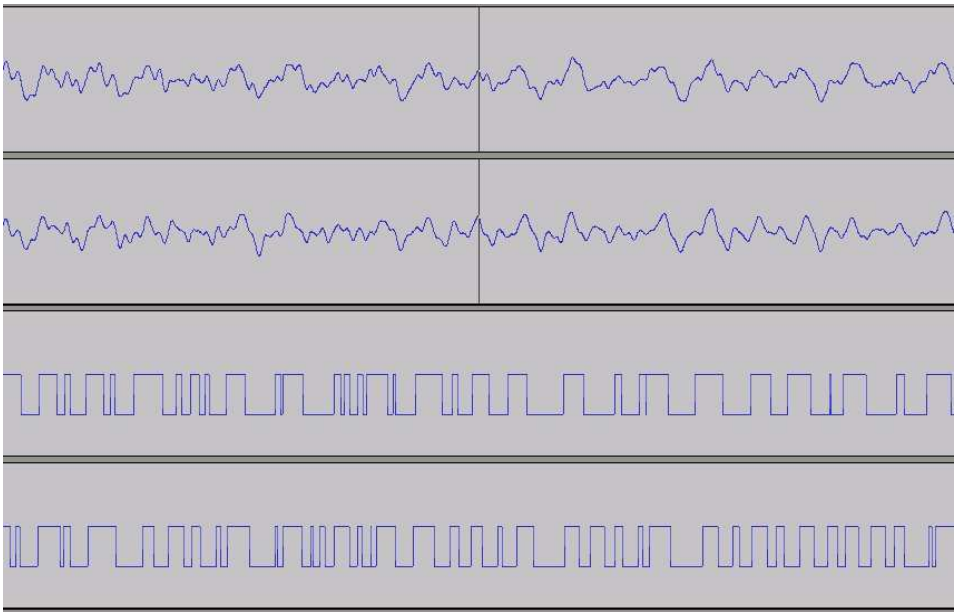
$$x_n + ax_{n-1} = e_n, \quad |a| < 1,$$

$$Y_n = \text{sgn}\,(X_n).$$

Example 2.: A short rate model (CIR):

$$dr_t = (a - br_t)dt + \sigma\sqrt{r_t}dw_t, \quad b > 0.$$

QUANTIZED SIGNALS



HIDDEN MARKOV PROCESSES

Definition (X_n, Y_n) is a hidden Markov process if

(i) (X_n) is a homogenous Markov chain on $\mathcal{X}$

(ii) $Y_n = h(X_n, V_n), \quad Y_n \in \mathcal{Y}$

where (V_n) is an i.i.d. sequence and h is Borel-measurable.

Here $\mathcal{X}, \mathcal{Y}$ are separable metric spaces (Polish spaces). (X_n) is the state, (Y_n) is the read-out.

It follows that

$$P(Y_n \dots Y_0 | X_n \dots X_0) = \prod_{i=0}^{n} P(Y_i | X_i).$$

EXAMPLE: GAUSSIAN MIXTURE

Let $\mathcal{X} = \{1, 2\}$, let (X_n) be i.i.d. and let

$$Y_n = h(X_n) + \sigma(X_n)V_n,$$

where (V_n) is Gaussian i.i.d.

HIDDEN MARKOV MODELS (HMM)

The setup: Primary interest in cases when $\mathcal{X}$ is discrete and $\mathcal{Y}$ is continuous.

The transition probabilities are

$$Q(x, x') = P(X_{n+1} = x' | X_n = x),$$

the read-out probabilities are

$$b^x(y) = P(Y_n = y | X_n = x).$$

Condition 1: $Q(x, x') > 0$ for all x, x'.

Example: Discretize

$$dx = Axdt + Bdw_t.$$

Condition 2: $b^x(y) > 0$ for all x, y.

Note: Condition 2 excludes quantization.

CONTINUOUS READ-OUTS

The conditional density of Y given $X = x$ is

$$b^x(y) = p(y|x).$$

Condition 2: We assume that $b^x(y) > 0$ for all x, y.

Example:

$$Y_n = h(X_n) + \sigma(X_n)V_n,$$

where (V_n) is a Gaussian white noise.

THE PREDICTIVE FILTER

A key quantity is the predictive filter:

$$p_{n+1}^j = P(X_{n+1} = j | Y_n, \ldots, Y_0). \qquad (1)$$

Proposition: *The filter process satisfies*

$$p_{n+1} = \pi(Q^T B(Y_n) p_n), \qquad (2)$$

where $B(y) = \mathrm{diag} b^x(y)$.

Here π is the normalizing operator: set for $x \geq 0,\ x \neq 0$

$$\pi(x)^i = x^i / \sum_j x^j.$$

Equation (2) is called the Baum-equation.

Reference: Baum and Petrie, Ann. Math. Stat., 1966

ESTIMATION OF HMM-S

Problem: Estimate Q and b from (Y_n).

Parameterization: Let

$$Q = Q(\theta) \quad \text{and} \quad b = b(\theta)$$

with $\theta \in D \subset \mathbb{R}^p$, where D is an open set. The true parameter is θ^*.

THE NATURAL PARAMETRIZATION

Take the entries of Q and b. If $|\mathcal{X}| = m, |\mathcal{Y}| = s$ then

$$\theta = \{q_{ij}, b_{il}\}$$

with

$$j = 1, \ldots, m-1 \quad \text{and} \quad l = 1, \ldots, s-1,$$

$$\sum_{j=1}^{m-1} q_{ij} < 1, \qquad q_{ij} > 0,$$

$$\sum_{l=1}^{s-1} b_{il} < 1, \qquad b_{il} > 0.$$

THE LOG-LIKELIHOOD FUNCTION I.

The conditional log-likelihood function is

$$L_N(\theta, q) = \log P(y_N, \ldots, y_0; \theta, q),$$

with $P(X_0) = q$.

A key problem: establish that

$$\lim_N \frac{1}{N} L_N(\theta, q)$$

exists w.p.1., uniformly in θ.

Relevant results:

Shannon-McMillan-Breiman theorem.

Extension to divergence rates:

Barron, Annals of Prob., 1985.

A HMM CONTEXT

Leroux: Subadditive ergodic theorem, SPA, 1992.

LeGland and Mevel: Geometric ergodicity, MCSS, 2000.

Gerencsér, Molnár-Sáska, Michaletzky, Tusnády: L-mixing, 2004.

THE LOG-LIKELIHOOD FUNCTION II.

Write $L_N(\theta, q)$ as a sum of the terms

$$\log P(y_n|y_{n-1}, \ldots, y_0; \theta, q).$$

Express this via

$$P(X_n = x|y_{n-1}, \ldots y_0, \theta) = p_n^x(\theta, q).$$

Get

$$\log \sum_x b^x(y_n, \theta) p_n^x(\theta, q).$$

Ultimately we work with

$$g(y, p, \theta) = \log \sum_x b^x(y, \theta) p^x. \qquad (3)$$

THE ASSUMED FILTER

For any assumed θ and q, define $p_n = p_n(\theta, q)$ via the Baum-equation:

$$p_{n+1} = \pi(Q(\theta)^T B(Y_n, \theta) p_n), \qquad (4)$$

with $p_0 = q$.

STABILITY OF THE FILTER I

Question: A sensitivity problem: does $p_n(\theta, q)$ forget its initial condition q?

Proposition: *Assume that* $Q > 0$ *and* $b^x(y) > 0$ *for any* x, y. *Then here exists* $0 < \rho < 1$ *and a deterministic constant* C *such that for any observation sequence* $y = (y_n)$

$$\|p_n(q) - p_n(q')\|_{TV} \leq C\rho^n \|q - q'\|_{TV}. \quad (5)$$

Remark: Note that this is a non-probabilistic statement.

LeGland and Mevel, MCSS, 2000.

FROM POSITIVE TO PRIMITIVE

If Q is only primitive, i.e. $Q^r > 0$ with some positive integer $r > 1$, then (5) holds with a random $C = C(\omega)$:

Proposition: *Assume that Q is primitive and $b^x(y) > 0$ for any x, y. Then*

$$\|p_n(q) - p_n(q')\|_{TV} \leq C(\omega)\rho^n \|q - q'\|_{TV}, \quad (6)$$

where $0 < \rho < 1$, with some random $C(\omega)$.

L-MIXING

Classical mixing concepts: let $\mathcal{G}, \mathcal{H}$ be σ-subalgebras in $(\Omega, \mathcal{F}, \mathbf{P})$. Require that

$$\sup_{A \in \mathcal{G}} |P(A|\mathcal{H}) - P(A)|$$

be small. Alternatively: write $\xi = \chi_A - P(A)$. Then require that

$$E(\xi|\mathcal{H})$$

be small.

L-mixing: let

$$\mathcal{H}^+ \perp \mathcal{H} \quad \text{and} \quad \sigma(\mathcal{H}, \mathcal{H}^+) = \mathcal{F}.$$

Require that

$$\xi - E(\xi|\mathcal{H}^+)$$

be small.

THE EXTENDED PROCESS (X_n, Y_n, p_n)

Theorem: *Let $Q > 0$. Assume that for all $i, j \in \mathcal{X}$ and for all $q \geq 1$ we have that*

$$\int |\log b^j(y, \theta)|^q b^i(y, \theta^*) \lambda(dy) < \infty. \qquad (7)$$

Then the process $g(Y_n, p_n)$ is L-mixing.

An alternative result:

LeGland & Mevel, MCSS, 2000: $g(Y_n, p_n)$ is geometrically ergodic.

Remark:

$$y_n = h(x_n) + \sigma(x_n) V_n$$

is not fully covered by LeGland and Mevel.

EXPONENTIAL STABILITY

Let us write

$$x_n \quad \text{for} \quad (x_n, y_n),$$
$$z_n \quad \text{for} \quad p_n.$$

Write the Baum-equation as

$$z_{n+1} = f(x_n, z_n) \qquad z_0 = \xi.$$

Let the solution be $(z_n(\xi))$. In general $z_n \in \mathcal{Z}$ Banach space and $x_n \in \mathcal{H}$ an abstract set.

Definition: The mapping f is uniformly exponentially stable if there exists $C, \delta > 0$ such that for any (x_n)

$$\|z_n(\xi) - z_n(\xi')\| \leq C(1-\delta)^n \|\xi - \xi'\|. \quad (8)$$

MARKOVIAN SWITCHING

Let

$$Z_{n+1} = f(X_n, Z_n), \quad Z_0 = \xi, \qquad (9)$$

where (X_n) is a Markov chain satisfying the Doeblin-condition.

Proposition: *The Markov chain* $U_n = (X_n, Z_n)$ *has a unique stationary distribution.*

Remark: Writing (9) as

$$z_{n+1} = T_{x_n} z_n$$

the stationary distribution is obtained by setting

$$z_0^* = \lim_k T_{x_{-1}} \circ \cdots \circ T_{x_{-k}} z.$$

OFF-LINE ESTIMATOR

Solve

$$\frac{\partial}{\partial\theta}\log p(y_N,\ldots,y_1;\theta,q)=0$$

for θ. Get $\widehat{\theta}_N$.

Reminder: the conditional log-likelihood function is the sum of terms

$$p(y_n|y_{n-1},\ldots,y_1,\theta,q)=\log\sum_x b^x(y_n,\theta)p_n^x(\theta,q).$$

We had the notation

$$g(y,p,\theta)=\log\sum_x b^x(y,\theta)p^x.$$

THE ASYMPTOTIC LIKELIHOOD FUNCTION

Let

$$\mu = \mu_\theta$$

be the unique invariant measure for (X_n, Y_n, p_n) under θ.

Define the asymptotic log-likelihood function as

$$J(\theta) = E_{\mu_\theta} g(Y_n, p_n, \theta).$$

The asymptotic log-likelihood equation is

$$\frac{\partial}{\partial \theta} J(\theta) = 0.$$

THE GRADIENT OF THE FILTER

Write the Baum-equation generating the assumed filter $p_n = p_n(\theta, q)$ as

$$p_{n+1} = f(y_n, p_n, \theta) \qquad p_0 = q.$$

Then the dynamics for the gradient process is

$$r_{n+1} = f_p r_n + f_\theta \qquad r_0 = 0.$$

Proposition: *There exists a unique invariant measure* $\mu = \mu(\theta)$ *for the Markov process*

$$U_n = (X_n, Y_n, p_n, r_n).$$

Let it be denoted again by $\mu = \mu(\theta)$.

THE ASYMPTOTIC LIKELIHOOD EQUATION

The asymptotic likelihood equation can be written as

$$E_{\mu_\theta} H(U_n, \theta) = 0.$$

Let the dynamics of (U_n) be written as

$$U_{n+1} = F(U_n, \theta, W_n),$$

where (W_n) is i.i.d.

STATE SPACE DESCRIPTION

Proposition. *If (X_n) is a homogeneous Markov process with values in $\mathcal{X}$, then it can be realized as*

$$X_{n+1} = f(X_n, W_n) \qquad X_0 = \xi$$

where (W_n) is an i.i.d. sequence.

Remark. This is a realization in weak sense.

RECURSIVE ESTIMATION

Problem: solve

$$E_{\mu_\theta} H(U_n, \theta) = 0$$

recursively using real-time data.

Benveniste, Metivier and Priouret (BMP), Adaptive algorithms and stochastic approximations, 1990.

It gives the recursion

$$\widehat{\widehat{\theta}}_{n+1} = \widehat{\widehat{\theta}}_n + \frac{1}{n+1} H(\widehat{\widehat{\theta}}_n, U_{n+1}),$$

where

$$U_{n+1} = F(U_n, \theta_n, W_{n+1}).$$

Modification: take a Newton step.

REFERENCES

- Legland and Mevel, IEEE Conference on Decision and Control (CDC), 1995.

- Gerencsér and Molnár-Sáska, IEEE Conference on Decision and Control (CDC), 2005.

OFF-LINE VS. ON-LINE

Proposition: *Assume that* $Q > 0$, $b^x(y) > 0$ *and* θ^* *is in* D, *where* D *is an open subset of* R^p. *Then we have for* $\widehat{\widehat{\theta}}_N$ *generated by a stochastic Newton method*

$$\widehat{\theta}_N - \widehat{\widehat{\theta}}_N = O_M(\frac{\log N}{N}).$$

Gerencsér, System and Control Letters, 1993.
Gerencsér, SIAM J. Control and Optimization, 2005.

FROM POSITIVE TO PRIMITIVE

Reminder: Assume that Q is a primitive matrix. Then

$$\|p_n(q)-p_n(q')\|_{TV} \leq C(\omega)\rho^n\|q-q'\|_{TV}, \quad (10)$$

where $0 < \rho < 1$, with some random $C(\omega)$.

Theorem: *For an arbitrary* $s > 1$ *we can choose* $C(\omega) > 0, 0 < \rho < 1$ *such that*

$$E|C(\omega)|^s < \infty.$$

(Gerencsér, Michaletzky, Molnár-Sáska, 2005.)

CONTINUOUS TIME MODELS

Multivariable linear stochastic systems

$$\dot{y} = H\dot{w},$$

where $\dot{w}$ is a Gaussian white noise. Let $H = H(\theta^*)$. Invert the system and define

$$\dot{\epsilon}(\theta) = H^{-1}(\theta)\dot{y}.$$

The likelihood-equation:

$$V_{\theta T}(\theta) = \int_0^T \dot{\epsilon}_{\theta t}(\theta) d\epsilon_t(\theta) = 0.$$

We can get an exact but uncomputable recursion for $\widehat{\theta}_T$ using Ito-Wentzel formula.

An approximate, computable recursion gives $\widehat{\widehat{\theta}}_T$.

PARTIAL RESULTS

A heuristic proof for

$$\widehat{\theta}_T - \widehat{\widehat{\theta}}_T = O_M\left(\frac{\log T}{T}\right)$$

is given in

Gerencsér, Gyöngy, Michaletzky, IFAC World Congress, 1984.

Rigorous results on recursive estimation:

Levanony, Zeitouni, Schwartz, SPA, 1994.

CONCLUSION

- The main objective: to prove strong approximation results for recursive estimators for HMM-s.

- Key tool: L-mixing and BMP-theory.

- Extensions to primitive transition matrices.

FUTURE POTENTIALS

- Estimation of time-varying system.
- Continuous time recursive estimation.
- Estimation of Markovian switching systems (from static to dynamic read-out).

∞

http://www.sztaki.hu/sztaki/ake/applmath/stoch