MLE in the partially observed setting: the continuous time case

P. Chigansky

The Hebrew University, Israel

2009-03-16/ SAPS VII

Inference in the partially observed setting

The model

$$Y_t = \int_0^t h(X_s, \theta) ds + W_t, \quad t \in [0, T],$$

where

- $\theta \in \Theta \subseteq \mathbb{R}^d$ is the unknown parameter;
- $X = (X_t)_{t \geq 0}$ is a Markov proc., whose law depends on θ;
- $h(\cdot, \cdot)$ is a known function;
- $W = (W_t)_{t \geq 0}$ is the Wiener process, independent of X;

The objective: estimate θ, given trajectory $Y^T = \{Y_t, t \in [0, T]\}$.

Maximum Likelihood Estimation

The likelihood:

$$L_T(Y^T;\theta) = \exp\left\{\int_0^T \pi_t^\theta(h)dY_t - \frac{1}{2}\int_0^T \left(\pi_t^\theta(h)\right)^2 dt\right\}, \ \theta \in \Theta,$$

with

$$\pi_t^\theta(h) := \mathsf{E}_\theta\Big(h(X_t)\Big|Y^t\Big) = \int_0^t h(x,\theta)\pi_t^\theta(dx),$$

where $\pi_t^\theta(\cdot) = \mathsf{P}_\theta(X_t \in \cdot | Y^t)$ is the *filtering* process.

The MLE of θ is any maximizer $\hat{\theta}_T$ of $L_T(Y^T,\theta)$ over Θ

$$L_T(Y^T,\hat{\theta}_T) = \sup_{\theta\in\Theta} L_T(Y^T,\theta).$$

Maximum Likelihood Estimation

The likelihood:

$$L_T(Y^T;\theta) = \exp\left\{\int_0^T \pi_t^\theta(h)dY_t - \frac{1}{2}\int_0^T \left(\pi_t^\theta(h)\right)^2 dt\right\},\ \theta \in \Theta,$$

with

$$\pi_t^\theta(h) := \mathsf{E}_\theta\Big(h(X_t)\Big| Y^t\Big) = \int_0^t h(x,\theta)\pi_t^\theta(dx),$$

where $\pi_t^\theta(\cdot) = \mathsf{P}_\theta(X_t \in \cdot | Y^t)$ is the *filtering* process.

The MLE of θ is any maximizer $\hat{\theta}_T$ of $L_T(Y^T,\theta)$ over Θ

$$L_T(Y^T,\hat{\theta}_T) = \sup_{\theta\in\Theta} L_T(Y^T,\theta).$$

Maximum Likelihood Estimation

How to compute ?

- E-M algorithm (A.Dembo & O.Zeitouni '86, R.Elliott '93)
- Direct maximization (F.Campillo & F.LeGland, '89)
- Monte Carlo methods, etc.

Asymptotic properties as $T \to \infty$:

- consistency
- asymptotic normality
- convergence of moments
- asymptotic efficiency

Maximum Likelihood Estimation

How to compute ?

- E-M algorithm (A.Dembo & O.Zeitouni ’86, R.Elliott ’93)
- Direct maximization (F.Campillo & F.LeGland, ’89)
- Monte Carlo methods, etc.

Asymptotic properties as $T \to \infty$:

- consistency
- asymptotic normality
- convergence of moments
- asymptotic efficiency

Linear Gaussian case

The model: OU stationary signal X + linear observations:

$$\begin{aligned} dX_t &= -a(\theta)X_t dt + b(\theta)dV_t \\ dY_t &= \quad c(\theta)X_t dt + dW_t \end{aligned}$$

A.V. Balakrishnan '73,

Y.Kutoyants '80,
A.Bagchi & V.Borkar '84
G.Kallianpur & R.Selukar '91

A.Broust & M. Kleptsyna '09 (system with fBm noises)

Linear Gaussian case: a typical result

Theorem (Y.Kutoyants, ’04)

Assume [...], then the MLE $\hat{\theta}_T$, $T \to \infty$ is uniformly consistent, asymptotically normal

$$\mathrm{Law}_{\theta_0}\{\sqrt{T}(\hat{\theta}_T - \theta_0)\} \implies N(0, I^{-1}(\theta_0))$$

and the moments of $\sqrt{T}(\hat{\theta}_T - \theta_0)$ converge. Moreover it is asymptotically efficient for polynomial loss functions. □

- [...] stand for appropriate (natural!) identifiability and smoothness conditions on the coefficients
- the Fisher information $I(\theta_0)$ has an *explicit* formula
- $\hat{\theta}_T$ doesn’t have an explicit formula (as in the direct observation case)

Linear Gaussian case: a typical result

Theorem (Y.Kutoyants, '04)

Assume [...], then the MLE $\hat{\theta}_T$, $T \to \infty$ is uniformly consistent, asymptotically normal

$$\mathrm{Law}_{\theta_0}\{\sqrt{T}(\hat{\theta}_T - \theta_0)\} \implies N(0, I^{-1}(\theta_0))$$

and the moments of $\sqrt{T}(\hat{\theta}_T - \theta_0)$ converge. Moreover it is asymptotically efficient for polynomial loss functions. □

- [...] stand for appropriate (natural!) identifiability and smoothness conditions on the coefficients
- the Fisher information $I(\theta_0)$ has an *explicit* formula
- $\hat{\theta}_T$ doesn't have an explicit formula (as in the direct observation case)

Linear Gaussian case: a typical result

Theorem (Y.Kutoyants, '04)

Assume [...], then the MLE $\hat{\theta}_T$, $T \to \infty$ is uniformly consistent, asymptotically normal

$$\mathrm{Law}_{\theta_0}\{\sqrt{T}(\hat{\theta}_T - \theta_0)\} \implies N(0, I^{-1}(\theta_0))$$

and the moments of $\sqrt{T}(\hat{\theta}_T - \theta_0)$ converge. Moreover it is asymptotically efficient for polynomial loss functions. □

- [...] stand for appropriate (natural!) identifiability and smoothness conditions on the coefficients
- the Fisher information $I(\theta_0)$ has an *explicit* formula
- $\hat{\theta}_T$ doesn't have an explicit formula (as in the direct observation case)

Linear Gaussian case: a typical result

Theorem (Y.Kutoyants, '04)

Assume [...], then the MLE $\hat{\theta}_T$, $T \to \infty$ is uniformly consistent, asymptotically normal

$$\mathrm{Law}_{\theta_0}\left\{\sqrt{T}(\hat{\theta}_T - \theta_0)\right\} \implies N\left(0, I^{-1}(\theta_0)\right)$$

and the moments of $\sqrt{T}(\hat{\theta}_T - \theta_0)$ converge. Moreover it is asymptotically efficient for polynomial loss functions. □

- [...] stand for appropriate (natural!) identifiability and smoothness conditions on the coefficients
- the Fisher information $I(\theta_0)$ has an *explicit* formula
- $\hat{\theta}_T$ doesn't have an explicit formula (as in the direct observation case)

Linear Gaussian case: a typical proof

Implementation of the Ibragimov-Khas'minskii method, using the following main features of the linear Gaussian mdoel:

- The filtering distribution $\pi_t^\theta(dx)$ is Gaussian, with the mean and variance

$$m_t(\theta) = \mathrm{E}_\theta(X_t|Y^t), \quad V_t(\theta) = var_\theta(X_t|Y^t)$$

  which satisfy the Kalman-Bucy equations.
- the analysis of the likelihood relies on the Gaussian property of the conditional mean process $m_t(\theta)$ (Cameron-Martin formula, etc.)

Linear Gaussian case: a typical proof

Implementation of the Ibragimov-Khas'minskii method, using the following main features of the linear Gaussian mdoel:

- The filtering distribution $\pi_t^\theta(dx)$ is Gaussian, with the mean and variance

$$m_t(\theta) = \mathsf{E}_\theta(X_t|Y^t), \quad V_t(\theta) = var_\theta(X_t|Y^t)$$

which satisfy the Kalman-Bucy equations.

- the analysis of the likelihood relies on the Gaussian property of the conditional mean process $m_t(\theta)$ (Cameron-Martin formula, etc.)

Linear Gaussian case: a typical proof

Implementation of the Ibragimov-Khas'minskii method, using the following main features of the linear Gaussian mdoel:

- The filtering distribution $\pi_t^\theta(dx)$ is Gaussian, with the mean and variance

$$m_t(\theta) = \mathrm{E}_\theta(X_t|Y^t), \quad V_t(\theta) = var_\theta(X_t|Y^t)$$

  which satisfy the Kalman-Bucy equations.
- the analysis of the likelihood relies on the Gaussian property of the conditional mean process $m_t(\theta)$ (Cameron-Martin formula, etc.)

Finite state case: the model

The signal is a finite state Markov chain on $\{1, ..., d\}$

- initial distribution $\nu_i = \mathrm{P}_\theta(X_0 = i)$, $i = 1, ..., d$
- transition rates matrix $R = \{r_{ij}(\theta)\}$:

$$\mathrm{P}_\theta(X_{t+\delta} = j | X_t = i) = \begin{cases} r_{ij}(\theta)\delta + o(\delta), & i \neq j \\ 1 - \sum_{k \neq i} r_{ik}(\theta)\delta + o(\delta), & i = j \end{cases}$$

- the chain is irreducible

The observation process:

$$Y_t = \int_0^t h(X_s, \theta) ds + W_t, \quad t \in [0, T].$$

Finite state case: the model

The signal is a finite state Markov chain on $\{1, ..., d\}$

- initial distribution $\nu_i = \mathrm{P}_\theta(X_0 = i)$, $i = 1, ..., d$
- transition rates matrix $R = \left\{r_{ij}(\theta)\right\}$:

$$\mathrm{P}_\theta(X_{t+\delta} = j | X_t = i) = \begin{cases} r_{ij}(\theta)\delta + o(\delta), & i \neq j \\ 1 - \sum_{k \neq i} r_{ik}(\theta)\delta + o(\delta), & i = j \end{cases}$$

- the chain is irreducible

The observation process:

$$Y_t = \int_0^t h(X_s, \theta) ds + W_t, \quad t \in [0, T].$$

Finite state case: identifiability

The question of identifiability:

$$\theta_1 \neq \theta_2 \quad \stackrel{?}{\Longrightarrow} \quad \mathrm{Law}_{\theta_1}\{Y^T\} \neq \mathrm{Law}_{\theta_2}\{Y^T\}$$

is equivalent to

$$\theta_1 \neq \theta_2 \quad \stackrel{?}{\Longrightarrow} \quad \mathrm{Law}_{\theta_1}\{h(X^T, \theta_1)\} \neq \mathrm{Law}_{\theta_2}\{h(X^T, \theta_2)\}$$

and can be checked explicitly

Example: if h doesn't depend on θ and is one-to-one, the transition rates are identifiable (modulo states permutations).

Finite state case: identifiability

The question of identifiability:

$$\theta_1 \neq \theta_2 \quad \overset{?}{\Longrightarrow} \quad \mathrm{Law}_{\theta_1}\{Y^T\} \neq \mathrm{Law}_{\theta_2}\{Y^T\}$$

is equivalent to

$$\theta_1 \neq \theta_2 \quad \overset{?}{\Longrightarrow} \quad \mathrm{Law}_{\theta_1}\{h(X^T, \theta_1)\} \neq \mathrm{Law}_{\theta_2}\{h(X^T, \theta_2)\}$$

and can be checked explicitly

Example: if h doesn't depend on θ and is one-to-one, the transition rates are identifiable (modulo states permutations).

Finite state case: consistency

Proposition

Assume Θ is compact and

- *the chain is irreducible for all $\theta \in \Theta$*
- *the parametrization is identifiable*

Then the MLE is consistent:

$$\lim_{T\to\infty} \hat{\theta}_T = \theta_0, \quad \mathrm{P}_{\theta_0} - a.s. \quad \square$$

Notice: the conditions are also necessary !

Finite state case: consistency

Proposition

Assume Θ is compact and

- *the chain is irreducible for all $\theta \in \Theta$*
- *the parametrization is identifiable*

Then the MLE is consistent:

$$\lim_{T \to \infty} \hat{\theta}_T = \theta_0, \quad \mathsf{P}_{\theta_0} - a.s. \quad \square$$

Notice: the conditions are also necessary !

Finite state case: consistency proof

Implementation of the classical A.Wald's method, following B.Leroux's '92 ideas in discrete time:

- the likelihood is (almost) a subadditive and (almost) stationary process
- Kingman's LLN for subadditive stationary processes is applicable and provides all the necessary ingredients of Wald's program

Notice: appealing to the ergodic properties of the filtering process is (almost) avoided !

Finite state case: asymptotic normality

Theorem (P.Chigansky, '09)

Assume

- *h does not depend on* θ
- $\min_{i \neq j} \min_{\theta \in \bar{\Theta}} r_{ij}(\theta) > 0$
- *the model is identifiable in the sense [...]*
- *the Fisher information* $I(\theta_0) := \lim_{T \to \infty} \mathrm{E}_{\theta_0} \left(\partial_\theta \pi_T^{\theta_0}(h) \right)^2$ *is positive in the sense [...]*

Then the MLE $\hat{\theta}_T$ *is uniformly consistent, asymptotically normal (with converging moments).* □

Example: symmetric telegraph process (binary Markov chain) with the rate $\theta \in [a, b] \subset (0, \infty)$.

Finite state case: asymptotic normality

Theorem (P.Chigansky, '09)

Assume

- *h does not depend on* θ
- $\min_{i\neq j}\min_{\theta\in\bar\Theta} r_{ij}(\theta) > 0$
- *the model is identifiable in the sense [...]*
- *the Fisher information* $I(\theta_0) := \lim_{T\to\infty} \mathrm{E}_{\theta_0}\Big(\partial_\theta \pi_T^{\theta_0}(h)\Big)^2$ *is positive in the sense [...]*

Then the MLE $\hat\theta_T$ *is uniformly consistent, asymptotically normal (with converging moments).* □

Example: symmetric telegraph process (binary Markov chain) with the rate $\theta \in [a, b] \subset (0, \infty)$.

Finite state space: asymptotic normality

- Implementation of the Ibragimov-Khas'minskii program, heavily using the ergodic properties of the filtering process π_t^θ (with lots of technical nuances on the way!)
- There is a big gap in model assumptions between consistency and asymptotic normality (as in the analogous discrete time setting, Bickel et al, '98)
- No explicit formula for $I(\theta_0)$ even in the "telegraph signal" example (only numerical computation via PDE)

Finite state space: asymptotic normality

- Implementation of the Ibragimov-Khas'minskii program, heavily using the ergodic properties of the filtering process π_t^θ (with lots of technical nuances on the way!)
- There is a big gap in model assumptions between consistency and asymptotic normality (as in the analogous discrete time setting, Bickel et al, '98)
- No explicit formula for $I(\theta_0)$ even in the "telegraph signal" example (only numerical computation via PDE)

Finite state space: asymptotic normality

- Implementation of the Ibragimov-Khas'minskii program, heavily using the ergodic properties of the filtering process π_t^θ (with lots of technical nuances on the way!)
- There is a big gap in model assumptions between consistency and asymptotic normality (as in the analogous discrete time setting, Bickel et al, '98)
- No explicit formula for $I(\theta_0)$ even in the "telegraph signal" example (only numerical computation via PDE)