The Variational EM Algorithm for On-line Identification of Extended AR Models

Václav Šmídl

UTIA, Czech Academy of Science

Anthony Quinn

Trinity College Dublin, Ireland

March 22, ICASSP 05
Univariate AR model

AR model:

\[ d_t = -a_1 d_{t-1} - a_2 d_{t-2} - \ldots - a_q d_{t-q} + \sigma e_t, \quad e_t \sim \mathcal{N}(0,1) \]

with

AR parameters \( a = [a_1, \ldots, a_q]' \), and \( \sigma \),

regressor \( x_t = [d_{t-1}, \ldots, d_{t-q}]' \),

history \( D_t = [d_1, d_2, \ldots, d_t] \),

Observation model:

\[ f(d_t|a, \sigma, D_{t-1}) = f(d_t|a, \sigma, x_t) = \mathcal{N}(-a'x_t, \sigma^2). \]
Recursive Bayesian Inference

Conjugacy:

\[ f(a, \sigma | D_t) \propto f(d_t | a, \sigma, x_t) f(a, \sigma | D_{t-1}), \]

same functional form
Conjugacy:

\[ f(a, \sigma|D_t) \propto f(d_t|a, \sigma, x_t) f(a, \sigma|D_{t-1}), \]

same functional form

\[ f(a, \sigma|D_t) = f(a, \sigma|s(D_t)) \]

\[ s_t = s(D_t) \text{ is time-invariant function generating sufficient statistics.} \]
Recursive Bayesian Inference

Conjugacy:

\[ f(a, \sigma | D_t) \propto f(d_t | a, \sigma, x_t) f(a, \sigma | D_{t-1}), \]

same functional form

\[ f(a, \sigma | D_t) = f(a, \sigma | s(D_t)) \]

\[ s_t = s(D_t) \text{ is time-invariant function generating sufficient statistics.} \]
Recursive Bayesian Inference of AR parameters

Observation model for AR:

\[ f \left( d_t | a, \sigma, x_t \right) = \mathcal{N} \left( -a' x_t, \sigma^2 \right), \]

Conjugate distribution is Normal-inverse-Gamma, \( \mathcal{N}i\mathcal{G} \):

\[ f \left( a, \sigma | D_t \right) = \mathcal{N}i\mathcal{G} \left( V_t, \nu_t \right) \]

sufficient statistics, \( s_t = \{ V_t, \nu_t \} \):

\[
\begin{align*}
V_t &= V_{t-1} + \begin{bmatrix} d_t \\ x_t \end{bmatrix} \begin{bmatrix} d_t \\ x_t \end{bmatrix} = \begin{bmatrix} V_{d,d;t} & v'_{x,d;t} \\ v_{x,d;t} & v_{x,x;t} \end{bmatrix} \\
\nu_t &= \nu_{t-1} + 1,
\end{align*}
\]

Moments:

\[ \mathbb{E}_{f(a|D_t)} \left[ a \right] = V_{x,x;t}^{-1} v_{x,d;t}. \]
Extending AR model

**AR**

\[ e_t \xrightarrow{\sigma} x_{1:t} \xrightarrow{a_1} z^{-1} \xrightarrow{x_{2:t}} z^{-2} \xrightarrow{a_2} x_{q:t} \xrightarrow{a_q} z^{-q} \xrightarrow{a_q} d_t \]

**Extended AR (EAR)**

\[ e_t \xrightarrow{\sigma} y_t \xrightarrow{g_0^{-1}} d_t \]

\[ e_t \xrightarrow{\sigma} x_{1:t} \xrightarrow{a_1} x_{2:t} \xrightarrow{a_2} x_{q:t} \xrightarrow{a_q} z^{-1} \xrightarrow{z^{-1}} d_t \]

Posterior:

\[ f(a, \sigma | D_t) = \mathcal{N}iG(V_t, \nu_t) , \]

\[ z_t = [d_t, d_{t-1}, d_{t-2}, \ldots, d_{t-q}]' \]

\[ V_t = V_{t-1} + z_t z_t' \]

\[ \nu_t = \nu_{t-1} + 1 \]

\[ z_t = g'(D_t) = [g_0(D_t), g_1(D_{t-1}), \ldots, g_q(D_{t-1})]' \]
Mixture-based Extension of AR (MEAR)

Relaxing the assumption of known time-variant transformation. Treating $g$ as time-variant discrete random variable:

$$g \rightarrow g_t \in \{g_1, g_2, \ldots, g_c\} \leftrightarrow l_t \in \{l_1, l_2, \ldots, l_c\}$$

The observation model is

$$f(d_t|a, \sigma, l_t, D_{t-1}) = \prod_{i=1}^{c} f(d_t|a, \sigma, x_{i;t}, g_i)^{l_{i;t}},$$

$$f(l_{i;t}) = \alpha_i, \quad \sum_{i=1}^{c} \alpha_i = 1, \quad i = 1, \ldots, c,$$

$$f(d_t|a, \sigma, D_{t-1}) = \sum_{i=1}^{c} \alpha_i f(d_t|a, \sigma, x_{i;t}, g_i),$$

EAR

MEAR
Conjugacy Lost and Conjugacy Regained

Bayes’ rule:

\[ f(a, \sigma, l_t|D_t) \propto f(d_t|a, \sigma, l_t, D_{t-1}) f(l_t) f(a, \sigma|D_{t-1}), \]

Conditional independence approximation:

\[ f(a, \sigma, l_t|D_t) \approx \tilde{f}(a, \sigma, l_t|D_t) = \tilde{f}(a, \sigma|D_t) \tilde{f}(l_t|D_t), \]

Then:

\[ \tilde{f}(a, \sigma|D_t) \tilde{f}(l_t|D_t) \propto f(d_t|a, \sigma, l_t, D_{t-1}) f(l_t) f(a, \sigma|D_{t-1}). \]
Traditionally, conditional independence was achieved heuristically:

\[
\begin{align*}
  f(a, \sigma, l_t|D_t) & \approx \tilde{f}(a, \sigma|D_t) \tilde{f}(l_t|D_t), \\
  \tilde{f}(l_t|D_t) & = \int_{a,\sigma} f(a, \sigma, l_t|D_t) \, da \, d\sigma, \\
  \tilde{f}(a, \sigma|D_t) & = f(a, \sigma|D_t, \hat{l}_t).
\end{align*}
\]

**Quasi-Bayes:** 
\[\hat{l}_t = E_{\tilde{f}(l_t|D_t)}[l_t],\]

**Viterbi-like:** 
\[\hat{l}_t = \arg \max_{l_t} \tilde{f}(l_t|D_t).\]

We favour free-form functional optimization.
The Variational Bayes (VB) method

Two ingredients:

1. Conditional independence assumption:

\[ \tilde{f}(a, \sigma, l_t|D_t) = \tilde{f}_1(a, \sigma|D_t) \tilde{f}_2(l_t|D_t), \]

2. Minimization of the KL divergence,

\[ \tilde{f}_1(\cdot), \tilde{f}_2(\cdot) = \arg\min_{\tilde{f}_1, \tilde{f}_2} KL \left( \tilde{f}(a, \sigma|D_t) \tilde{f}(l_t|D_t) \| f(a, \sigma, l_t|D_t) \right), \]

Necessary condition for minimum:

\[ \tilde{f}_1(a, \sigma|D_t) \propto \exp \left( E_{\tilde{f}_2(l_t|D_t)} [\ln f(a, \sigma, l_t, D_t)] \right), \]

\[ \tilde{f}_2(l_t|D_t) \propto \exp \left( E_{\tilde{f}_1(a, \sigma|D_t)} [\ln f(a, \sigma, l_t, D_t)] \right). \]

Solved iteratively via the Variational EM (VEM) algorithm.
The Variational Bayes (VB) method

Two ingredients:

1. Conditional independence assumption:

\[ \tilde{f}(\mathbf{a}, \sigma, \mathbf{l}_t|D_t) = \tilde{f}_1(\mathbf{a}, \sigma|D_t) \tilde{f}_2(\mathbf{l}_t|D_t), \]

2. Minimization of the KL divergence,

\[ \tilde{f}_1(\cdot), \tilde{f}_2(\cdot) = \arg \min_{\tilde{f}_1, \tilde{f}_2} KL \left( \tilde{f}(\mathbf{a}, \sigma|D_t) \tilde{f}(\mathbf{l}_t|D_t) \| f(\mathbf{a}, \sigma, \mathbf{l}_t|D_t) \right), \]

Necessary condition for minimum:

\[ \tilde{f}_1(\mathbf{a}, \sigma|D_t) \propto \exp \left( E_{\tilde{f}_2(\mathbf{l}_t|D_t)} [ \ln (f(\mathbf{a}, \sigma, \mathbf{l}_t, D_t)) ] \right), \]

\[ \tilde{f}_2(\mathbf{l}_t|D_t) \propto \exp \left( E_{\tilde{f}_1(\mathbf{a}, \sigma|D_t)} [ \ln (f(\mathbf{a}, \sigma, \mathbf{l}_t, D_t)) ] \right). \]

Solved iteratively via the Variational EM (VEM) algorithm.
The Variational Bayes (VB) method

Two ingredients:

1. Conditional independence assumption:
   \[ \tilde{f}(a, \sigma, l_t | D_t) = \tilde{f}_1(a, \sigma | D_t) \tilde{f}_2(l_t | D_t), \]

2. Minimization of the KL divergence,
   \[ \tilde{f}_1(\cdot), \tilde{f}_2(\cdot) = \arg \min \limits_{\tilde{f}_1, \tilde{f}_2} KL \left( \tilde{f}(a, \sigma | D_t) \tilde{f}(l_t | D_t) \| f(a, \sigma, l_t | D_t) \right), \]

Necessary condition for minimum:

\[ \tilde{f}_1(a, \sigma | D_t) \propto \exp \left( E_{\tilde{f}_2(l_t | D_t)} \left[ \ln \left( f(a, \sigma, l_t, D_t) \right) \right] \right), \]
\[ \tilde{f}_2(l_t | D_t) \propto \exp \left( E_{\tilde{f}_1(a, \sigma | D_t)} \left[ \ln \left( f(a, \sigma, l_t, D_t) \right) \right] \right). \]

Solved iteratively via the Variational EM (VEM) algorithm.
\[ \tilde{f}_1 (a, \sigma | D_t) \propto \exp \left( \mathbb{E}_{\tilde{f}(l_t|D_t)} \left[ \ln \left( f(d_t|a, \sigma, l_t) f(l_t) \tilde{f}(a, \sigma | D_{t-1}) \right) \right] \right), \]

\[ \tilde{f}_1 (a, \sigma | D_t) \propto \text{red} \tilde{f}(d_t|a, \sigma) \tilde{f}(a, \sigma | D_{t-1}) \]

VB the generates a partial VB-observation model for each variable.
\[ \tilde{f}_1(a, \sigma | D_t) \propto \exp \left( \mathbb{E}_{\tilde{f}(l_t | D_t)} \left[ \ln \left( f(d_t | a, \sigma, l_t) f(l_t) \tilde{f}(a, \sigma | D_{t-1}) \right) \right] \right), \]

\[ \tilde{f}_1(a, \sigma | D_t) \propto \text{red} \{ f(d_t | a, \sigma) \tilde{f}(a, \sigma | D_{t-1}) \} \]

VB the generates a partial VB-observation model for each variable.
$$
\tilde{f}_1(a, \sigma | D_t) \propto \exp \left( E_{\tilde{f}(l_t | D_t)} \left[ \ln \left( f(d_t | a, \sigma, l_t) f(l_t) \tilde{f}(a, \sigma | D_{t-1}) \right) \right] \right), \\
\tilde{f}_1(a, \sigma | D_t) \propto \text{red}\{\tilde{f}(d_t | a, \sigma) \tilde{f}(a, \sigma | D_{t-1})\}
$$

VB the generates a partial VB-observation model for each variable.

$$f(a, \sigma | D_{t-1}) \xrightarrow{BT} \tilde{f}(a, \sigma | D_t)$$

$$\tilde{f}_1(d_t | a, \sigma, D_{t-1})$$

$$\tilde{f}_2(d_t | l_t, D_{t-1})$$

$$f(l_t | D_{t-1}) \xrightarrow{BT} \tilde{f}(l_t | D_t)$$
\[ \tilde{f}_1(a, \sigma | D_t) \propto \exp \left( \mathbb{E}_{\tilde{f}(I_t | D_t)} \left[ \ln \left( f(d_t | a, \sigma, I_t) f(I_t) \tilde{f}(a, \sigma | D_{t-1}) \right) \right] \right), \]
\[ \tilde{f}_1(a, \sigma | D_t) \propto \text{red}_1 \tilde{f}(d_t | a, \sigma) \tilde{f}(a, \sigma | D_{t-1}) \]

VB the generates a partial VB-observation model for each variable.

Observation models interact with posteriors via moments.
VB-posterior distributions:

\[ \tilde{f}(a, \sigma | D_t) = \mathcal{NI}(V_t, \nu_t), \]
\[ \tilde{f}(l_t | D_t) = \mathcal{Mu}(w_t), \]

with statistics expressed implicitly:

\[ V_t = V_{t-1} + \sum_{i=1}^{c} w_{i;t} z_{i;t} z_{i;t}', \]
\[ \nu_t = \nu_{t-1} + 1, \]
\[ w_{i;t} \propto |J_{i;t}| \exp \left[ -\frac{1}{2\hat{\sigma}} z_{j;t} \begin{bmatrix} -1, \hat{a}' \end{bmatrix}' \begin{bmatrix} -1, \hat{a}' \end{bmatrix} z_{j;t} \right. \]
\[ \left. -\frac{1}{2} z_{j;t}' V_{x,x;t}^{-1} z_{j;t} \right], \]

Solved iteratively.

Note: Statistics \( V_t \) is updated by rank-c structure.
VB-posterior distributions:

\[
\tilde{f}(a, \sigma | D_t) = N_iG(V_t, \nu_t),
\]
\[
\tilde{f}(l_t | D_t) = Mu(w_t),
\]

with statistics expressed implicitly:

\[
V_t = V_{t-1} + \sum_{i=1}^{c} W_{i;t} z_{i;t} z'_{i;t} \]
\[
\nu_t = \nu_{t-1} + 1,
\]
\[
w_{i;t} \propto |J_{i;t}| \exp \left[ - \frac{1}{2\hat{\sigma}} z'_{j;t} \left[ -1, \hat{a}' \right] \left[ -1, \hat{a}' \right] z_{j;t} - \frac{1}{2} z'_{j;t} V_{x;x;t}^{-1} z_{j;t} \right], \tag{1}
\]

Solved iteratively.

Note: Statistics \( V_t \) is updated by rank-\( c \) structure.
VB for MEAR

VB-posterior distributions:

\[
\tilde{f}(a, \sigma | D_t) = \mathcal{N}(V_t, \nu_t),
\]

\[
\tilde{f}(l_t | D_t) = \mathcal{M}u(w_t),
\]

with statistics expressed implicitly:

\[
V_t = V_{t-1} + \sum_{i=1}^{c} w_{i;t} z_{i;t} z'_{i;t}
\]

\[
\nu_t = \nu_{t-1} + 1,
\]

\[
w_{i;t} \propto |J_{i;t}| \exp \left[ -\frac{1}{2\hat{\sigma}} z'_{j;t} \left[ -1, \hat{a}' \right]' \left[ -1, \hat{a}' \right] z_{j;t} - \frac{1}{2} z'_{j;t} V_{x, x; t}^{-1} z_{j;t} \right],
\]

Solved iteratively.

Note: Statistics $V_t$ is updated by rank-c structure.
Consider a Markov model of label evolution:

\[
f(l_t \mid l_{t-1}, Q) = \prod_{i,j=1}^{C} q_{i,j}^{l_i l_j t^{-1}},
\]

with unknown transition matrix \( Q \).

Conditional independence is imposed between the weights:

\[
f(a, \sigma, l_t, l_{t-1} \mid D_t) \approx \tilde{f}(\sigma \mid D_t) \tilde{f}(l_t \mid D_t) \tilde{f}(l_{t-1} \mid D_t),
\]

Posterior distributions are again Multinomial on both labels, \( l_t, l_{t-1} \), and Dirichlet on \( Q \).
Since conjugacy was restored using VB, the technique of forgetting can be used for dealing with time-variant parameters.

\[
\tilde{f}(a_t, \sigma_t | D_{t-1}) \propto \left[ \tilde{f}(a_{t-1}, \sigma_{t-1} | D_{t-1}) \right]^{\phi} \quad a_{t-1} \rightarrow a_t \quad \sigma_{t-1} \rightarrow \sigma_t
\]

In combination with time-variant weights we use two approximations of time-update step of stochastic filtering: (i) VB for weights, (ii) forgetting for \(a\) and \(\sigma\).
Outlier corrupted AR

Outlier:

\[ y_t = -ax_t + \sigma e_t, \]
\[ d_t = y_t + \xi_t, \quad \xi_t = \begin{cases} he_t \\ 0 \end{cases} \]

Transformations:

\[ g_1 : z_t = [d_t, \ldots, d_{t-q}]', \]
\[ g_2 : z_t = \frac{1}{h} [d_t, \ldots, d_{t-q}]', \]
\[ g_i : i = 3 \ldots q \]
\[ z_t = [d_t, \ldots, \hat{y}_{t-i}, \ldots d_{t-q}]', \]
Burst noise corrupted AR

Burst noise:

\[ y_t = -ax_t + \sigma e_t, \]
\[ d_t = y_t + \xi_t, \quad \xi_t \sim \mathcal{N}(0, h), \]

with unknown \( h \).

Transformations:

\[ g_1 : z_t = [d_t, \ldots, d_{t-q}]', \]
\[ g_i : z_t = \frac{1}{\kappa_i} [d_t, \hat{x}_{i;t}], \]

where \( \kappa_i \) and \( \hat{x}_{i;t} \) are from Kalman filter.

Experiment was performed with forgetting factor \( \phi = 0.95 \).
Conclusion

- Mixture-based extension of the AR for dealing with unknown transformations was presented.
- The Variational Bayes technique was used to achieve conjugate update of parameter statistics.
- Applied to reconstruction of speech corrupted by burst noise.
The chosen transformation (filter-bank) forms nodes of a simplex of possible dyads. Update by a linear combination of these dyads allows us to span interior of the simplex.

desired behaviour

\[ g_1 \quad g_2 \quad g_3 \]

undesired behaviour

\[ g_1 \quad g_2 \quad g_3 \]

This should be remembered when designing a filter-bank.
Computational flow

Reminiscent of multiple model approach. In our approach, however, we propagate statistics.