Penalized Likelihood Estimation in High-Dimensional Time Series Models

1 Introduction

Aim: Construct a general estimation method for high-dim. time series models by penalized QML that gives sparse estimates.

Examples: \( K \)-dim. VAR(\( r \)) model is defined by

\[
y_{t} = \Phi_{1}y_{t-1} + \cdots + \Phi_{r}y_{t-r} + \varepsilon_{t},
\]

which has \( K^2r \) parameters. \( K \)-dim. MGARCH(1,1) is given by

\[
y_{t} = \Sigma_{t}^{1/2}\varepsilon_{t}, \quad \Sigma_{t} = C\Sigma_{t-1}^\top + A^\top y_{t-1} + A + B^\top \Sigma_{t-1} B,
\]

which has \( K(5K + 1)/2 \) parameters.

2 General Theory

2.1 The model and its PQML estimator

Model: Let \( \{y_t\}_{t=1} = (y_t, y_{t-1}, \ldots) \) be a vector stationary process with a continuous conditional density \( g(y_t|y_{t-1}, y_{t-2}, \ldots) \). Consider a parametric family of densities \( \{ f(y_t|y_{t-1}, y_{t-2}, \ldots; \theta) : \theta \in \Theta \} \) s.t.: \( p := \dim(\Theta) = \text{tr}(\Sigma) \) for some \( \delta > 0 \), so possibly \( p > n \); the “true value” \( \theta^{0} \), the unique minimizer of the KLIC of \( g \) relative to \( f \), is sparse.

Define some notation more precisely:

- \( \mathcal{M}_0 = \{ j \in \{1, \ldots, p \} : \theta_j^0 \neq 0 \} \) and \( \bar{\mathcal{M}}_0 = \{1, \ldots, p\} \setminus \mathcal{M}_0 \);
- \( \theta_{(q)}^0 \) is the \( q \)-dim. subvector of \( \theta^0 \) composed of the nonzero elements \( \{ \theta_j^0 \} : j \in \mathcal{M}_0 \);
- \( \hat{\theta}_{(p-q)}^0 \) is the \( (p-q) \)-dim. subvector of \( \theta^0 \) composed of zeros.

Estimator: The PQML estimator \( \hat{\theta} \) of \( \theta^0 \) is defined by

\[
Q_n(\hat{\theta}) = \max Q_n(\theta) \quad \text{with} \quad Q_n(\theta) := L_n(\theta) - P_n(\theta),
\]

where \( L_n(\theta) := n^{-1}\sum_{t=1}^{n} \log f(y_t|y_{t-1}, \ldots; \theta) \) is the quasi-log-likelihood and \( P_n(\theta) := \sum_{t=1}^{n} p_{\lambda}(\theta_{(q)}) \) is the penalty term such as L1-penalty (lasso), SCAD, MCP, etc., with \( \lambda(=\lambda_n) \to 0 \).

2.2 Theoretical results

Theorem 1 (Weak oracle property) Under regularity conditions, there is a local maximizer \( \hat{\theta} = (\hat{\theta}_{(q)}^0, \hat{\theta}_{(p-q)}^0) \) of \( Q_n(\theta) \) s.t.:

(1) \( P(\hat{\theta}_{(q)}^0 = 0) \to 1 \); (2) \( \|\hat{\theta}_{(q)} - \theta_{(q)}^0\|_{\lambda} = O_p(n^{-1/2}) \).

Corollary 1 (L1-penalized QML estimator) Under regularity conditions in Theorem 1, there is a local maximizer \( \hat{\theta} = (\hat{\theta}_{(q)}^0, \hat{\theta}_{(p-q)}^0) \) of \( Q_n(\theta) \) s.t. Thm. 1 (a) and (b) hold.

Theorem 2 (Oracle property) Under regularity conditions, there is a local maximizer \( \hat{\theta} = (\hat{\theta}_{(q)}^0, \hat{\theta}_{(p-q)}^0) \) of \( Q_n(\theta) \) s.t.:

(1) \( P(\hat{\theta}_{(q)}^0 = 0) \to 1 \); (2) \( \|\hat{\theta}_{(q)} - \theta_{(q)}^0\|_{\lambda} = O_p(n^{-1/2}) \). If a stronger assumption is added to the penalty, we have (Asy. N) \( n^{1/2}(\hat{\theta}_{(q)}^0 - \theta_{(q)}^0) \to d N(0, (\theta_{(q)}^0)^{-1} P_{(q)}(\theta_{(q)}^0)^{-1}) \).

3 Application to VAR

3.1 Theoretical result for VAR

Consider (1) with \( \varepsilon_t \sim i.i.d. (0, \Sigma_\varepsilon) \). Let \( \theta^0 = \text{vec}(\Phi_1^0, \ldots, \Phi_r^0) \in \mathbb{R}^p \) with \( p = K^2r \), which is supposed sparse. Using some appropriate \( \Sigma \) instead of unknown \( \Sigma_n \), we have:

Proposition 1 Under some moment and stability conditions, Thm. 2 (a) (c) hold for \( \hat{\theta} \) in (1), where \( \hat{\theta}_{(q)}^0 = P_{(q)}(\Gamma \otimes \Sigma_\varepsilon^{-1}\Sigma_n^{-1})P_{(q)} \) and \( P_{(q)} = P_{(q)}(\Sigma_\varepsilon^{-1}\Sigma_n^{-1})P_{(q)} \) with \( \Gamma = E[x_i x_i^\top] \).

3.2 Empirical study

Compare performances of sparse VAR and dynamic Nelson-Siegel (DNS) model in terms of yield curve forecasting.

Data: Zero-coupon US government bond yields that are:

- monthly from January 1986 to December 2007;
- made of 8 maturities \( \tau = 3, 6, 12, 24, 36, 60, 84, 120 \) months.

Model 1: DNS model is defined by

\[
y_{\tau i} = \beta_{1 i} + \beta_{2 i} \left(1 - e^{-\eta_i \tau} \right) + \beta_{3 i} \left(1 - e^{-\eta_i \tau} / e^{-\eta_i \tau} \right),
\]

\[
\beta_{1 i} = a_i + b_i \beta_{1, i-1} + u_i \quad \text{for each} \quad i = 1, 2, 3,
\]

where \( \beta_{1 i}, \beta_{2 i}, \beta_{3 i} \) can be interpreted as latent dynamic factors and \( \eta_i \) is a sequence of tuning parameters.

Model 2: In sVAR strategy, the model is specified as 8-dim. VAR(12) below and is estimated by SCAD penalized QML.

\[
\begin{pmatrix}
\Delta y_{3, t} \\
\Delta y_{6, t}
\end{pmatrix}
= \Phi_1
\begin{pmatrix}
\Delta y_{3, t-1} \\
\Delta y_{6, t-1}
\end{pmatrix}

+ \cdots + \Phi_{12}
\begin{pmatrix}
\Delta y_{3, t-12} \\
\Delta y_{6, t-12}
\end{pmatrix}
+ \varepsilon_t.
\]

Forecasting strategy: The two models are estimated recursively, using the data from Jan. 1986 to the time \( h \in \{1, 3, 6, 12\} \)-month-ahead forecast is made, beginning in Jan. 2001 and extending through Dec. 2007.

Result: The comparison result is summarized below:

<table>
<thead>
<tr>
<th>( h ) (month)</th>
<th>3</th>
<th>6</th>
<th>12</th>
<th>24</th>
<th>36</th>
<th>60</th>
<th>84</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.356</td>
<td>0.301</td>
<td>0.288</td>
<td>0.279</td>
<td>0.266</td>
<td>0.254</td>
<td>0.258</td>
<td>0.275</td>
</tr>
<tr>
<td>3</td>
<td>0.418</td>
<td>0.393</td>
<td>0.358</td>
<td>0.345</td>
<td>0.333</td>
<td>0.324</td>
<td>0.329</td>
<td>0.356</td>
</tr>
<tr>
<td>6</td>
<td>0.557</td>
<td>0.513</td>
<td>0.443</td>
<td>0.405</td>
<td>0.391</td>
<td>0.379</td>
<td>0.381</td>
<td>0.400</td>
</tr>
<tr>
<td>12</td>
<td>0.625</td>
<td>0.591</td>
<td>0.540</td>
<td>0.492</td>
<td>0.468</td>
<td>0.442</td>
<td>0.435</td>
<td>0.445</td>
</tr>
</tbody>
</table>