A Non-Parametric Bayesian AR Model – Application to DNA-sequencing

Maria Anna Di Lucca\textsuperscript{1} Peter Müller\textsuperscript{2} Yuan Ji\textsuperscript{3}

\textsuperscript{1}Department of Decision Science Bocconi University, Italy
\textsuperscript{2}Department of Biostatistics MD Anderson Cancer Center, Texas
\textsuperscript{3}Department of Bioinformatics MD Anderson Cancer Center, Texas

November, 8th 2010
High-throughput sequencing

- High-throughput sequencing reads $10^9$ base pairs per run.
- Many problems calls for solutions, we concentrate on Base Calling.
- Base-Calling: map (4 dim) intensity measurements, $y_t$ into base $k_t \in \{A, C, G, T\}$.
- Ideally the intensity corresponding to the correct base is highest.
- DNA is arranged in short segments of length 36 bases pairs (= colony)
**Observed DNA-Sequencing dataset**

- dataset references: short DNA fragments of 36 bases each, \( i = 1, 2, \ldots, 36 (=\text{colony}) \).
- Each colony corresponds to a DNA segment with 36 bases.
- For each base there are four fluorescence intensities 
  \( y_t = (y_{t1}, y_{t2}, y_{t3}, y_{t4}) \) (ideally the intensity corresponding to \( k_t \) is the highest).

There are three kind of noises: *phasing*, *fading* and *cross-talk* between the channels.
Compare DNA-Sequencing Dataset with true bases

- Data: 4 dim measurements $y_t \in \mathbb{R}^4$
- Truth: time basis $k_t \in \{A, C, G, T\}$

Figure: It is the subset $(350 \times 4)$ of the complete dataset with $(35000 \times 4)$. The different colors come from the true sequences. There is evidence of cross-talk between the channels.
Previous Parametric Studies

Ji and all (2010) defined a multivariate mixture of 4 Normals to model the DNA-sequencing:

\[ y_t \sim \sum_{j=1}^{4} P(k_t = j) N_4(\mu_j, \alpha y_{t-1,j}, \Sigma_j) \]

for \( t = 2, \ldots, 36 \)

- We will replace \( N(.,.) \) assumption by a non-parametric model.
We propose a non-parametric Bayesian autoregression for a sequence \( \{y_t\} \).

\[
y_t \mid y_{t-1} = y \sim \sum_{j=1}^{4} p_j F_j^y
\]

We introduce a family of unknown random probability measures

\[
\mathcal{F}_j \equiv \{F^j_{y,i}(\cdot); y \in \mathbb{R}^4, i \in 1, 2, \ldots, 36, j \in 1, 2, 3, 4\}
\]

with a DDP prior \( P(\mathcal{F}_j) \)

\[
F^j_{y,i} = \sum_{h=1}^{\infty} \omega^j_h N_4(\theta^j_h(y, i), \Sigma_j)
\]

DDP with common weights across \( y \)
Summary

- $y_t \mid y_{t-1} = y \sim \sum_{j=1}^{4} p_j F_{y,i}^j$

or $y_t \mid y_{t-1} = y, k_t = j \sim F_{y,i}^j$

where $P(k_t = j) = p_j$

- $F_{y,i}^j = \sum_{h=1}^{H} \omega_{h}^j N_4(\theta_{h}^j(y, i), \Sigma_j)$

or

$y_t \mid y_{t-1} = y, r_t = h \sim \sum_{j=1}^{J} p_j N_4(\alpha_{h}^j \times y_{t-1} + e^{-\lambda i \gamma} \beta_{h}^j, \Sigma_j)$
DDP-VAR(1) model for DNA-Sequencing

\[ y_t \mid y_{t-1} = y, r_t = h, k_t = j \sim N(\theta^j_h(y, i), \Sigma_j) \]

\[ \theta^j_h(y, i) = \alpha^j_h \times y + \beta^j_h e^{-\lambda t^\gamma} \]

\[ P(r_t = h) = w_h \]
\[ P(k_t = j) = p_j \]
\[ (\beta^j_h, \alpha^j_h) \sim G^0_j(\beta^j_h, \alpha^j_h) \]

- \( \beta^j_h \sim N_4(m_\beta, \Sigma_\beta) \)
- \( \alpha^j_h \sim N_4(m_\alpha, \Sigma_\alpha) \)
- \( \beta^j_h \) and \( \alpha^j_h \) are independent
First Results on the Dataset

Figure: The elipses = a posterior draw for $\alpha_h^j, \beta_h^j, \Sigma_j$; the yellow are for $A$ ($j = 1$), The light blue are for $C$ ($j = 2$), the grey for $G$ ($j = 3$) and the pink for $T$ ($j = 4$).
A Simulated Dataset

Figure: These plots show simulated data. The form and the position show the absence of exchangeability.
Correct vs Wrong Labels $k_t$

Figure: The blue barplot counts the number of true base calls and the red barplot the number of the wrong labels.
Remarks

- This is a natural consequence of the ANOVA DDP for linear models.

- The innovation is the extension to the time series problems.

- The flexibility of the model is useful for real problems.