Time Series Prediction via Density Ratio Modeling

Benjamin Kedem
Department of Mathematics & ISR University of Maryland, College Park, MD

Joint with: Richard E. Gagnon and Haiming Guo.
A Density Ratio

Our starting point is a well known fact regarding $k$-parameter exponential families:

$$\theta = (\theta_1, ..., \theta_k)$$

$$g(x, \theta) = d(\theta)S(x) \exp \left\{ \sum_{i=1}^{k} c_i(\theta)T_i(x) \right\}$$

Define:

$$\alpha = \log[d(\theta_1)/d(\theta_2)]$$

$$\beta = (c_1(\theta_1) - c_1(\theta_2), ..., c_k(\theta_1) - c_k(\theta_2))'$$

$$h = (T_1(x), ..., T_k(x))'$$

Then

$$\frac{g_1(x)}{g_2(x)} \equiv \frac{g(x, \theta_1)}{g(x, \theta_2)} = \exp\{\alpha + \beta' h(x)\} \quad (1)$$

or

$$g_1(x) = \exp\{\alpha + \beta' h(x)\} g_2(x) \quad (2)$$
Normal distribution: 
\( \theta = (\mu, \sigma^2) \),

\[
\alpha = \log \left( \frac{\sigma_2}{\sigma_1} \right) + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2} \\
\beta = \left( \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}, \frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2} \right)'
\]

\( h(x) = (x, x^2)' \)

Gamma distribution: 
with shape \( r \) and scale \( \lambda \), \( \theta = (r, \lambda) \),

\[
\alpha = \log \frac{\lambda_1^{r_1} \Gamma(r_2)}{\lambda_2^{r_2} \Gamma(r_1)} \\
\beta = (\lambda_2 - \lambda_1, r_1 - r_2)' \\
\color{black}{ h(x) = (x, \log x)'}
\]

Rayleigh distribution: 
\( \alpha = \log \frac{\theta_2^2}{\theta_1^2}, \quad \beta = \frac{1}{2\theta_2^2} - \frac{1}{2\theta_1^2}, \quad h(x) = x^2 \)
A Density Ratio Model for Time Series

Assume we have $m$ regressions:

$$x_{1t} = f_1(z_{1,t-1}) + \epsilon_{1t}, \quad t = 1, ..., n_1$$

$$\vdots$$

$$x_{qt} = f_q(z_{q,t-1}) + \epsilon_{qt}, \quad t = 1, ..., n_q$$

$$x_{mt} = f_m(z_{m,t-1}) + \epsilon_{mt}, \quad t = 1, ..., n_m$$

(3)

The $x_{jt}$ may be nonstationary, and the $f_i$ may be non-linear.

In applications the $f_i$ are estimated and the $\epsilon_{kt}$ are the residuals. They may be dependent.
Suppose that for each $t$,
\[ \epsilon_{jt} \sim g_j(x), \quad j = 1, \ldots, q, m \]
Define the *reference* density:
\[ g(x) = g_m(x) \]
Emboldened by (2), we assume:
\[ g_j(x) = \exp\{\alpha_j + \beta_j' h(x)\} g(x) \quad j = 1, \ldots, q \tag{4} \]
Define the *combined data*
\[ \tau \equiv \{(\epsilon_{1t}, \ldots, \epsilon_{1n_t}), \ldots, (\epsilon_{q1}, \ldots, \epsilon_{qn_q}), (\epsilon_{m1}, \ldots, \epsilon_{mn_m})\} \]
Problem:
From the combined data estimate all the $\alpha_j, \beta_j$, and $g(x), G(x)$, for the purpose of predicting the future reference value $x_{m,t+1}$. 
Remark
$g(x)$ is estimated from all the combined residual set of size $n = n_1 + n_2 + \cdots + n_m$ and not just from “its own” residuals $(\epsilon_{m1}, \ldots, \epsilon_{mn_m})$ of size $n_m < n$.

Remark
The exponential “tilt” relationships (4) relative to a reference or baseline density $g$ enable a semiparametric inference about the $\alpha_j, \beta_j$, and about $g, G$ based on the combined data set.

Remark
Using the estimator $\hat{G}$ of $G$ we can estimate future probabilities of events formulated in terms of the “reference” $x_{m,t+1}$ conditional on $z_{m,t}$.
Example: Bivariate AR.

The linear system

\[ x_t = a_1x_{t-1} + a_2y_{t-1} + \epsilon_t \]
\[ y_t = b_1x_{t-1} + b_2y_{t-1} + \eta_t \]

\( t = 1, \ldots, N \), with independent Gaussian noise components \( \epsilon_t \sim N(0, \sigma_1^2) \) and \( \eta_t \sim N(0, \sigma_2^2) \), satisfies the density ratio model. We have

\[
\frac{g_\epsilon(x)}{g_\eta(x)} = \exp\left\{ \log \frac{\sigma_2}{\sigma_1} + \left( \frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2} \right) x^2 \right\}
\]

\[ \equiv e^{\alpha + \beta x^2} \]

or \( m = 2, q = 1 \), and (4) reduces to

\[ g_\epsilon(x) = e^{\alpha + \beta x^2} g_\eta(x) \]  

(5)
Semiparametric Estimation

A maximum likelihood estimator of $G(x)$ can be obtained by maximizing the empirical likelihood over the class of step cdf’s with jumps at the values $\tau_1, \ldots, \tau_n$: Qin and Lawless(1994), Qin and Zhang(1997), GLV(1999), FKQS(2001).

Let $w_j(\tau) = \exp \{ \alpha_j + \beta_j' h(\tau) \}$, $j = 1, \ldots, q$, and $p_i = dG(\tau_i)$, $i = 1, \ldots, n$.

Then the empirical likelihood becomes (Owen(2001)),

$$L(\alpha, \beta_1, \ldots, \beta_q, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} w_1(\epsilon_{1j}) \cdots \prod_{j=1}^{n_q} w_q(\epsilon_{qj})$$  \hspace{1cm} (6)
Define:
\[ \hat{w}_j(\tau) = \exp \{ \hat{\alpha}_j + \hat{\beta}'_j h(\tau) \} \]

We obtain FKQS(2001),
\[ \hat{p}_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 \exp(\hat{w}_1(\tau_i)) + \cdots + \rho_q \exp(\hat{w}_q(\tau_i))} \]

Therefore, with \( I(B) \) the indicator of \( B \),
\[ \hat{G}(\tau) = \sum_{i=1}^{n} \hat{p}_i I(\tau_i \leq \tau) = \]
\[ \frac{1}{n_m} \cdot \sum_{i=1}^{n} \frac{I(\tau_i \leq \tau)}{1 + \rho_1 \exp(\hat{w}_1((\tau_i)) + \cdots + \rho_q \exp(\hat{w}_q((\tau_i)))} \]

Asymptotic properties of \( \hat{G} \), and its optimality over the empirical distribution function obtained only from the reference sample
\[ \epsilon_{m1}, \cdots, \epsilon_{mn_m} \]

ignoring all the other samples, are discussed in a sequence of papers by Zhang (2000abc).
Prediction

Since

\[ x_{m,t+1} = f_m(z_{m,t}) + \epsilon_{m,t+1} \]

and \( \epsilon_{m,t+1} \sim G \), we have the useful approximation of the predictive probability at \( t + 1 \) conditional on \( z_{m,t} \),

\[
P(x_{m,t+1} \leq x \mid z_{m,t}) = G(x - f_m(z_{m,t})) \approx \hat{G}(x - f_m(z_{m,t}))
\]

(7)

From (7) we can get the predicted values

\[
\hat{E}(x_{m,t+1} \mid z_{m,t})
\]
Bivariate AR Example

The bivariate AR(1) system (5) was simulated with parameters

\[(a_1, a_2, b_1, b_2, \sigma_1, \sigma_2) = (0.6, -0.5, 0.4, 0.5, 1, 0.5)\]

and \(n_1 = n_2 = 500\).

The true “tilt” parameters are:

\[(\alpha_1, \beta_1) = (-0.693, 1.5)\]

and the reference \(g\) is the pdf of \(\eta_t\),

\[\eta_t \sim N(0, 0.5^2)\]
Plots of $x_t$ (solid) and $y_t$ (dashed), $t = 1, \ldots, 100$.

Pairs of $(G(x), \hat{G}(x))$, $-2 < x < 2$. 
Bivariate AR: a. Estimated reference cdf $\hat{G} = \hat{G}_\eta$. b. Estimated spline-smoothed reference pdf $\hat{g} = \hat{g}_\eta$ and a histogram from $\eta_t$. c. Estimated spline-smoothed tilted pdf $\hat{g}_\epsilon$ and a histogram from $\epsilon_t$. d. Estimated spline-smoothed reference pdf (dashed line) and the true $g = g_\eta$ (solid line).
Prediction: \( \hat{P}(y_t > a \mid x_{t-1}, y_{t-1}), t = 400, \ldots, 501. \)
Prediction: \( \hat{P}(y_t > 1 \mid x_{t-1}, y_{t-1}) \) versus \( I[y_t > 1] \), \( t = 400, \ldots, 501 \).
Prediction: $\hat{P}(y_t > 0.5 \mid x_{t-1}, y_{t-1})$ versus $I[y_t > 0.5]$, $t = 400, ..., 501$.

Estimated autocorrelation of the residual series resulting from the estimated bivariate AR model (5).
Top: Prediction (dashed line) of $y_t$ from the bivariate AR model (5) using the conditional expectation computed from (7). Bottom: The conditional expectation $E[y_t \mid x_{t-1}, y_{t-1}]$ and its estimate (dashed line) obtained from (7).
LA Mortality Prediction

We shall now apply the semiparametric method to sampled filtered mortality data from Los Angeles County, 01.01.1970 to 12.31.1979, discussed in Shumway et al (1988). The original daily data, consisting of the response series (total mortality) and its covariate series (two weather and six pollution series), were filtered and then sampled weekly to produce series of length $N = 508$ each.

$y =$ Total mortality
$T =$ Temperature (centered at 74.26 F$^\circ$)
$CO =$ Carbon monoxide
Kedem and Fokianos (2002): \textbf{Model 1}

\[ y_t = \exp \{ \hat{\beta}_0 + \hat{\beta}_1 y_{t-1} + \hat{\beta}_2 y_{t-2} + \hat{\beta}_3 T_{t-1} + \hat{\beta}_4 \log(CO_t) \} + \eta_t \]

\[ (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4) = (4.505, 0.00189, 0.00184, -0.00133, 0.04683) \]

The exponential model outperforms several competitors, giving close to white noise residual \( \eta_t \).

Additional regression: \textbf{Model 2}

\[ T_t = \hat{\phi}_1 T_{t-1} + \hat{\phi}_2 T_{t-2} + \hat{\phi}_3 T_{t-3} + \hat{\phi}_4 T_{t-4} + \epsilon_t \]

\[ (\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3, \hat{\phi}_4) = (0.272, 0.283, 0.096, 0.183) \]

The sample autocorrelation of the residual \( \epsilon_t \) has somewhat significant values at lags 8,9,22, so that \( \epsilon_t \) is probably not white noise.
LA Mortality: Histograms of $\epsilon_t$ and $\eta_t$ residual series.

The figure suggests the tilt model (5),

$$g_\epsilon(x) = e^{\alpha + \beta x^2} g_\eta(x)$$

where $\eta_t$ is the reference with density $g = g_\eta$ and distribution function $G$. In the present case

$$(\hat{\alpha}, \hat{\beta}) = (0.145(0.04), -0.003(0.00086))$$
Estimated $g, G$, comparison of $g$ and the corresponding histogram from $\eta_t$, and prediction for the filtered/sampled mortality data.

Bottom left: Note the fit.
Bottom right: The estimated prediction probabilities obtained from $\hat{G}$ of exceeding 200 “deaths”,

$$P(Y_t > 200 \mid Y_{t-1}, Y_{t-2}, T_{t-1}, \log(CO_t)) \approx 1 - \hat{G}(200 - \exp \{ \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 T_{t-1} + \beta_4 \log(CO_t) \})$$

t=3,...,202, pointing to a strong annual cycle.
Semiparametric Forecasting of U.S. Mortality

a. For each age 1-85 we have a short ts: 1970-2001 (size 32).


c. Each age is predicted from 5 time series corresponding to 5 ages.

d. Autoregression for each age:

\[ x_t = b x_{t-1} + c + \epsilon_t, \quad t = 1970, ..., 2001 \]  

(8)

e. All Age MSE: Lee & Carter (1992) SVD vs SP.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Case</th>
<th>LC</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Step</td>
<td>Total Pop</td>
<td>0.297</td>
<td>0.104</td>
</tr>
<tr>
<td>1 Step</td>
<td>Female</td>
<td>0.619</td>
<td>0.187</td>
</tr>
<tr>
<td>1 Step</td>
<td>White Female</td>
<td>0.645</td>
<td>0.249</td>
</tr>
<tr>
<td>2 Step</td>
<td>Total Pop</td>
<td>0.389</td>
<td>0.180</td>
</tr>
</tbody>
</table>

True (solid), SP Model (dotted), L-C model (dash-dot), 95% CI bounds (dash).
**VaR: SP vs GARCH(1,1)**

GARCH-Normal case:

\[ \tilde{r}_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim N(0, 1) \]
\[ \sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \]

TGARCH-Gamma case:

\[ \tilde{r}_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim \text{Gamma}(4, 2) \]
\[ \sigma_t^2 = \begin{cases} 
\alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \epsilon_{t-1}^2 & : \epsilon_t > 0 \\
\alpha_0 + \alpha_2 \epsilon_{t-1}^2 + \beta_1 \epsilon_{t-1}^2 & : \epsilon_t \leq 0 
\end{cases} \]

GARCH uses \( \tilde{r}_t \) only, assuming it is normal.
SP Uses \( r_t, \tilde{r}_t \), assuming only a quadratic tilt’:

\[ g(x) = \exp(\alpha + \beta x^2) \tilde{g}(x) \]

Define: \( r_t = 0.8 \tilde{r}_t + 0.2 y_t \), with \( y_t \) a normal GARCH(1,1).
The tilt assumption is not satisfied!

Problem: Estimate VaR of the reference \( \tilde{r}_t \).
Normal GARCH case:

<table>
<thead>
<tr>
<th>Given p</th>
<th>SP</th>
<th>Historical</th>
<th>GARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>MSE</td>
<td>0.0709</td>
<td>0.0709</td>
</tr>
<tr>
<td>0.05</td>
<td>MSE</td>
<td>0.0121</td>
<td>0.0158</td>
</tr>
<tr>
<td>0.01</td>
<td>BIAS</td>
<td>1.3295</td>
<td>0.9745</td>
</tr>
<tr>
<td>0.05</td>
<td>BIAS</td>
<td>0.0604</td>
<td>0.0680</td>
</tr>
<tr>
<td>0.01</td>
<td>ER</td>
<td>0.0097</td>
<td>0.0131</td>
</tr>
<tr>
<td>0.05</td>
<td>ER</td>
<td>0.0513</td>
<td>0.0541</td>
</tr>
</tbody>
</table>

TGARCH Case:

<table>
<thead>
<tr>
<th>Given p</th>
<th>SP</th>
<th>Historical</th>
<th>GARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>MSE</td>
<td>0.0066</td>
<td>0.0129</td>
</tr>
<tr>
<td>0.05</td>
<td>MSE</td>
<td>0.0046</td>
<td>0.0047</td>
</tr>
<tr>
<td>0.01</td>
<td>BIAS</td>
<td>0.1243</td>
<td>0.2356</td>
</tr>
<tr>
<td>0.05</td>
<td>BIAS</td>
<td>0.0005</td>
<td>0.0028</td>
</tr>
<tr>
<td>0.01</td>
<td>ER</td>
<td>0.0111</td>
<td>0.0113</td>
</tr>
<tr>
<td>0.05</td>
<td>ER</td>
<td>0.0577</td>
<td>0.0515</td>
</tr>
</tbody>
</table>
We close by noting the curious fact that the tilt model 
(4) is reminiscent of the celebrated relationship between 
the spectral densities of the input, \( f(\omega) \), and output, 
\( f_H(\omega) \), in a linear system with squared gain \( |H^2(\omega)| \), 
\[
    f_H(\omega) = |H(\omega)|^2 f(\omega)
\]
If the input series is operated on by a sequence of \( m \) linear filters, then with output spectral densities \( f_1(\omega), \ldots, f_m(\omega) \)
\[
    f_1(\omega) = |H_1(\omega)|^2 f(\omega) \\
    \vdots \\
    \vdots \\
    f_m(\omega) = |H_m(\omega)|^2 f(\omega)
\]
That is, \( m \) distortions of the same \( f \) as in (4).
**Derivation of SE's**

The asymptotic covariance matrix of the parameter estimates is

\[ \Sigma = S^{-1}V S^{-1} \]  \hspace{1cm} (9)

The matrices \( S, V \) are derived by repeated differentiation of the profile log-likelihood.

First define

\[ \nabla \equiv \left( \frac{\partial}{\partial \alpha_1}, ..., \frac{\partial}{\partial \alpha_q}, \frac{\partial}{\partial \beta_1}, ..., \frac{\partial}{\partial \beta_q} \right)' \]  \hspace{1cm} (10)

Then \( E[\nabla l(\alpha_1, ..., \alpha_q, \beta_1, .., \beta_q)] \equiv 0 \). To obtain the score second moments it is convenient to define \( \rho_m \equiv 1, \ w_m(t) \equiv 1, \)

\[ E_j[h(t)] \equiv \int h(t)w_j(t)dG(t) \]  \hspace{1cm} (11)

and,

\[ A_0(j, j') \equiv \int \frac{w_j(t)w_{j'}(t)dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)} \]  \hspace{1cm} (12)

\[ A_1(j, j') \equiv \int \frac{h(t)w_j(t)w_{j'}(t)dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)} \]  \hspace{1cm} (13)

\[ A_2(j, j') \equiv \int \frac{h(t)h'(t)w_j(t)w_{j'}(t)dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)} \]  \hspace{1cm} (14)

for \( j, j' = 1, ..., q \).
Then, the entries in

\[ V \equiv \text{Var} \left[ \frac{1}{\sqrt{n}} \nabla l(\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_q) \right] \]  

are,

\[
\frac{1}{n} \text{Var} \left( \frac{\partial l}{\partial \alpha_j} \right) = \frac{\rho_j^2}{1 + \sum_{k=1}^{q} \rho_k} \{ A_0(j, j) - \sum_{r=1}^{m} \rho_r A_0^2(j, r) \} 
\]

\[
\frac{1}{n} \text{Cov} \left( \frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \alpha_{j'}} \right) = \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^{q} \rho_k} \{ A_0(j, j') \} - \sum_{r=1}^{m} \rho_r A_0(j, r) A_0(j', r) \}
\]

\[
\frac{1}{n} \text{Cov} \left( \frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \beta_{j'}} \right) = \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^{q} \rho_k} \{ A_0(j, j') E_j[h'(t)] \} - \sum_{r=1}^{m} \rho_r A_0(j, r) A_1'(j, r) \}
\]

\[
\frac{1}{n} \text{Cov} \left( \frac{\partial l}{\partial \beta_{j}}, \frac{\partial l}{\partial \beta_{j'}} \right) = \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^{q} \rho_k} \{ - A_2(j, j') + E_j[h(t)] A_1'(j, j') \} + \sum_{r=1}^{m} \rho_r A_1(j, r) A_1'(j', r) \}
\]
\[ + \frac{1}{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n_j'} \text{Cov}[h(\epsilon_{ji}), h(\epsilon_{jk'})] \] (20)

The last term is 0 for \( j \neq j' \) and \((n_j/n)\text{Var}[h(\epsilon_{j1})]\) for \( j = j' \).
Next, as $n \to \infty$,
\[-\frac{1}{n} \nabla \nabla' l(\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_q) \to S \tag{21}\]

where $S$ is a $q(1 + p) \times q(1 + p)$ matrix with entries corresponding to $j, j' = 1, \ldots, q$,
\[
-\frac{1}{n} \frac{\partial^2 l}{\partial \alpha_j^2} \to \frac{\rho_j}{1 + \sum_{k=1}^q \rho_k} \int \frac{[1 + \sum_{k \neq j}^q \rho_k w_k(t)]w_j(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t) \tag{22}\]
\[
-\frac{1}{n} \frac{\partial^2 l}{\partial \alpha_j \partial \beta_j'} \to \frac{-\rho_j \rho_j'}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(t)w_j'(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t) \tag{23}\]
\[
-\frac{1}{n} \frac{\partial^2 l}{\partial \alpha_j \partial \beta_j} \to \frac{\rho_j}{1 + \sum_{k=1}^q \rho_k} \int \frac{[1 + \sum_{k \neq j}^q \rho_k w_k(t)]w_j(t)h'(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t) \tag{24}\]
\[
-\frac{1}{n} \frac{\partial^2 l}{\partial \alpha_j \partial \beta_j'} \to \frac{-\rho_j \rho_j'}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(t)w_j'(t)h'(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t) \tag{25}\]
\[
-\frac{1}{n} \frac{\partial^2 l}{\partial \beta_j \partial \beta_j'} \to \frac{\rho_j}{1 + \sum_{k=1}^q \rho_k} \int \frac{[1 + \sum_{k \neq j}^q \rho_k w_k(t)]w_j(t)h(t)h'(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t) \tag{26}\]
\[
-\frac{1}{n} \frac{\partial^2 l}{\partial \beta_j \partial \beta_j'} \to \frac{-\rho_j \rho_j'}{1 + \sum_{k=1}^q \rho_k} \int \frac{w_j(t)w_j'(t)h(t)h'(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t) \tag{27}\]


