REGULARIZATION IN TIME SERIES

Yulia R. Gel

Department of Statistics and Actuarial Science,

University of Waterloo, Canada

joint with Peter Bickel, Berkeley

Supported by a grant from
the National Science and Engineering Research Council (NSERC) of Canada

Outline

1. Online Modeling and Forecasting

2. Regularization

3. Asymptotic Consistency of Banded Estimates

4. Selection of an Optimal Regularizer

5. Performance Study

6. Conclusions and Future Research

Online Modeling and Forecasting

Suppose that we observe $y_1, y_2, \ldots$ at every time point t and the goal is to construct a simple and reliable model for online (real time) filtering and forecasting.

The simple and easy choice is to employ an autoregressive (AR) model

$$a(B)y_t = v_t, \tag{1}$$

where

$$a(\lambda) = 1 + a_1\lambda + \ldots + a_q\lambda^q, \qquad a(\lambda) \neq 0, \ \forall\lambda \leq 1, \tag{2}$$

B is a backshift operator ($By_t = y_{t-1}$) and v_t is white noise ($Ev_t = 0$, $Ev_t^2 = \sigma^2$) with $E|v_t|^\beta \leq C < \infty$, $\beta \geq 4$.

Online Modeling and Forecasting: contd

The “true” AR order q is typically unknown. Moreover, when (1) serves only as an approximation of $\{y_t\}$, there exists no “true” AR model. Hence, potentially $q = \infty$ and we get:

$$a(\lambda) = 1 + a_1\lambda + \ldots, \tag{3}$$

where

$$a(\lambda) \neq 0, \ \forall\lambda \leq 1, \quad \sum_{i=1}^{\infty} |a_i| < \infty. \tag{4}$$

Notice that $\{y_t\}$ can be also represented as a causal moving average model of infinite order, MA(∞), i.e.

$$y_t = b(B)v_t, \tag{5}$$

where

$$b(\lambda) = \frac{1}{a(\lambda)} = 1 + b_1\lambda + \ldots \neq 0, \ \forall\lambda \leq 1, \quad \text{and} \quad \sum_{\mathrm{i}=1}^{\infty} |\mathrm{b_i}| < \infty. \tag{6}$$

Hence, the spectral density of $f(\lambda)$ of $\{y_t\}$ satisfies:

$$0 < F_1 < f(\lambda) < F_2, \quad F_1, F_2 > 0 \tag{7}$$

Estimation

To estimate unknown AR(∞), we re-write it in a state-space form:

$$y_t = \Phi'_{t-1}\tau_* + v_t, \tag{8}$$

where $\Phi_{t-1} = (y_{t-1}, y_{t-2}, \ldots, y_1, 0, \ldots)$ and $\tau_* = -(a_1, a_2, \ldots)$ are in $\ell_2(\mathbb{N})$. Since

$$\begin{aligned} &R[k-1]a_1 + R[k-2]a_2 + \ldots = R[k], \\ &R[k] = E\{y_t y_{t+k}\}, \qquad k = 1, 2, \ldots, \end{aligned} \tag{9}$$

we form a Yule-Walker (YW) system of infinite order

$$R\tau = r, \tag{10}$$

where

$$R = \begin{pmatrix} R[0] & R[1] & \ldots \\ R[1] & R[0] & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix} \qquad r = (R[1], R[2], \ldots)'.$$

Since the population covariance R is a positive definite linear operator: $\ell_2(\mathbb{N}) \to \ell_2(\mathbb{N})$, we get a unique solution in $\ell_2(\mathbb{N})$:

$$\tau = R^{-1}r. \tag{11}$$

Truncated Estimation

Clearly, in practice we cannot solve an infinite number of equations (11). Let us choose some truncation order p. Then using an orthogonal projector P_p in $\ell_2(\mathbb{N})$, we obtain a truncated Yule-Walker system

$$P_p R P_p \tau_k = P_p r. \tag{12}$$

Let $R_p = P_p R P_p$. Similarly $r_p = P_p r$ and $\tau_p = P_p \tau$. Clearly, $R_p > 0$. Hence,

$$\tau_p = R_p^{-1} r_p. \tag{13}$$

If $\widehat{R}_{p,n}$ and $\widehat{r}$ are the sample estimates of $R_{p,n}$ and r, then the sample estimate of τ_p, given observations up to time n, is provided by

$$\widehat{\tau}_p = \widehat{R}_{p,n}^{-1} \widehat{r}_p. \tag{14}$$

Model Selection and A Path to Regularization

When we predict, we typically use AIC to estimate p, i.e.

$$\mathrm{AIC(p)} = \ln \widehat{\sigma_{\mathrm{p}}}^2 + \frac{\mathrm{2p}}{\mathrm{n}},$$

where $\widehat{\sigma_p}^2$ is the sample variance of 1-step ahead fcst. For $\{y_t\}$ with short memory, typically $p^{\mathrm{AIC}} = O(\log(n))$.

However, if $\{y_t\} \sim \mathrm{AR}(\infty)$, all finite AR approximations are wrong. To reduce the error of approximation, we need $p \to \infty$ when $n \to \infty$. Hence, for on-line modeling and forecasting, p^{AIC} frequently needs to be refined upon arrival of new data. Thus, all the AR parameters need to be recalculated for a new $\mathrm{AR}(p^{\mathrm{AIC}'})$ model ($p^{\mathrm{AIC}} \neq p^{\mathrm{AIC}'}$), which increases the computational costs.

Model Selection and A Path to Regularization: contd

A model selection criterion is called to be **asymptotically efficient for a same-realization prediction** of an AR(∞) process if the yielded model order p_n^* satisfies the following condition (Ing and Wei, 2005):

$$\lim_{n \to \infty} \sup \frac{E\Big\{(y_{n+1} - \widehat{y}_{n+1}(p_n^*))^2 - \sigma^2\Big\}}{\min_{1 \le p \le K_n} E\Big\{(y_{n+1} - \widehat{y}_{n+1}(p))^2 - \sigma^2\Big\}} \le 1. \tag{15}$$

Similarly, a model selection criterion is said to be **strongly asymptotically efficient** for a same-realization prediction of an AR(∞) process if

$$\lim_{n \to \infty} \sup \frac{E\Big\{(y_{n+1} - \widehat{y}_{n+1}(p_n^*))^2 - \sigma^2\Big\}}{\inf_{\widehat{I}_n \in J_n} E\Big\{(y_{n+1} - \widehat{y}_{n+1}(\widehat{I}_n))^2 - \sigma^2\Big\}} = 1. \tag{16}$$

Here $C_l \le K_n^{2+\delta}/n \le C_u$ for some positive δ, C_l and C_u; J_n is the family of all F_n-measurable random variables taking on values on $\{1, 2, \ldots K_n\}$, and F_n is the σ-algebra generated by $\{y_1, y_2, \ldots, y_n\}$

Model Selection and A Path to Regularization: contd

Ing and Wei (2005) show that AIC is **asymptotically efficient** but not **strongly asymptotically efficient** for AR(∞), and instead it is suggested to select an approximating model based on the conditional mean squared prediction error (MSPE)

$$E\Big\{(y_{n+1} - \widehat{y}_{n+1}(p_n)|y_1, \ldots, y_n)^2 - \sigma^2\Big\}. \tag{17}$$

or its sample estimate. Studies of Ing and Wei (2005) suggest that higher order models that are selected by MSPE can be much more efficient than models selected by AIC for the same sample size. However, a sample covariance matrix of a high order might be a poor estimate of the population covariance matrix.

Hence, an alternative idea is to “overestimate” p, i.e. from the beginning to use a sample cov. matrix of up to a greatest possibly “expected” order that can be employed for modeling and forecasting the observed process $\{y_t\}$, and then to regularize the “oversized” matrix, to ensure that it provides a good estimate of the population covariance matrix.

Regularization

E.g., we can consider a **banding** operator B_k:

$$B_k(R) = (R[i-j]\mathbf{1}_{|i-j|\le k}) = \begin{pmatrix} R[0] & R[1] & \dots & R[k] & 0 & \dots \\ R[1] & R[0] & \dots & R[k-1] & R[k] & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \dots \\ R[k] & R[k-1] & \dots & R[0] & R[1] & \dots \\ 0 & R[k] & \dots & R[1] & R[0] & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

or a **thresholding** operator $T_s = (R[i-j]\mathbf{1}_{|R[i-j]|\ge s})$

Banding B_k preserves the symmetric Toeplitz structure of R and, hence, is a more natural choice for time series. Hence, we get a banded Yule-Walker system

$$B_k(\widehat{R}_{p,n})\widehat{\tau}^{b}_{p,n} = \widehat{r}_p. \tag{18}$$

Note that AIC or other IC correspond to applying an orthogonal projector $P_{p^{\text{AIC}}}$ to $\widehat{R}$, while banding can be viewed as a superposition of two operators, i.e. B_k and P_p.

Asymptotic Consistency of Banded Estimates

Theorem 1. Let y_t be generated by the model (1)-(4) and $\beta \geq 4$. Let $0 < \delta_1 < 0.5$, $0 < \delta_2 < 1$ and $1 - 2\delta_1 - \delta_2 > 0$. If $k \sim (p/n)^{-\delta_2}$, then

$$||B_k(\hat{R}_{p,n}) - R_p||_F = O\left(\left[\frac{p}{n}\right]^{\delta_1}\right), \tag{19}$$

where $\hat{R}_{p,n} = \frac{1}{n}\sum_{k=1}^{n} P_p \Phi_k \Phi_k' P_p$.

In view of Theorem 1, we can now show that banded regularized YW estimate $\hat{\tau}_{p,n}^b$ is a consistent estimate of the truncated vector of population parameters τ_p.

Corollary 1. Under the conditions of Theorem 1, if $p/n \to 0$ as $p \to \infty$ and $n \to \infty$, then

$$||\tau_p - \hat{\tau}_{p,n}^b|| \to 0. \tag{20}$$

Selection of an Optimal Band

We choose an optimal band k by a cross-validation (Bickel and Levina, 2008a and 2008b; Chen and Gel, 2009). We divide the training set Ω of size n into two consecutive segments, Ω_1 and Ω_2 of size n_1 and n_2, where $n_1 \sim n/3$. Then we compare the banded "target" quantity $\widehat{g}_B(\Omega_1)$ with the "target" quantity $\widehat{g}(\Omega_2)$. Here $\widehat{g}(\Omega_2)$ can be viewed as **a proxy to the population "target" quantity** g. The selected "target" quantities of interest can be a covariance matrix or a mean squared error of h-step ahead forecasts.

The optimal band is then selected as a minimizer of the empirical loss function over N splits, i.e.

$$\widehat{k} = \arg\min_k \frac{1}{N} \sum_{\nu=1}^{N} ||\widehat{g}_B(\Omega_{1,\nu}) - \widehat{g}(\Omega_{2,\nu})||. \tag{21}$$

Since $\{y_t\}$ is a time series, random splitting does not work. We can either employ forward–backward splits or select Ω such that $\Omega_1 \bigcup \Omega_2$ is a proper subset of Ω, i.e. $n_1 + n_2 < n$ with $n_2 \approx 2n_1$, and then applying random selection of $\Omega_1 \bigcup \Omega_2$ within Ω.

Selection of an Optimal Band: contd

Our goal now is to show that the rates of convergence for an empirical loss function

$$\frac{1}{N}\sum_{\nu=1}^{N} ||\widehat{g}_B(\Omega_{1,\nu}) - \widehat{g}(\Omega_{2,\nu})||, \tag{22}$$

and the oracle loss function

$$E||\widehat{g}_B(\Omega_{1,\nu}) - \widehat{g}(\Omega_{2,\nu})|| \tag{23}$$

are of the same order and, hence, asymptotically the empirical band $\widehat{k}$ performs as well as the oracle band selection.

Selection of an Optimal Band: contd

Assume w.l.g. that the number of splits $N = 1$. Then we get the following result for banded estimates of covariance matrices:

Theorem 3. Let $\widehat{k}$ and k^o be the band selected from minimizing the empirical and oracle loss functions (22) and (23) respectively. Then, under the conditions of Theorem 1 and for $\beta \geq 4$,

$$||B_{\widehat{k}}\Big(\widehat{R}_p\Big) - R_p||_F = ||B_{k^0}\Big(\widehat{R}_p\Big) - R_p||_F(1 + o(1)). \tag{24}$$

Hence,

$$||B_{\widehat{k}}\Big(\widehat{R}_p\Big) - R_p||_F = O\left(\frac{p}{n}\right)^{\delta}, \quad 0 < \delta < 0.5. \tag{25}$$

Selection of an Optimal Band: contd

A similar result holds if we consider a loss function for prediction. In particular, let

$$\begin{aligned} \text{loss}_\text{o} &= E|y_{t+h} - \Phi'_{p,t}\hat{\tau}_p^b|^2, & (26) \\ \text{loss}_\text{e} &= \frac{1}{N}\sum_{t=T_0}^{T_0+N} |y_{t+h} - \Phi'_{p,t}\hat{\tau}_p^b|^2, & (27) \end{aligned}$$

be a loss function based on the oracle and empirical banding estimates respectively. Let $\hat{b}$ and b^o be the band selected from minimizing the empirical and oracle loss functions (26) and (27) respectively. Then we can state the following asymptotic result on the banded prediction.

Corollary 2. Under the conditions of Theorem 2 and for $\beta \geq 4$,

$$\frac{1}{N}\sum_{t=T_0}^{T_0+N} |y_{t+h} - \Phi'_{p,t}\hat{\tau}_p^{\hat{b}}|^2 = E|y_{t+h} - \Phi'_{p,t}\hat{\tau}_p^{b^o}|^2(1 + o(1)). \quad (28)$$

Remarkably, the empirical loss function (27) for banded prediction can be viewed as a sample approximation to MSPE in (17).

Performance. Simulation Example

Consider the stationary stochastic process described by the transfer function (Mari J. et al., 2000):

$$W(z) = \frac{1 - 0.8762z + 0.0184z^2 + 0.0197z^3 + 0.8591z^4 - 0.7491z^5}{1 - 0.6281z + 0.3597z^2 + 0.2634z^3 - 0.5322z^4 + 0.7900z^5}$$

We choose an optimal model based on AIC and the first 200 observations as a training set:

$$p_{200}^{\mathrm{AIC}} = \arg \min_{0 \le p \le 24} \mathrm{AIC(p)}, \tag{29}$$

and the optimal banding parameter k is selected by cross-validation as follows

$$k^* = \arg \min_{0 \le j \le 24} \frac{1}{50} \sum_{t=150}^{200} [y_{t+1} - \widehat{y}_{t+1}^{j}]^2, \tag{30}$$

i.e. the first 150 observations of the training set are used to estimate the banded model and the next 50 observations are used for cross-validation. Here we apply banding to a 24 × 24-sample information matrix.

Performance. Simulation Example: contd

Based on the models (29)–(30), we construct 1-step ahead out-of-sample predictions of the next 300 observations, i.e. $y_{201}, y_{202}, \ldots, y_{500}$, and calculate the respective root mean squared prediction error (RMSPE):

$$\left\{\frac{1}{300}\sum_{t=201}^{500}[y_{t+1} - \widehat{y}_{t+1}^{\text{reg,p}^*}]^2\right\}^{1/2} - 1. \tag{31}$$

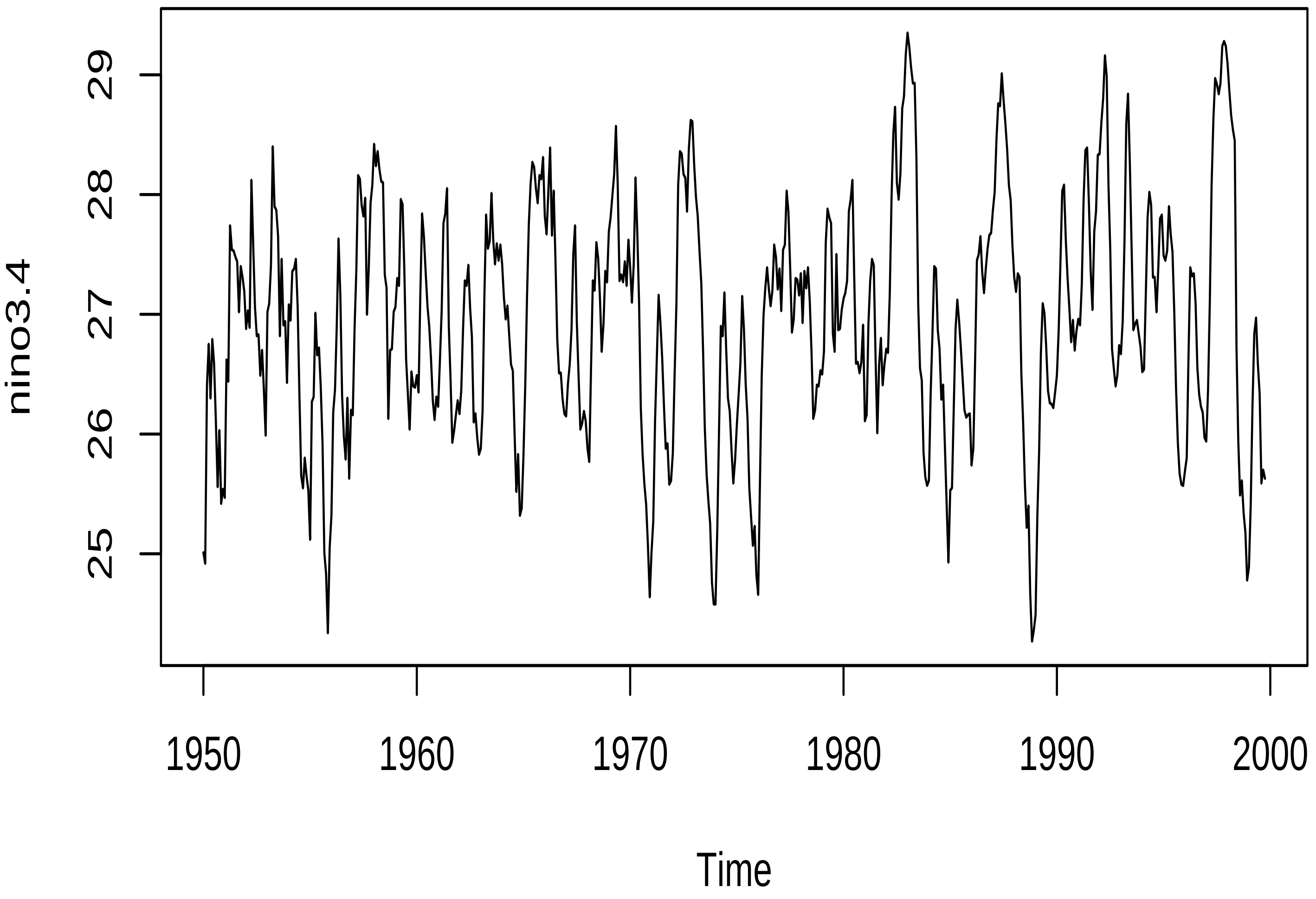
We then apply this procedure to 100 Monte Carlo simulations of $\{y_t\}_{t=1}^{500}$.

MSPE of 1-step prediction	
AR(6)_AIC	B_{21}
1.78	1.10

Hence, banded forecasts provide about 38% smaller MSPE than the AIC-based models.

Performance. Forecasting Niño Region Sea Surface Temperature (SST) Indexes

El Niño is a phenomenon in the equatorial Pacific Ocean manifested in increase of sea surface temperatures (SST). El Niño has a strong impact on local and global climate by affecting atmospheric circulation and, hence, rainfall and temperature. To characterize the nature of El Niño, sea surface temperature (SST) anomalies, i.e. positive deviations from the mean temperature, in certain regions of the Pacific are recorded and analyzed. SST is also used for prediction of future state of the ocean and, hence, El Niño. One of the widely employed SST indexes refers to measurements of ocean in the Nino 3.4 region, bounded by 120W–170W and 5S–5N.



The time series plot of Nino Sea Surface Temperature Indexes

Performance. Forecasting Niño Region Sea Surface Temperature (SST) Indexes

We consider monthly observations of the Niño 3.4 SST index (in C^0) from January 1950 to April 1993. We select the training window of the 120 observations, where the first 90 observations are used to estimate the banded 10×10-covariance matrix and the next 30 observations are employed for cross-validation to obtain an optimal banding parameter. This selection procedures yields an optimal band of 9. The optimal AR(3) model by means of AIC is selected using the whole training set of 120 observations, i.e. from January 1950 to December 1959.

The next 400 observations, i.e. from January 1960 to April 1993, are employed for out-of-sample verification of one-step ahead forecasts, i.e. we compare MSPE

$$\left\{\frac{1}{400}\sum_{t=121}^{520}[y_{t+1}-\widehat{y}_{t+1}^{\text{reg,p}^*}]^2\right\}^{1/2}.$$

We find that MSPE of the AIC-based model is $5.54C^o$ while the respected banded MSPE is $3.67C^o$, which about 34% smaller.

Conclusions and Future Research

- consistency and efficiency properties of a banded regularization as an information criterion;

- a combination of re-enforced Toeplitz-ation and thresholding in order to caputure a more general class of sparse models;

- banding for long memory processes, i.e. AR(∞) with hyperbolically decaying coefficients, nonlinear and locally linear time series.