ASPECTS OF PREDICTION

N. H. BINGHAM, Imperial College London

Birkbeck College, 11 April 2012

This talk is based on joint work with Akihiko Inoue (Hiroshima U.) and Yukio Kasahara (Hokkaido U.). See my surveys
Szegö's theorem and its probabilistic descendants, arXiv:1108.0379,
Multivariate prediction and matrix Szegö theory, arXiv:1203:0962;
BIK, An explicit representation of Verblunsky coefficients, *Statistics and Probability Letters* 82.2 (2012), 403-410.
Background areas: Time series in Statistics; Hardy spaces in Analysis. See e.g.
N. K. Nikolskii, *Operators, functions and systems: an easy reading*. Vol. 1: *Hardy, Hankel and Toeplitz*; Vol 2, *Model operators and systems*, AMS, 2002.

Abstract

The talk concerns prediction theory for stationary stochastic processes in discrete time, in one dimension or (motivated by financial portfolios) many. The basic tool is orthogonal polynomials on the unit circle (OPUC), or its recent multivariate extension (MOPUC). The partial autocorrelation function (PACF), or Verblunsky coefficients, and the Levinson-Durbin algorithm, play key roles.

Setting

Prediction theory for a stationary stochastic process $X = (X_t)$ in discrete time.

Stationarity is a strong assumption!

C. W. J. (Sir Clive) Granger (1934-2009).

Scalar case: orthogonal polynomials on the unit circle (OPUC).

Matrix case (MOPUC): X_t a d-vector (e.g., a portfolio in math. finance – Markowitzian diversification).

There are two main themes.

1. Strong and weak conditions – corresponding to the strong and weak forms of Szegö's limit theorem; needed to exclude long memory (in time), or long-range dependence (in space).

2. A hierarchy: strong, weak and intermediate conditions. The Goldilocks Principle: not too hot/hard/high/..., not too cold/soft/low/..., but just right.

The Kolmogorov Isomorphism Theorem
Let $X = (X_n : n \in Z)$ be a discrete-time, zero-mean, (wide-sense) stationary stochastic process, with autocovariance function $\gamma = (\gamma_n)$, $\gamma_n = E[X_n \overline{X_0}]$ (w.l.o.g., take the variance as 1, so (auto)covariance = (auto)correlation).
Let $\mathcal{H}$ be the Hilbert space spanned by $X = (X_n)$ in the L_2-space of the underlying probability space, with inner product $(X, Y) := E[X\overline{Y}]$ and norm $\|X\| := [E(|X|^2)]^{1/2}$. Write T for the unit circle, the boundary of the unit disc D, parametrised by $z = e^{i\theta}$; unspecified integrals are over T.

Kolmogorov Isomorphism Theorem, KIT. There is a process Y on T with orthogonal increments and a probability measure μ on T with
(i) $X_n = \int e^{in\theta} dY(\theta)$;
(ii) $E[dY(\theta)^2] = d\mu(\theta)$.

(iii) The autocorrelation function γ then has the *spectral representation*
$\gamma_n = \int e^{-in\theta} d\mu(\theta)$.
(iv) One has the *Kolmogorov isomorphism* between $\mathcal{H}$ (the *time domain*) and $L_2(\mu)$ (the *frequency domain*) given by

$$X_t \leftrightarrow e^{it\cdot}, \qquad (KIT)$$

for integer t (as time is discrete).

(i), (ii): Cramér representation of 1942 (Doob X.4; Cramér and Leadbetter 7.5).
(iii) (Herglotz, 1911) follows from (i) and (ii) (Doob; Brockwell & Davis 4.3).
(iv): Kolmogorov, 1941.
This rests on Stone's theorem of 1932 (spectral representation of groups of unitary transformations of linear operators on Hilbert space); see Doob 636-7, Dunford & Schwartz X.5 for spectral theory. //

To avoid trivialities, we suppose in what follows that μ is *non-trivial* – has infinite support. Since for integer t the $e^{it\theta}$ span polynomials in $e^{i\theta}$, prediction theory for stationary processes reduces to approximation by polynomials. This is the classical approach to the main result of the subject, Szegö's theorem.
We write

$$d\mu(\theta) = w(\theta)d\theta/2\pi + d\mu_s(\theta),$$

so w is the *spectral density* (w.r.t. normalized Lebesgue measure), μ_s the *singular part* of μ. By stationarity,

$$E[X_m\overline{X_n}] = \gamma_{|m-n|}.$$

The *Toeplitz matrix* for X, or μ, or γ, is

$$\Gamma := (\gamma_{ij}), \quad \text{where} \quad \gamma_{ij} := \gamma_{|i-j|}.$$

Principal minors of Γ: T_n.

Orthogonal Polynomials on the Unit Circle (OPUC)

Gabor Szegö (1895-1985)

Szegö limit theorem, 1915, Math. Ann.

OPUC, 1920, 1921, MZ

Orthogonal polynomials. AMS Colloquium Publications 23, 1939 [orthogonal polynomials on the real line, OPRL; OPUC, Ch. XI]

U. Grenander and G. Szegö, *Toeplitz forms and their applications*. U. Calif. Press, 1958

Barry Simon (1946-)

OPUC on one foot, 2005, BAMS [survey]

Orthogonal polynomials on the unit circle. Part 1: Classical theory, Part 2: Spectral theory, AMS Colloquium Publications 54.1, 54.2, 2005.

The sharp form of the strong Szegö theorem, 2005, *Contemporary Math.*

Szegö's theorem and its descendants. Princeton UP, 2011.

A. Inoue, 2000, J. Analyse Math., 2008, PTRF

A. Inoue and Y. Kasahara, 2006, Ann. Stat.

Verblunsky's theorem and partial autocorrelation.

$\mathcal{H}_{[-n,-1]}$: subspace of $\mathcal{H}$ spanned by $\{X_{-n}, \ldots, X_{-1}\}$ (finite past at time 0 of length n),

$P_{[-n,-1]}$: projection onto $\mathcal{H}_{[-n,-1]}$ (best linear predictor of X_0 based on the finite past),

$P^{\perp}_{[-n,-1]} := I - P_{[-n,-1]}$: orthogonal projection ($P^{\perp}_{[-n,-1]} X_0 := X_0 - P_{[-n,-1]} X_0$ is the prediction error).

For prediction based on the infinite past:

$\mathcal{H}_{(-\infty,-1]}$: closed lin. span (cls) of X_k, $k \leq -1$,

$P_{(-\infty,-1]}$: corresponding projection, etc.

$\mathcal{H}_n := \mathcal{H}_{(-\infty,n]}$: (subspace generated by) the past up to time n

$\mathcal{H}_{-\infty} := \bigcap_{n=-\infty}^{\infty} \mathcal{H}_n$: *remote past*.

Partial autocorrelation function (PACF):

$\alpha_n := corr(X_n - P_{[1,n-1]} X_n, X_0 - P_{[1,n-1]} X_0)$: correlation between the residuals at times 0, n resulting from (linear) regression on the intermediate values $X_1, \ldots, X_{n-1}$.

$\alpha = (\alpha_n)_{n=1}^{\infty}$.

Unrestricted parametrization of PACF: the only restrictions on the α_n are the obvious ones resulting from their being correlations – $|\alpha_n| \leq 1$ (or avoiding degeneracy, $|\alpha_n| < 1$): the α fill out the infinite-dimensional cube.
Statistics: Barndorff-Nielsen & Schou, 1973, J. Multiv. An., F. L. Ramsey, 1974, Ann. Stat.
Analysis: Samuel Verblunsky, 1935, 1936, JLMS.
By contrast, the correlation function $\gamma = (\gamma)_n$ again has each $|\gamma_n| \leq 1$, but the γ fill out only part of the inf-dim cube (specified by determinental inequalities).
Szegö recursion (= Levinson-Durbin algorithm). Let P_n be the orthogonal polynomials on the unit circle (OPUC) w.r.t. m. Then

$$P_{n+1}(z) = zP_n(z) - \bar{\alpha}_{n+1}P_n^*(z),$$

where for any polynomial Q_n of degree n,

$$Q_n^*(z) := z^n\overline{Q_n(1/\bar{z})}$$

are the *reversed polynomials*. Herglotz and Verblunsky theorems:

$$\alpha \leftrightarrow \mu \leftrightarrow \gamma.$$

Weak condition: Szegö's condition.
Write σ^2 for the one-step mean-square prediction error:

$$\sigma^2 := E[(X_0 - E(X_0|X_k, k<0))^2].$$

Call X *non-deterministic* (ND) if $\sigma > 0$, *deterministic* if $\sigma = 0$ (i.e. iff $X_n \in \mathcal{H}_{-\infty}$ for each n – the remote past dominates).
Wold decomposition (von Neumann in 1929, Wold in 1938):

$$X_n = U_n + V_n,$$

with V deterministic and U_n a moving average:

$$U_n = \sigma \sum_0^\infty m_j \xi_{n-j},$$

ξ_j iid $N(0,1)$ (so U absent if $\sigma = 0$).
Kolmogorov's formula (1941):

$$\sigma^2 = \exp(\frac{1}{2\pi}\int \log w(\theta) d\theta) =: G(\mu) > 0, \quad (K)$$

(μ_s plays no role; on the right, $G(\mu)$ is the *geometric mean* of μ. So:

Szegö's theorem: $\sigma > 0$ iff

$$\log w \in L_1. \qquad (Sz)$$

When also the remote past is trivial –

$$\mathcal{H}_{-\infty} = \{0\}, \quad i.e. \quad \mu_s = 0$$

– call X *purely non-deterministic*, or (PND):

$$(PND) = (ND) + (\mu_s = 0) = (Sz) + (\mu_s = 0).$$

Hardy spaces (see e.g. P. L. Duren, *Theory of H^p spaces*, AP, 1974). Define the *Szegö function*

$$h(z) := \exp(\frac{1}{4\pi}\int(\frac{e^{i\theta}+z}{e^{i\theta}-z})\log w(\theta)d\theta) \qquad (z \in D). \qquad (OF)$$

Because $\log w \in L_1$ by (Sz), $H := h^2$ is an *outer function* for H_1 (whence the name (OF) above). By Beurling's canonical factorization theorem,
(i) $h \in H_2$.
(ii) The radial limit

$$H(e^{i\theta}) := \lim_{r\uparrow 1} H(re^{i\theta})$$

exists a.e., and

$$|H(e^{i\theta})| = |h(e^{i\theta})|^2 = w(\theta)$$

(thus h may be regarded as an 'analytic square root' of w). The following are equivalent:
(i) Szegö condition $(Sz) = (ND)$, i.e. $\sigma > 0$;
(ii) PACF $\alpha = (\alpha_n) \in \ell_2$. Then
(iii) MA coefficients $m = (m_n) \in \ell_2$;
(iv) Szegö function $h(z) := \sum_{n=0}^{\infty} m_n z^n \in H_2$.

Strong condition 1: Baxter's theorem
We also have the following stronger equivalent conditions (Glen Baxter, 1961, 1962, 1963; Simon Vol. 1, Ch. 5):
(i) PACF $\alpha \in \ell_1$ (Baxter's condition, (B));
(ii) autocorrelation $\gamma \in \ell_1$, and μ is abs. cts with continuous positive density:

$$\min_\theta w(\theta) > 0.$$

(iii) MA coefficients $m \in \ell_1$, $\mu_s = 0$ and μ is ac with continuous positive density w.

Long-range dependence (LRD)

Physics: spatial LRD, phase transitions.

Statistics: LRD in time; see e.g. Cox's survey of 1984 (*Selected Papers* Vol. 2 (2005), TS3), or

Jan Beran, *Statistics for long-memory processes*, Ch&H, 1994.

There was no precise definition of LRD, but two leading candidates, both involving the covariance γ:

(i) LRD is non-summability: $\gamma \notin \ell_1$.

(ii) LRD is covariance decaying like a power: $\gamma_n \sim c/n^{1-2d}$ as $n \to \infty$, for some parameter $d \in (0, 1/2)$ (d for differencing) and constant $c \in (0, \infty)$ (and so $\sum \gamma_n = \infty$).

Motivated by Baxter's theorem, one now has

Definition (L. Debowski, 2007, SPL; Inoue, 2008, PTRF): LRD is $\alpha \notin \ell_1$.

Note. 1. (ii) above may be generalized to γ_n regularly varying, or $w(t)$ regularly varying.

2. Hurst parameter $H := d + 1/2 \in (1/2, 1)$.

3. For $d \in (0, \frac{1}{2})$, $\ell(.)$ slowly varying, the following class of prototypical long-memory examples is considered in Inoue-Kasahara 2006:

$$\gamma_n \sim \ell(n)^2 B(d, 1-2d)/n^{1-2d},$$

$$m_n \sim \ell(n)/n^{1-d},$$

$$r_n \sim \frac{d \sin(\pi d)}{\pi} . \frac{1}{\ell(n)} . 1/n^{1+d}$$

($r = (r_n)$: autoregressive (AR) coefficients).
4. They also consider $FARIMA(p, d, q)$.

Strong condition 2: strong Szegö condn
This is motivated by two areas of physics.
1. The cepstrum.
J. W. Tukey and collaborators, 1963: distinguishing the signature of the underground explosion in a nuclear weapon test from that of an earthquake. Used the cepstrum $L = (L_n)$: Fourier coefficients of $\log w$ (cepstrum: spectrum + reflection, for echo: hard c). This was

used by Bloomfield in his time-series models (alternative to Box-Jenkins $ARMA(p,q)$).
2. The strong Szegö limit theorem, Szegö (1952):

$$\frac{\det\ T_n}{G(\mu)^n} \to E(\mu) := \exp\{\sum_1^\infty kL_k^2)\} \quad (n \to \infty).$$

Taking logs gives the (weak) Szegö limit theorem of 1915:

$$(\log \det\ T_n)/n \to G(\mu).$$

Motivation: Onsager's work in the two-dimensional Ising model, and in particular *Onsager's formula*, giving the existence of a critical temparature T_c and the decay of the magnetization as the temperature $T \downarrow T_c$.
Write $H^{1/2}$ for the subspace of ℓ_2 of sequences $a = (a_n)$ with

$$\|a\|^2 := \sum_n (1 + |n|)|\alpha_n|^2 < \infty$$

('1' on the right to give a norm, or $\|.\|$ vanishes on the constant functions) – a Sobolev space (also a Besov space, whence the alternative notation $B_2^{1/2}$). This plays the role here of ℓ_2 for Szegö's theorem and ℓ_1 for Baxter's theorem. Note that, although ℓ_1 and $H^{1/2}$ are close in that a sequence (n^c) of powers belongs to both or neither, neither contains the other (consider $a_n = 1/(n \log n)$, $a_n = 1/\sqrt{n}$ if $n = 2^k$, 0 otherwise).

Ibragimov's version of the strong Szëgo limit theorem: if $(Sz) = (ND)$ holds and $\mu_s = 0$, then

$$G(\mu) = \prod_{j=1}^{\infty} (1 - |\alpha_j|^2)^{-j} = \exp(\sum_{n=1}^{\infty} nL_n^2)$$

(all may be infinite). The infinite product converges iff the *strong Szegö condition* holds:

$$\alpha \in H^{1/2}, \qquad (sSz)$$

or equivalently by above

$$L \in H^{1/2}. \qquad (sSz')$$

The Golinski-Ibragimov theorem states that, under (Sz), finiteness forces $\mu_s = 0$.
Borodin-Okounkov formula (2000; Geronimo & Case, 1979).
This turns the strong Szegö limit theorem above from analysis to algebra. In terms of operator theory and in Widom's notation, the result is

$$\frac{det\ T_n(a)}{G(a)^n} = \frac{det(I - Q_n H(b) H(\tilde{c}) Q_n)}{det(I - H(b)H(\tilde{c}))},$$

for a a sufficiently smooth function without zeros on the unit circle and with winding number 0. Then a has a Wiener-Hopf factorization $a = a_- a_+$; $b := a_- a_+^{-1}$, $c := a_-^{-1} a_+$; $H(b)$, $H(\tilde{c})$ are the Hankel matrices $H(b) = (b_{j+k+1})_{j,k=0}^{\infty}$, $H(\tilde{c}) = (c_{-j-k-1})_{j,k=0}^{\infty}$, and Q_n is the orthogonal projection of $\ell^2(1, 2, \ldots)$ onto $\ell^2(\{n, n+1, \ldots\})$. By Widom's formula,

$$1/det(I - H(b)H(\tilde{c})) = \exp\{\sum_{k=1}^{\infty} k L_k^2\} =: E(a)$$

(see e.g. Simon 1, Th. 6.2.13), and $Q_n H(b) H(\tilde{c}) Q_n \to 0$ in the trace norm, whence

$$det\ T_n(a)/G(a)^n \to E(a),$$

the strong Szegö limit theorem.

ϕ-mixing

$$\phi(n) := E \sup\{|P(A|\mathcal{F}^0_{-\infty}) - P(A)| : A \in \mathcal{F}^\infty_n\};$$

$$\rho(n) := \rho(\mathcal{F}^0_{-\infty}, \mathcal{F}^\infty_n),$$

$$\rho(\mathcal{A}, \mathcal{B}) := \sup\{\|E(f|\mathcal{B}) - Ef\|_2 / \|f\|_2 : f \in L_2(\mathcal{A})\}.$$

Call X *ϕ-mixing* if $\phi(n) \to 0$ as $n \to \infty$, *ρ-mixing* if $\rho(n) \to 0$.

We quote: ϕ-mixing (regarded here as 'strong') implies ρ-mixing ('intermediate' – below). The spectral characterization for ϕ-mixing is

$$\mu_s = 0, \qquad w(\theta) = |P(e^{i\theta})|^2 w^*(\theta),$$

where P is a polynomial with its roots on the unit circle and the cepstrum $L^* = (L^*_n)$ of w^* satisfies (sSz).

Intermediate conditions (four, in decreasing order of strength)

1. Complete regularity (or ρ-mixing): ρ-mixing coefficients $\rho(n) \to 0$. Spectral characterization

$$\mu_s = 0, \qquad w(\theta) = |P(e^{i\theta})|^2 w^*(\theta),$$

where P is a polynomial with its roots on the unit circle and for all $\epsilon > 0$,

$$\log\ w^* = r_\epsilon + u_\epsilon + \tilde{v}_\epsilon,$$

where r_ϵ is continuous, u_ϵ, v_ϵ are real and bounded, and $\|u_\epsilon\| + \|v_\epsilon\| < \epsilon$ (Ibragimov-Rozanov, V.2 Th. 3; cf. Fefferman-Stein decomposition). Alternatively,

$$\mu_s = 0, \qquad w(\theta) = |P(e^{i\theta})|^2 w^*(\theta),$$

where P is a polynomial with its roots on the unit circle and

$$\log\ w^* = u + \tilde{v},$$

with u, v real and continuous (Sarason; Helson and Sarason).

2. Positive angle: the Helson-Szegö and Helson-Sarason conditions.
For subspaces A, B of $\mathcal{H}$, the *angle* between A and B is defined as

$$\cos^{-1} \sup\{|(a,b)| : a \in A, b \in B\}.$$

Then A, B are at a *positive angle* iff this supremum is < 1. X satisfies the *positive angle condition*, (PA), if for some time lapse k the past $cls(X_m : m < 0)$ and the future $cls(X_{k+m} : m \geq 0)$ are at a positive angle, i.e. $\rho(0) = \ldots \rho(k-1) = 1, \rho(k) < 1$, which we write as $PA(k)$ (Helson and Szegö, $k = 1$; Helson and Sarason, $k > 1$). Spectral characterization:

$$\mu_s = 0, \qquad w(\theta) = |P(e^{i\theta})|^2 w^*(\theta),$$

where P is a polynomial of degree $k-1$ with its roots on the unit circle and

$$\log\ w^* = u + \tilde{v},$$

where u, v are real and bounded and $\|v\| < \pi/2$ ([IR] V.2, Th. 3, Th. 4). The Helson-Szegö condition ($PA(1)$) coincides with *Muckenhoupt's condition* A_2 in analysis:

$$\sup_I((\frac{1}{|I|}\int_I w(\theta)d\theta)(\frac{1}{|I|}\int_I \frac{1}{w(\theta)}d\theta)) < \infty, \quad (A_2)$$

where $|.|$ is Lebesgue measure and the supremum is taken over all subintervals I of the unit circle T. See e.g. Hunt, Muckenhoupt and Wheeden [HMW]. Reducing $PA(k)$ to $PA(1)$ (by sampling every kth time point), we then have complete regularity ($\rho(n) \to 0$) implies $PA(1) = (A_2)$.

3. Pure minimality

Interpolation problem: find best linear interpolation of a missing value, X_0 say, from the others. Write $H_n' := cls\{X_m : m \neq n\}$ for the

closed linear span of the values at times other than n. X is *minimal* if $X_n \notin H_n'$, *purely minimal* if $\bigcap_n H_n' = \{0\}$. Spectral condition for minimality is (Kolmogorov in 1941) $1/w \in L_1$ (and for pure minimality, this $+ \mu_s = 0$). Under minimality, the relationship between the moving-average coefficients $m = (m_n)$ and the autoregressive coefficients $r = (r_n)$ becomes symmetrical, and one has the equivalences (i) minimal; (ii) AR coefficients $r = (r_n) \in \ell_2$; (iii) $1/h \in H_2$.

4. Rigidity; (LM), (CND), (IPF).

Rigidity; the Levinson-McKean condition.

Call $g \in H^1$ *rigid* if is determined by its phase:

$$f \in H^1 \quad (f \text{ not identically 0}), \quad f/|f| = g/|g| \quad a.e.$$

implies $f = cg$ for some positive constant c (Sarason, Nakazi, de Leeuw and Rudin, Levinson and McKean). Call the condition that μ be ac with spectral density $w = |h|^2$ with h^2 rigid, or determined by its phase, the *Levinson-McKean condition*, (LM).

Complete non-determinism; intersection of past and future.
(i) *complete non-determinism*,

$$\mathcal{H}_{(-\infty,-1]} \cap \mathcal{H}_{[0,\infty)} = \{0\}, \qquad (CND)$$

(ii) the *intersection of past and future* property,

$$\mathcal{H}_{(-\infty,-1]} \cap \mathcal{H}_{[-n,\infty)} = \mathcal{H}_{[-n,-1]} \qquad \forall n \ (IPF)$$

$$(CND) \quad \Leftrightarrow \quad (IPF) \quad \Leftrightarrow \quad (LM).$$

First equivalence: Inoue & Kasahara 2006; second equivalence: (LM) Kasahara & Bingham, 2012. They are stronger than (PND), itself stronger than the weak condition $(ND) = (Sz)$.

Multivariate prediction

If X is *vector-valued* (e.g., the price vector in a portfolio in math. finance), then μ, α are *matrix-valued*, and OPUC becomes matrix OPUC or MOPUC. Most of the above still goes through; see my sequel *Multivariate prediction and matrix Szegö theory* ((LM), CND), (IPF): work in progress).

Stochastic v. independent v deterministic If all the X_n are independent, no prediction is possible: the *free case*, μ normalised Lebesgue measure, $\gamma_n = \delta_{n0}$, $\alpha_n \equiv 0$: a boundary between our stochastic case and deterministic chaos (dynamical systems, non-linearity etc.) Distinguishing these: Lucas Lacasa, visibility graph, 2008, 2009, 2012.

Non-stationarity. Extensions are possible. Probability: theory of *harmonisable* processes (Cramér; M. M. Rao); KIT: Y. Kakihara, 2001. Statistics: see e.g. Dégerine & Lambert-Lacroix, JTSA 2002.

Implementation: statistics. PACF estimation (Dégerine et al.); density estimation on the circle.
Implementation: financial applications. Lots of good problems – e.g., hedging with options of different maturities (Andrea Macrina, KCL).