ORTHOGONAL POLYNOMIALS ON THE UNIT CIRCLE AND PREDICTION

N. H. BINGHAM, Imperial College London

Sussex, 28 October 2010

Joint work with

Akihiko INOUE, Hiroshima University and

Yukio Kasahara, U. Hokkaido, Sapporo

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 (Review, Paul Nevai, BAMS 44 (2007), 447-470).
The sharp form of the strong Szegö theorem, 2005, *Contemporary Math.*
A. Inoue, 2000, J. Analyse Math., 2008, PTRF
A. Inoue and Y. Kasahara, 2006, Ann. Stat.
Background areas: Time series in Statistics; Hardy spaces in Analysis

Verblunsky's theorem and partial autocorrelation.

$X = (X_n : n \in Z)$: discrete-time, zero-mean, real, (wide-sense) stationary stochastic process, with autocovariance function $\gamma = (\gamma_n)$,

$$\gamma_n = E[X_0 X_n].$$

Herglotz' theorem: *spectral representation*

$$\gamma_n = \int e^{-in\theta} d\mu(\theta)$$

for some *spectral measure* μ on the unit circle $z = e^{i\theta}$ (boundary of unit disk D). Throughout, take μ *non-trivial* – with infinite support. Take $var X_n = 1$: γ is the *autocorrelation function* and μ is a *probability measure*.

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

w is the *spectral density* and μ_s is the *singular part* of μ. By stationarity,

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

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

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

positive definite. Principal minors T_n (below).
$\mathcal{H}$: Hilbert space spanned by $X = (X_n)$ in the L_2-space of the underlying probability space, with inner product $(X, Y) := E[XY]$ and norm $\|X\| := [E(X^2)]^{1/2}$.
$\mathcal{H}_{[-n,-1]}$: 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 – $|\alpha_n| \in [-1, 1]$ (or avoiding degeneracy, $|\alpha_n| \in (-1, 1)$): 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 $|\gamma_n| \in [-1, 1]$, but the γ fill out only part of the inf-dim cube (specified by determinental inequalities).
Szegö recursion.
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*. Szegö recursion (called the *Durbin-Levinson algorithm* in the time-series literature is the OPUC analogue of Favard's theorem (three-term recurrence relation) for OPRL.

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): if $\sigma > 0$,

$$X_n = U_n + V_n,$$

with V deterministic and U_n a moving average:

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

ξ_j iid $N(0, \sigma^2)$.
Kolmogorov's formula (1941):

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

(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*

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

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 $(Sz) = (ND)$;
(ii) PACF $\alpha = (\alpha_n) \in \ell_2$;
(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 (Inoue, 2008, PTRF; L. Debowski, preprint): 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):

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

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

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

(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,

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

(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.

Absolute regularity and β-mixing

Weak dependence: *mixing* conditions (general); *regularity* conditions (Gaussian case). We assume for simplicity that our process is Gaussian, which brings the two hierarchies of mixing

and regularity conditions together. Gaussian case: Ibragimov & Rozanov; [IR] IV.3 (information regularity), IV.4 (absolute regularity), V (complete regularity).
Mutual information at lag n:

$$I_n := I(\{X_t : t < 0\}, \{X_t \geq n\});$$

for the definition of the mutual information on the right, see e.g. Ibragimov & Linnik, IV.1. If $I_n \to 0$ as $n \to \infty$, the process is called *information regular* (or I-mixing). The mutual information may be infinite, but is finite iff $L \in H^{1/2}$, and then

$$I = \frac{1}{2}\|L\|^2 = \frac{1}{2}\sum_{k=1}^{\infty} kL_k^2.$$

The strongest of the conditions we study here is *information regularity* (or I-mixing), equivalently, *absolute regularity* (or β-mixing). For their equivalence, see [IR] IV.3, p. 128. Spectral characterization for abs. regularity:

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

where P is a polynomial with its roots on the unit circle and the cepstrum $L^* = (L_n^*)$ of w^* satisfies (sSz) ([IR] IV.4, p. 129). Thus absolute regularity is weaker than (sSz).

Intermediate conditions (four, in decreasing order of strength)

1. Complete regularity (or ρ-mixing): ρ-mixing coefficients $\rho(n) \to 0$ [IR]. 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 $\|u_\epsilon\| + \|v_\epsilon\| < \epsilon$ ([IR], 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).

Complete regularity (ρ-mixing) is equivalent to strong mixing (α-mixing) ([IR] IV.1, (1.9) and (1.16)). Since $\alpha(.) \leq \beta(.)$ ([IR], 109-110), this shows that absolute regularity (itself weaker than (sSz)) implies complete regularity. Since under Baxter's condition (B) w is continuous and positive, $\log w$ is bounded, so (B) implies complete regularity by the above. This justifies our characterization of (B), (sSz) and absolute regularity as strong but complete regularity as intermediate.

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:

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

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 (IK06).
(i) *complete non-determinism*,

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

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

$$\mathcal{H}_{(-\infty,-1]} \cap \mathcal{H}_{[-n,\infty)} = \mathcal{H}_{[-n,-1]} \qquad (n = 1, 2, \ldots) \qquad (IPF)$$

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

These are weaker than pure minimality, but stronger than (PND), itself stronger than the weak condition $(ND) = (Sz)$.