Optimal designs for correlated observations: A reliable asymptotic theory

Anatoly Zhigljavsky (Cardiff)

Collaborators:

Holger Dette (Bochum),

Andrey Pepelyshev (Aachen ?),

Karl Michael Schmidt, Nikolai Leonenko (Cardiff)

Cambridge, Newton Institute, July 19, 2011

Generic model:

$$y(t) = \theta^T f(t) + \varepsilon(t) = \theta_1 f_1(t) + \ldots + \theta_m f_m(t) + \varepsilon(t), \quad t \in \mathbb{T}$$

where

$$\mathbb{E}\varepsilon(t) = 0, \quad \mathbb{E}\varepsilon^2(t) = \sigma^2, \quad \mathbb{E}\varepsilon(t)\varepsilon(t') = \sigma^2\rho(t - t') \text{ for } t, t' \in \mathbb{T}$$

Design: a set of points $t_j \in \mathbb{T}$ $(j = 1, \ldots, N)$.
Aim: Efficient estimation of parameters θ.

Some issues to think about before we start designing:

- Is N large?
- Is $\mathbb{T}$ an interval?
- Is the autocorrelation function $\rho(t)$ known exactly?
- Which estimator can/should we use?
- What exactly is the model?
- What is the design optimality criterion?
- etc

Models:

$$y(t) = \theta + \varepsilon(t), \quad t \in \mathbb{T}$$

$$y(t) = \theta f(t) + \varepsilon(t), \quad t \in \mathbb{T}$$

$$y(t) = \theta^T f(t) + \varepsilon(t) = \sum_{i=1}^{m} \theta_i f_i(t) + \varepsilon(t), \quad t \in \mathbb{T}$$

$$y(t) = \eta(\theta, t) + \varepsilon(t), \quad t \in \mathbb{T}$$

$$y(t) = \eta(t) + \varepsilon(t), \quad t \in \mathbb{T}, \quad \eta \in \mathcal{F}$$

$$y(t) = \zeta(\theta, t), \quad t \in \mathbb{T}$$

More: ask Henry Wynn, Rainer Schwabe, Radoslav Harman, Werner Müller and many others

Estimators:

(i) **BLUE** = MLE (Gaussian errors) = Weighted LSE:

$$\hat{\theta} = (\mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Sigma}^{-1} Y$$

where $\mathbf{X} = (f_i(t_j))_{j=1,\dots,N}^{i=1,\dots,m}$, $\mathbf{\Sigma} = (\rho(t_i - t_j))_{i,j=1,\dots,N}$.
The covariance matrix is

$$\mathbb{D}(\hat{\theta}) = \sigma^2 (\mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{X})^{-1} .$$

(ii) Standard (non-weighted) **LSE**:

$$\tilde{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y$$

$$\mathbb{D}(\tilde{\theta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Sigma} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$$

Gauss-Markov theorem:

$$\mathbb{D}(\tilde{\theta}) - \mathbb{D}(\hat{\theta}) \geq 0$$

How large is the difference $\mathbb{D}(\tilde{\theta}) - \mathbb{D}(\hat{\theta})$?

BLUE asymptotically ($N \to \infty$) coincides with LSE for the linear models with certain correlation functions, see Rao (1967), Kruskal (1968).

Moreover, consider the model $y(x) = \theta + \varepsilon(x)$ and define the (linear unbiased) estimate $\hat{\theta}(G) = \int y(x) dG(x)$, where $G(x)$ is a c.d.f. of a signed probability measure.

Grenander (1950): $\hat{\theta}(G^*)$ is BLUE if and only if $\int \rho(u-t) dG^*(u)$ is constant for all $t \in \mathbb{T}$. Consequently, if $G^*(t)$ is a c.d.f. of a probability measure, then LSE coincides with BLUE and an asymptotic optimal design for LSE is also an asymptotic optimal design for BLUE.

Hajek (1956) proved that $G^*(t)$ is a c.d.f. a non-signed probability measure if the correlation function $\rho(t)$ is convex on the interval $(0, \infty)$. AZ,HD,AP (2010) have shown that $G^*(t)$ is a proper c.d.f. for certain families of correlation functions including non-convex ones. Result of Grenander (1950) was extended by Näther (1985) to the case of random fields with constant mean.

$N \nrightarrow \infty$ **(non-asymptotic setting):**

Algorithms for construction of optimal designs (BLUE):
(i) Direct optimization;

(ii) Brimkulov, Krug, Savanov (1986); Näther (1985); Pazman, Müller (2001, 2003); Dette, Kunert, Pepelyshev (2007); Uciński, Atkinson (2010) as well as many other papers

Algorithms for construction of optimal designs (LSE):
similar but simpler; multiplicative algorithms are available

How large should N be anyway?

Approximate design ($N \to \infty$, asymptotic setting)

Sacks and Ylvisaker (1966, 1968, 1970), Näther (1985), Bickel-Herzberg etc.: the design points $\{t_1, \ldots, t_N\}$ are generated by the quantiles of a distribution function, that is

$$t_{iN} = a\Big((i-1)/(N-1)\Big),\ i = 1, \ldots, N,$$

where the function $a : [0,1] \to \mathbb{T}$ is the inverse of a c.d.f..
That is, design points are quantiles of a c.d.f., call this c.d.f. $G(t)$.

Sacks and Ylvisaker approach (1966, 1968, 1970):

Weighted LSE, $N \to \infty$,

$\mathbb{T}$ is a fixed interval,

$f_i(t) = \int \rho(t,s)\phi_i(s)ds$ (Hilbert space with reproducing kernel).

$$\mathbb{D}(\xi_N^*) \to \mathbb{D}^*$$

Three observations:

– the variances of the estimators $\hat{\theta}_i$ do not tend to zero as N increases;

– the optimal design depends only on the derivatives of $\rho(t,s)$ at $t = s$;

– there is no limiting transition to the case of uncorrelated errors.

LSE estimators

If ξ_N denotes a design with N points the covariance matrix of the estimate $\tilde{\theta} = \tilde{\theta}_{\xi_N}$ is

$$\mathbb{D}(\tilde{\theta}) = \sigma^2 \mathbb{D}(\xi_N) = \sigma^2 \, M^{-1}(\xi_N) B(\xi_N, \xi_N) M^{-1}(\xi_N),$$

For an arbitrary design ξ, the matrices $M(\xi)$ and $B(\xi, \xi)$ are

$$\begin{aligned} M(\xi) &= \int f(u) f^T(u) \xi(\mathrm{d}u), \\ B(\xi, \xi) &= \iint \rho(u - v) f(u) f^T(v) \xi(\mathrm{d}u) \xi(\mathrm{d}v). \end{aligned}$$

and the matrix

$$\mathbb{D}(\xi) = M^{-1}(\xi) B(\xi, \xi) M^{-1}(\xi),$$

is the covariance matrix of design ξ; it can be defined for any probability measure ξ supported on the design space $\mathbb{T}$ such that the matrices $B(\xi, \xi)$ and $M^{-1}(\xi)$ are well-defined.

Covariance (matrix) of a design ξ:

$$y(t) = \theta + \varepsilon(t) \quad \Rightarrow \quad \mathbb{D}(\xi) = \int\int \rho(u-v)\xi(du)\xi(dv)$$

$$y(t) = \theta f(t) + \varepsilon(t) \quad \Rightarrow \quad \mathbb{D}(\xi) = \frac{\int\int f(u)f(v)\rho(u-v)\xi(du)\xi(dv)}{\left(\int f^2(u)\xi(du)\right)^2}$$

$$y(t) = \theta^T f(t) + \varepsilon(t) \quad \Rightarrow \quad \mathbb{D}(\xi) = M^{-1}(\xi)B(\xi,\xi)M^{-1}(\xi),$$

where

$$M(\xi) = \int f(u)f^T(u)\xi(\mathrm{d}u), \quad B(\xi,\xi) = \int\int \rho(u-v)f(u)f^T(v)\xi(\mathrm{d}u)\xi(\mathrm{d}v).$$

How important is the convexity or strict convexity of the optimality criterion $\Phi(\mathbb{D}(\xi))$?

Convexity of the optimality criterion, Näther(1985)

Model $y(t_j) = \theta + \varepsilon(t_j)$,

$$\mathbb{D}(\xi) = \int\int \rho(x-y)\xi(dx)\xi(dy)$$

Consider $\xi_\alpha = (1-\alpha)\xi_0 + \alpha\xi_1$. Then

$$\mathbb{D}(\xi_\alpha) = (1-\alpha)\mathbb{D}(\xi_0) + \alpha\mathbb{D}(\xi_1) - \alpha(1-\alpha)A$$

where

$$\begin{aligned} A &= \int\int \rho(x-y)[\xi_0(dx)\xi_0(dy) + \xi_1(dx)\xi_1(dy) - 2\xi_0(dx)\xi_1(dy)] \\ &= \int\int \rho(x-y)\eta(dx)\eta(dy) \quad \text{with} \quad \eta(dx) = \xi_0(dx) - \xi_1(dx) \end{aligned}$$

The function $K(x,y) = \rho(x-y)$ is positive definite in view of the Bohner-Khintchine theorem. Hence $A > 0$ implying $\mathbb{D}(\xi)$ is a strictly convex functional.

Optimality (equivalence) theorem, Näther(1985)

Consider $\mathbb{D}(\xi) = \int\int \rho(x-y)\xi(dx)\xi(dy)$.

Set $\phi(\xi, x) = \int \rho(x-y)\xi(dy)$, $\xi_\alpha = (1-\alpha)\xi_0 + \alpha\xi_1$

Lemma.

$$\frac{\partial \mathbb{D}(\xi_\alpha)}{\partial \alpha}|_{\alpha=0} = 2\left[\int \phi(\xi_1, x)\xi_0(dx) - \mathbb{D}(\xi_0)\right]$$

Theorem. ξ^* is optimal iff

$$\min_x \phi(\xi^*, x) = \mathbb{D}(\xi^*)$$

or

$$\phi(\xi^*, x) = \text{const}, \quad \text{for } \xi^* - \text{almost all } x\,.$$

Behaviour of optimal designs: AZ, HD, AP

Let $\mathbb{D}(\xi) = \int\int \rho(x-y)\xi(dx)\xi(dy)$, $\mathbb{T} = [-1,1]$, $\rho(t) = \max\{0, 1-\lambda|t|\}$.

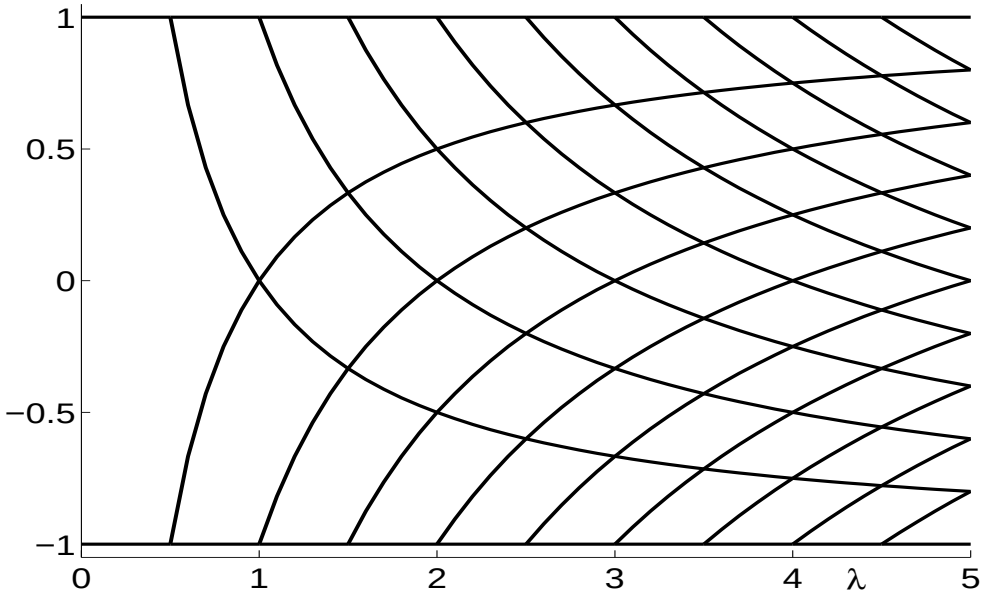

Figure 1: *Support points of the optimal designs for estimating mean.*

Approach of Bickel-Herzberg (1977, 1979)

The correlation function changes with N: $\rho_N(t) = \rho(Nt)$.
Alternatively we can fix $\rho(t)$ but expand the design interval proportionally to the number of observation points N. Asymptotically (as $N \to \infty$) the covariance matrix is proportional to

$$\mathbb{D}(\xi) = M^{-1}(\xi) R(\xi) M^{-1}(\xi)$$

where $\xi(dt) = p(t)dt$,

$$R(\xi) = \frac{1}{1-\alpha} \left(\int f(t) f^T(t) Q(1/p(t)) p(t)\, dt \right), \quad Q(u) = \sum_{j=1}^{\infty} \rho(ju)\,.$$

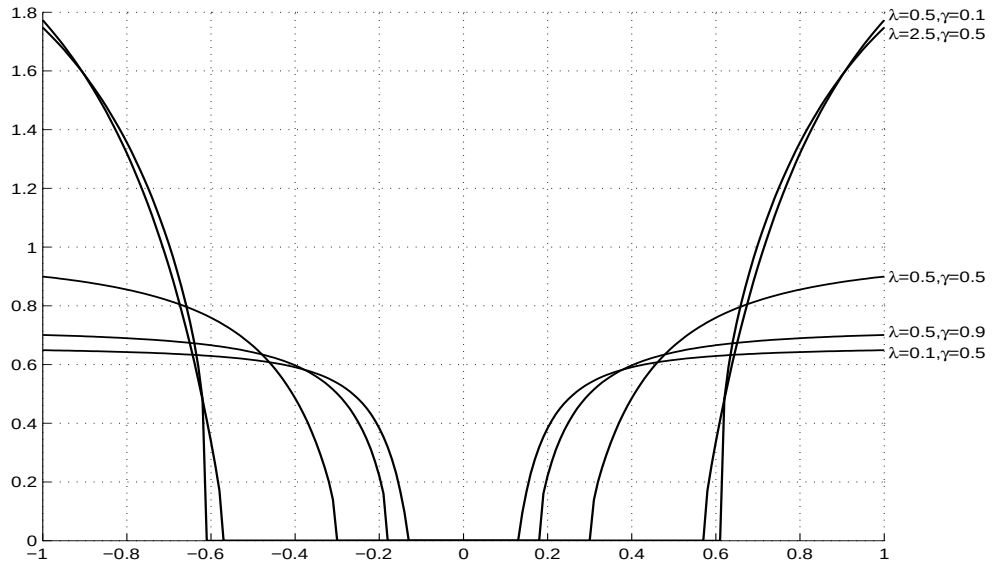

Figure 2: *Asymptotic optimal densities for the linear regression through the origin;* $\rho_\lambda(t) = e^{-\lambda|t|}$.

Long-range dependent error process (Dette, Leonenko, Pepelyshev & AZ (2008))

Correlation functions $(0 < \alpha < 1)$:

$$\rho_\alpha^{(1)}(t) = \frac{1}{(1+|t|^2)^{\alpha/2}}, \quad \rho_\alpha^{(2)}(t) = \frac{1}{1+|t|^\alpha}, \quad \rho_\alpha^{(3)}(t) = \frac{1}{(1+|t|)^\alpha}.$$

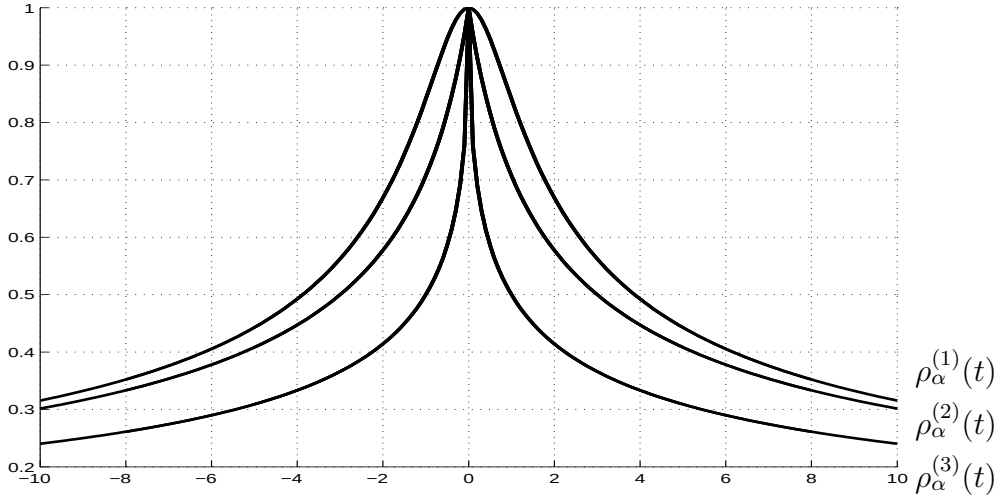

Figure 3: *The three correlation functions,* $\alpha = 0.5$.

Long-range dependent error process:

Asymptotically (as $N \to \infty$) the covariance matrix is proportional to $\mathbb{D}_\alpha(\xi) = M^{-1}(\xi) R_\alpha(\xi) M^{-1}(\xi)$,
where $R_\alpha(\xi) = \frac{1}{1-\alpha} \left(\int f(t) f^T(t) p^{1+\alpha}(t)\, dt \right)$, $\xi(dt) = p(t)dt$,

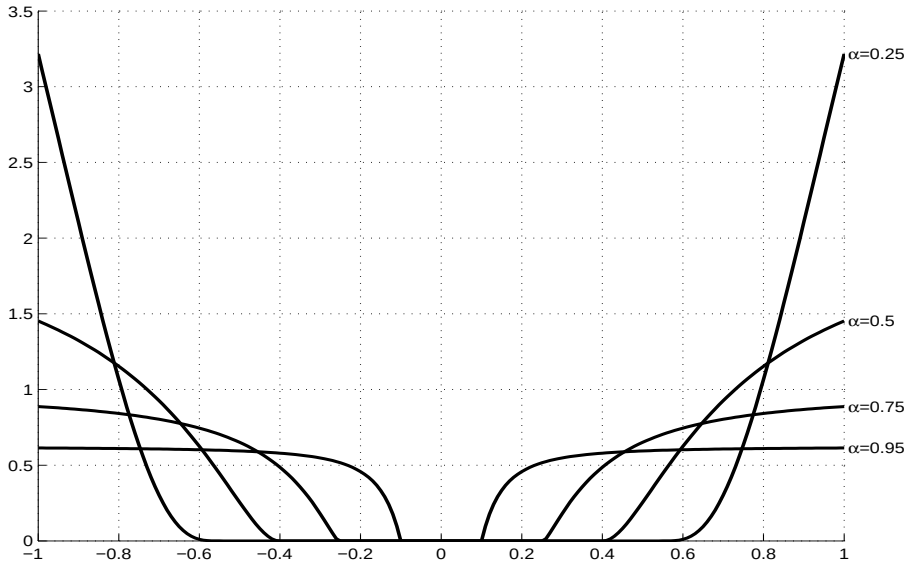

Figure 4: *Asymptotic optimal design densities for the linear regression through the origin.*

Alternative asymptotics (AZ, H.Dette and A.Pepelyshev (2010))

Consider

$$\mathbb{D}_N(\xi) = \sigma_N^2 \int\int f(x) f^T(y) \rho_N(x-y) \xi(dx) \xi(dy)\,.$$

As in the Bickel-Herzberg approach we assume

$$\rho_N(t) = \rho(Nt), \;\; \text{where, for example, } \; \rho(t) = \frac{1}{(1+|t|)^\alpha}\,.$$

In addition, we assume that the variance depends on N as well: $\sigma_N^2 = N^\alpha \sigma^2$.
Then the covariance function is

$$\sigma_N^2 \rho_N(t) = \sigma^2 N^\alpha \frac{1}{(1+|Nt|)^\alpha} = \sigma^2 \frac{1}{(1/N+|t|)^\alpha}$$

As $N \to \infty$, the sequence of matrices $\mathbb{D}_N(\xi)$ tends to

$$\mathbb{D}(\xi) = \sigma^2 \int\int f(x) f^T(y) r(x-y) \xi(dx) \xi(dy)$$

where $r(t)$ has singularity at 0; in this case $r(t) = 1/|t|^\alpha$.
The corresponding sequence of optimal designs converges too.

Covariance (matrix) of a design ξ:

$$y(t) = \theta + \varepsilon(t) \quad \Rightarrow \quad \mathbb{D}(\xi) = \int\int K(u,v)\xi(du)\xi(dv)$$

$$y(t) = \theta f(t) + \varepsilon(t) \quad \Rightarrow \quad \mathbb{D}(\xi) = \frac{\int\int f(u)f(v)K(u,v)\xi(du)\xi(dv)}{\left(\int f^2(u)\xi(du)\right)^2}$$

$$y(t) = \theta^T f(t) + \varepsilon(t) \quad \Rightarrow \quad \mathbb{D}(\xi) = M^{-1}(\xi)B(\xi,\xi)M^{-1}(\xi),$$

where

$$M(\xi) = \int f(u)f^T(u)\xi(\mathrm{d}u), \quad B(\xi,\xi) = \int\int K(u,v)f(u)f^T(v)\xi(\mathrm{d}u)\xi(\mathrm{d}v).$$

and $K(u,v)$ is either $\rho(u-v)$ or $r(u-v)$.

General optimality criteria: $\Phi(\mathbb{D}(\xi)) \to \min\limits_{\xi}$

Lemma 1. Let ξ and ν be two designs and Φ be a differentiable functional. Set $\xi_\alpha = (1-\alpha)\xi + \alpha\nu$ and assume that the matrices $M(\xi_0)$ and $B(\xi_0, \xi_0)$ are nonsingular. Then the directional derivative of Φ at the design ξ in the direction of $\nu - \xi$ is

$$\left.\frac{\partial \Phi(\mathbb{D}(\xi_\alpha))}{\partial \alpha}\right|_{\alpha=0} = 2[\mathbf{b}(\nu, \xi) - \boldsymbol{\varphi}(\nu, \xi)]$$

where

$$\boldsymbol{\varphi}(\nu, \xi) = \operatorname{tr} M(\nu) M^{-1}(\xi) B(\xi, \xi) M^{-1}(\xi)\, C(\xi) M^{-1}(\xi),$$

$$\mathbf{b}(\nu, \xi) = \operatorname{tr} M^{-1}(\xi)\, C(\xi) M^{-1}(\xi) B(\xi, \nu)$$

$$B(\xi, \nu) = \int\int K(u, v) f(u) f^T(v) \xi(\mathrm{d}u) \nu(\mathrm{d}v), \quad C = \frac{\partial \Phi(D)}{\partial D} = \left(\frac{\partial \Phi(D)}{\partial D_{ij}}\right)_{i,j=1,\ldots,m}$$

Proof. First:

$$\frac{\partial}{\partial\alpha}M^{-1}(\xi_\alpha)\Big|_{\alpha=0} = -M^{-1}(\xi_\alpha)\left(\frac{\partial}{\partial\alpha}M(\xi_\alpha)\right)M^{-1}(\xi_\alpha)\Big|_{\alpha=0} = M^{-1}(\xi) - M^{-1}(\xi)M(\nu)M^{-1}$$

Second:

$$\begin{aligned}
&\frac{\partial}{\partial\alpha}B(\xi_\alpha,\xi_\alpha)\Big|_{\alpha=0} = \\
&= \frac{\partial}{\partial\alpha}\Big((1-\alpha)^2B(\xi,\xi)+\alpha^2B(\nu,\nu)+\alpha(1-\alpha)\big(B(\xi,\nu)+B(\nu,\xi)\big)\Big)\Big|_{\alpha=0} \\
&= B(\xi,\nu)+B(\nu,\xi)-2B(\xi,\xi).
\end{aligned}$$

This implies

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\mathbb{D}(\xi_\alpha)\Big|_{\alpha=0} = &-M^{-1}(\xi)M(\nu)M^{-1}(\xi)B(\xi,\xi)M^{-1}(\xi) - M^{-1}(\xi)B(\xi,\xi)M^{-1}(\xi)M(\nu)M^{-} \\
&+M^{-1}(\xi)\Big(B(\xi,\nu)+B(\nu,\xi)\Big)M^{-1}(\xi) = 2[\mathbf{b}(\nu,\xi)-\boldsymbol{\varphi}(\nu,\xi)]\,.
\end{aligned}$$

The functions $\mathbf{b}(\nu, \xi)$ and $\boldsymbol{\varphi}(\nu, \xi)$ can be represented as

$$\mathbf{b}(\nu, \xi) = \int b(x, \xi)\nu(\mathrm{d}x), \quad \boldsymbol{\varphi}(\nu, \xi) = \int \varphi(x, \xi)\nu(\mathrm{d}x)$$

where and ξ_x is the probability measure concentrated at a point x,

$$\varphi(x, \xi) = \boldsymbol{\varphi}(\xi_x, \xi) = f^T(x)M^{-1}(\xi)B(\xi, \xi)M^{-1}(\xi)\, C(\xi)M^{-1}(\xi)f(x)\,,$$

Lemma 2. For any design ξ such that the matrices $M(\xi)$ and $B(\xi, \xi)$ are nonsingular we have

$$\int \varphi(x, \xi)\xi(\mathrm{d}x) = \int b(x, \xi)\xi(\mathrm{d}x) = \mathrm{tr}\, M^{-1}(\xi)B(\xi, \xi)M^{-1}(\xi)\, C(\xi)$$

Optimality theorem

Theorem. Let ξ^* be any design minimizing the functional $\Phi(\mathbb{D}(\xi))$. Then the inequality

$$\varphi(x, \xi^*) \leq b(x, \xi^*)$$

holds for all $x \in \mathbb{T}$. Moreover, this is an equality for ξ^*-almost all x.

Remark. In at least one paper, Torsney (1986), an optimality theorem for non-convex optimality criteria was used (non-correlated observations).

D-optimality $\Phi(\mathbb{D}(\xi)) = -\ln\det(\mathbb{D}(\xi))$

The analogue of the celebrated 'Equivalence Theorem' of Kiefer and Wolfowitz:

Theorem. Let ξ^* be any D-optimal design. Then for all $x \in \mathbb{T}$ we have

$$d(x, \xi^*) \leq b(x, \xi^*)$$

where the functions d and b are defined by

$$d(x, \xi) = f^T(x) M^{-1}(\xi) f(x),$$

$$b(x, \xi) = f^T(x) B^{-1}(\xi, \xi) \int K(u, x) f(u) \xi(\mathrm{d}u),$$

respectively. Moreover, this is an for ξ^*-almost all x.

Corollary. For any design ξ such that the matrices $M(\xi)$ and $B(\xi, \xi)$ are nonsingular we have

$$\int d(x, \xi) \xi(\mathrm{d}x) = \int b(x, \xi) \xi(\mathrm{d}x) = m\,.$$

Kernels with singularity at 0,

$y(t) = \theta + \varepsilon(t)$, $\mathbb{D}(\xi) = \int\int r(x-y)\xi(dx)\xi(dy)$

Is the function $r(t)$ a positive definite?

Logarithmic potential, arcsine density:

$$r(u) = -\log|u| \;\Rightarrow\; \mathbb{D}(\xi) = -\int\int \log|x-y|\xi(dx)\xi(dy)$$

Optimality condition:

$$\int \log|x-t|\xi(dx) = \text{const} \;\; \forall t \in \mathbb{T} \;\Rightarrow\; p^*(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \;\; \mathbb{T} = [0,1]$$

Beta-distribution:

$$r(u) = \frac{1}{|u|^\alpha} \;\Rightarrow\; p^*(x) = \frac{1}{B(\gamma,\gamma)} x^{\gamma-1}(1-x)^{\gamma-1}, \;\; \gamma = \frac{1+\alpha}{2}$$

$(0 < \alpha < 1)$.

Beta-density satisfies the condition: AZ,H.Dette, A.Pepelyshev (2010).

Uniqueness: K.-M.Schmidt and AZ (2011)

The same but in one formula:

Define the function

$$h_\alpha(t) = \begin{cases} (1-|t|^{\alpha-1})/(\alpha-1) & \text{if } \alpha \neq 1\,, \\ -\log|t| & \text{if } \alpha = 1\,. \end{cases}$$

Note $h_\alpha(t) > 0$ for $t \in (-1,0) \cup (0,1)$ and $\alpha \in (0,2)$.

Theorem. *Let ξ be a r.v. supported on $[0,1]$. This r.v. has the generalized arcsine density*

$$p_\gamma(t) = \frac{t^{\gamma-1}(1-t)^{\gamma-1}}{B(\gamma,\gamma)}\,, \quad 0 < t < 1\,,$$

if and only if the expectation $\mathbb{E}h_{2-2\gamma}(\xi - x)$ has the same value for a.a. $x \in [0,1]$.

Kernels with singularity at 0, model $y(t) = \theta f(t) + \varepsilon(t)$

$$\mathbb{D}(\xi) = \int\int r(x-y) f(x) f(y) \xi(dx) \xi(dy) / \left(\int f^2(x) \xi(dx) \right)^2$$

In general, this criterion is not convex.

Optimality condition: there exists $\lambda > 0$ such that

$$\lambda f(x) = \int r(u-x) f(u) \xi(\mathrm{d}u) \quad \forall\, x$$

Arcsine density: If $r(u) = -\log|u|$ and $f(x)$ is any Chebyshev polynomial of the first kind then the optimal design has the arcsine density.

Beta-density: If $r(u) = -|u|^\alpha$ and $f(x)$ is a Gegenbauer polynomial then the optimal design has the beta-density.

Kernels with singularity at 0, model $y(t) = \theta^T f(t) + \varepsilon(t)$

If $f(x) = (1, x, \ldots, x^{m-1})^T$ and $r(u) = -\log|u|$ then the arcsine design is D-optimal.

Example: Quadratic regression $y(x) = \theta_1 + \theta_2 x + \theta_3 x^2 + \varepsilon(x)$ $x \in [-1, 1]$.

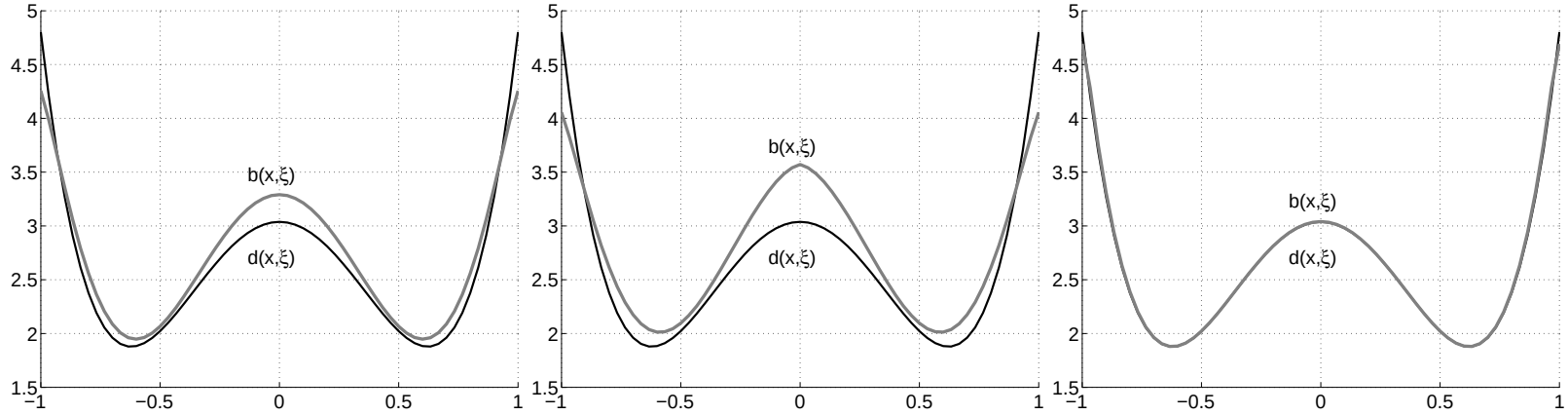

Figure 5: The functions $b(x, \xi)$ and $d(x, \xi)$ for the quadratic regression model and the covariance kernels $K(u, v) = e^{-|u-v|}$ (left), $K(u, v) = \max(0, 1 - |u - v|)$ (middle) and $K(u, v) = -\log(u - v)^2$ (right), and the arcsine design ξ.

Conclusions

1. A reliable but useless asymptotic theory is built
2. Nice mathematical problems remain

Thank you for attention