Lecture #7 Notes Summary

Kiefer-Wolfowitz Equivalence Theorem, Duality

Equivalence theorem for $D-$optimality

A special case of $\boldsymbol{c}-$optimality is when the experimenter wants to estimate a quantity $\zeta = \boldsymbol{a}(\boldsymbol{x})^T\boldsymbol{\theta}$ which could be observed by a single trial (here, the trial at $\boldsymbol{x} \in \mathcal{X}$ with regression vector $\boldsymbol{a}(\boldsymbol{x})$). In this case, the variance of the best estimator is $\sigma^2\boldsymbol{a}(\boldsymbol{x})^T M(\xi)^-\boldsymbol{a}(\boldsymbol{x})$. If $\boldsymbol{a}(\boldsymbol{x})$ is a vertex of the Elfving set E, this case is highly trivial (assign all the weight of the design to $\boldsymbol{x}$). However, an interesting case occurs when the experimenter is not interested in the observation of a single experiment $\boldsymbol{a}(\boldsymbol{x})^T\boldsymbol{\theta}$, but in the whole *regression surface* $\{\boldsymbol{a}(\boldsymbol{x})^T\boldsymbol{\theta},\ \boldsymbol{x} \in \mathcal{X}\}$. For example, recall the line fit model $y(x) = ax + b + \epsilon$, with $\boldsymbol{\theta} = [a, b]^T$. The experimenter might be interested to estimate the whole regression segment $\{ax + b, x \in \mathcal{X}\}$. A global criterion is needed to measure the performance of a design in this case. The *global criterion* (known as $G-$criterion) is

$$\Phi_G : M \to \max_{\boldsymbol{x}\in\mathcal{X}}\ \boldsymbol{a}(\boldsymbol{x})^T M^- \boldsymbol{a}(\boldsymbol{x})$$

and the $G-$optimal design guards one against the worst case, by minimizing the variance of every observation in the regression surface:

$$\begin{aligned} \min_{\xi} \quad & \max_{\boldsymbol{x}\in\mathcal{X}}\ \boldsymbol{a}(\boldsymbol{x})^T M(\xi)^- \boldsymbol{a}(\boldsymbol{x}) \\ \text{s.t.} \quad & M(\xi) = \sum_{i=1}^s w_i \boldsymbol{a}(\boldsymbol{x_i})\boldsymbol{a}(\boldsymbol{x_i})^T \\ & \sum_{i=1}^s w_i = 1, \quad \forall\ i \in [s], w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{1}$$

The Kiefer-Wolfowitz theorem establishes the equivalence between the $D-$ and the $G-$optimal designs:

Theorem 1 (Kiefer-Wolfowitz)**.** *Assume that the regression range $\{\boldsymbol{a}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ contains m linearly independent vectors. Then the following statements are equivalent:*

(i) The design ξ is $G-$optimal;

(ii) The design ξ is $D-$optimal for the full parameter $\boldsymbol{\theta}$ (i.e. with $K = I_m$);

(iii) For all $\boldsymbol{x}$ in $\mathcal{X}$, $\boldsymbol{a}(\boldsymbol{x})^T M(\xi)^- \boldsymbol{a}(\boldsymbol{x}) \leq m$.

Moreover, the bound provided by the inequality in (iii) is attained for the support points of the optimal design:

$$\boldsymbol{x_i} \in \operatorname{supp}(\xi) \Longrightarrow \boldsymbol{a}(\boldsymbol{x_i})^T M(\xi)^- \boldsymbol{a}(\boldsymbol{x_i}) = m.$$

Proof. We first show that for all design $\xi = \{\boldsymbol{x_k}, w_k\}$, we have $\Phi_G(\xi) \geq m$. If $M(\xi)$ is singular, then by assumption there is a regression vector $\boldsymbol{a}(\boldsymbol{x})$ which is not in the range of $M(\xi)$, and so $\Phi_G(\xi) = \infty \geq m$. If $M(\xi)$ is nonsingular, we write:

$$\begin{aligned} m = \text{trace}\ I_m = \text{trace}\ M(\xi)M(\xi)^{-1} &= \text{trace}\left(\sum_{i=1}^s w_i \boldsymbol{a}(\boldsymbol{x_i})\boldsymbol{a}(\boldsymbol{x_i})^T M(\xi)^{-1}\right) \\ &\leq \sum_{i=1}^s w_i\ \max_{\boldsymbol{x}\in\mathcal{X}} (\boldsymbol{a_x}^T M(\xi)^{-1} \boldsymbol{a_x}) \\ &= \Phi_G(\xi). \end{aligned}$$

This proves the part $(iii) \implies (i)$.

For the part $(ii) \implies (iii)$ we need this lemma:

Lemma 2. *Let $M \succ 0$. The directional derivative of* $\log\det$ *at M in the direction of $H \in \mathbb{S}^m$ is*

$$D\log\det(M)[H] := \lim_{\varepsilon \to 0^+} \frac{\log\det(M+\varepsilon H) - \log\det(M)}{\varepsilon} = \operatorname{trace}(M^{-1}H)$$

Now, we consider a $D-$optimal design ξ_D, and we show that $\boldsymbol{a_x}^T M(\xi_D)^- \boldsymbol{a_x} \leq m$ for every point $\boldsymbol{x} \in \mathcal{X}$, with equality when $\boldsymbol{x}$ is in the support of ξ_D. Note that a $D-$optimal design exists indeed, since we are maximizing a continuous function over a compact set. Moreover $M(\xi_D) \succ 0$. (otherwise $\det M(\xi_D) = 0$, and by assumption there is a nonsingular design, so at optimality the determinant must be > 0). $M(\xi_D)$ has the largest possible determinant, so $D\log\det(M(\xi_D))\Big[\boldsymbol{a}(\boldsymbol{x})\boldsymbol{a}(\boldsymbol{x})^T - M(\xi_D)\Big]$ must be ≤ 0; otherwise, there would exist a small $\varepsilon > 0$ such that $\log\det\Big((1-\varepsilon)M(\xi_D) + \varepsilon a(\boldsymbol{x})a(\boldsymbol{x})\Big) > \log\det\Big(M(\xi_D)\Big)$. So:

$$0 \geq D\log\det(M(\xi_D))\Big[\boldsymbol{a}(\boldsymbol{x})\boldsymbol{a}(\boldsymbol{x})^T - M(\xi_D)\Big] = \operatorname{trace} M(\xi_D)^{-1}(\boldsymbol{a}(\boldsymbol{x})\boldsymbol{a}(\boldsymbol{x})^T - M(\xi_D)) = \boldsymbol{a}(\boldsymbol{x})^T M(\xi_D)^- \boldsymbol{a}(\boldsymbol{x}) - m.$$

We further show that the latter inequality becomes an equality if $\boldsymbol{x}$ is a support point of ξ_D. We denote by $(\boldsymbol{x_i})_{i\in[s]}$ the support points of ξ_D and by $\boldsymbol{w}$ the vector of the associated weights, and we write:

$$m = \operatorname{trace}\ I_m = \operatorname{trace}\ M(\xi_D)M(\xi_D)^{-1} = \operatorname{trace}(\sum_{i=1}^{s} w_i \boldsymbol{a(x_i)}\boldsymbol{a(x_i)}^T M(\xi_D)^{-1}) = \sum_{i|w_i>0} w_i \boldsymbol{a(x_i)}^T M(\xi_D)^- \boldsymbol{a(x_i)}.$$

The latter expression is a weighted average of terms all smaller than m and takes the value m. Hence, $w_i > 0 \Rightarrow \boldsymbol{a(x_i)}^T M(\xi_D)^- \boldsymbol{a(x_i)} = m$.

Assume conversely that ξ is not $D-$optimal. If $M(\xi)$ is singular, then there is a regression vector $\boldsymbol{a}(\boldsymbol{x})$ which is not in the range of $M(\xi)$, and so (iii) does not hold. If $M(\xi)$ has full rank, then in view of the strict concavity of the $\log\det$ function over $\mathbb{S}^m_{++}$, and similarly to the previous discussion, there exists a design ξ' such that $\log\det(M(\xi))$ has a positive derivative in the direction of $M(\xi') - M(\xi)$:

$$\operatorname{trace} M(\xi)^{-1}(M(\xi') - M(\xi)) = \operatorname{trace} M(\xi)^{-1}M(\xi') - m > 0.$$

Denoting the support points and the weights of ξ' by $\boldsymbol{x_i}'$ and w'_i respectively, we obtain:

$$\operatorname{trace} M(\xi)^{-1}M(\xi') = \sum_{i|w'_i>0} w'_i \boldsymbol{a_{x'_i}}^T M(\xi)^- \boldsymbol{a_{x'_i}} > m.$$

This expression is a weighted average strictly larger than m, which implies the existence of a support point $\boldsymbol{x'}$ of ξ' such that $\boldsymbol{a_{x'_i}}^T M(\xi)^- \boldsymbol{a_{x'_i}} > m$. Hence, (iii) does not hold and we have proved the part $(iii) \implies (ii)$.

The existence of a $D-$optimal design, for which the Φ_G-criterion takes the value m, in conjunction with the fact that $\Phi_G(\xi) \geq m$ for all design ξ shows that a design ξ is $G-$optimal if and only if $\Phi_G(\xi) = m$. This proves the part $(i) \implies (iii)$ and the proof is complete. □

Duality

Definition 1 (Scalar product over $\mathbb{S}^m$)**.** The scalar product of two symmetrix matrices $A, B \in \mathbb{S}^m$ is

$$\langle A, B\rangle := \operatorname{trace} B^T A = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}b_{ij}.$$

Definition 2 (Polar information function)**.** Let $\Phi : \mathbb{S}^m \to \mathbb{R}$ be an information function. We define the polar conjugate of Φ as

$$\Phi^\star(D) := \inf_{C \succ 0} \frac{\langle C, D\rangle}{\Phi(C)}.$$

Proposition 3 (Polar of Kiefer's Φ_p-criterion)**.** *Let p and q be conjugate numbers on $[-\infty, 1]$, i.e. $p+q = pq$, or equivalently $\frac{1}{p} + \frac{1}{q} = 1$. The polar function of Kiefer's Φ_p-criterion (over $\mathbb{S}^m$) is*

$$\Phi_p^\star := m\Phi_q$$

Theorem 4 (Duality)**.** *Let $\Phi : \mathbb{S}^r \to \mathbb{R}$ be an information function and K be an $m \times r$ matrix of full column rank. Then,*

$$\max_{\xi \in \Xi(K)} \Phi\Big(M_K(\xi)\Big) = \begin{array}{ll} \min\limits_{N \succeq 0} & \dfrac{1}{\Phi^\star(K^T N K)} \\ s.t. & \forall \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{a}(\boldsymbol{x})^T N \boldsymbol{a}(\boldsymbol{x}) \leq 1. \end{array}$$

Moreover, for the optimal dual variable $N \succeq 0$ it holds that $\boldsymbol{x} \in \operatorname{supp}(\xi) \Longrightarrow \boldsymbol{a}(\boldsymbol{x})^T N \boldsymbol{a}(\boldsymbol{x}) = 1$.

Proof. We only proof the weak duality inequality ($\leq$). Let $\xi \in \Xi(K)$ be a feasible design, set $M := M(\xi)$, $M_K := M_K(\xi)$, and let $N \in \mathbb{S}_+^m$ be a feasible matrix for the dual problem. The weak duality is a consequence of the following three inequalities, which in fact become equalities for the optimal M and N:

(i) $1 \geq \langle M, N\rangle$

(ii) $\langle M, N\rangle \geq \langle M_K, K^T N K\rangle$

(iii) $\langle M_K, K^T N K\rangle \geq \Phi(M_K)\Phi^\star(K^T N K)$

The point (i) simply follows from the fact that $\boldsymbol{a}(\boldsymbol{x})^T N \boldsymbol{a}(\boldsymbol{x}) \leq 1$ for all design points $x \in \mathcal{X}$ (because N is feasible for the dual problem.) The point (iii) comes from the definition of the polar function $\Phi^\star$.

Now, consider a decomposition $M = A^T A$ for a $m \times m-$matrix A, and recall the Gauss-Markov theorem $KM^- K = K^T (A^T A)^- K = \min_{\preceq}\{H^T H : H \in R^{m \times r} A^T H = K\}$. Let H_0 be a minimizer of this problem. We have $A^T H_0 = K$ and $H_0^T H_0 = K^T M^- K$, so that

$$0 \preceq \begin{pmatrix} A^T \\ H_0^T \end{pmatrix} (A\ H_0) = \begin{pmatrix} M & K \\ K^T & K^T M^- K \end{pmatrix}.$$

The Schur complement lemma yields $M \succeq K(K^T M^- K)^{-1} K^T = K M_K K^T$. Now, we use the following

Lemma 5. *Let $U \succeq 0$. Then, $X \succeq Y \Longrightarrow \langle X, U\rangle \geq \langle Y, U\rangle$.*

This gives $\langle M, N\rangle \geq \langle K M_K K^T, N\rangle = \operatorname{trace}(K M_K K^T N) = \operatorname{trace}(M_K K^T N K) = \langle M_K, K^T N K\rangle$. □

Exercises

1. Prove Lemma 5

2. Let $\xi = \{\boldsymbol{x}, \boldsymbol{w}\}$ be a $D-$optimal design (with a support of size s, for the whole parameter $\boldsymbol{\theta} \in \mathbb{R}^m$).

   The goal of this exercise is to show that $w_i \leq \frac{1}{m}$ for all $i = 1, \ldots, s$. To simplify the notation, we write $\boldsymbol{a}_i$ instead of $\boldsymbol{a}(\boldsymbol{x}_i)$. Now, let i be an arbitrary index in $\{1, \ldots, s\}$.

   - What does the Kiefer-Wolfowitz theorem tell you about the quantity $\boldsymbol{a}_i^T M(\xi)^- \boldsymbol{a}_i$.
   - Show that $M(\xi)$ is invertible and conclude that $\boldsymbol{a}_i^T M(\xi)^- \boldsymbol{a}_i = \boldsymbol{a}_i^T M(\xi)^{-1} M(\xi) M(\xi)^{-1} \boldsymbol{a}_i$.
   - Rewrite $\boldsymbol{a}_i^T M(\xi)^- \boldsymbol{a}_i$ as a convex combination of the $(\boldsymbol{a}_i^T M(\xi)^- \boldsymbol{a}_k)^2$ $(k = 1, \ldots, s)$.
   - Conclude

3. Consider the polynomial regression model of degree d on $\mathcal{X} = [-1, 1]$:

$$\forall \boldsymbol{x} \in \mathcal{X}, \boldsymbol{a}(\boldsymbol{x}) = [1, x, x^2, \ldots, x^d]^T \in \mathbb{R}^{d+1}.$$

   - Show that if an information matrix $M(\xi) = \sum_{i=1}^{s} w_i \boldsymbol{a}(x_i) \boldsymbol{a}(x_i)^T$ is non singular, the design ξ must have at least $s = d + 1$ support points.
   - Let ξ be a $D-$optimal design. Show that there exists a matrix $N \succ 0$ such that $\boldsymbol{a}(x)^T N \boldsymbol{a}(x) = 1$ for all support points x of ξ.
   - What can you say about function $x \to \boldsymbol{a}(x)^T N \boldsymbol{a}(x)$ over $\mathcal{X} = [-1, 1]$? Conclude that ξ has exactly $d + 1$ support points $-1 = x_0 < \ldots < x_d = 1$.
   - Show moreover that $w_i = \frac{1}{d+1}$ for all $i = 0, \ldots, d$ (use Exercise 2).
   - By using a simple symmetry argument, find the $D-$optimal design for the quadratic fit model $(d = 2)$.