Notes on: Robustness of Design for Clinical Trials

Douglas P. Wiens[1]

July 2, 2019

1 An approximate model for clinical trials

Suppose that subjects arrive for treatment. With each is associated a d-dimensional vector $\boldsymbol{x}$ of prognostic factors, which may be included in the ensuing analysis via a q_1-dimensional vector $\boldsymbol{f}(\boldsymbol{x})$ of regressors (possibly) interacting with treatments and a q_2-dimensional vector $\boldsymbol{g}(\boldsymbol{x})$ of regressors whose effect is common to all treatments. Each is assigned, upon arrival, to one of p treatment groups (perhaps a control); possible purposes of the study are to (i) estimate the treatment effects, (ii) perhaps to predict these effects, given a particular treatment and covariate vector, (iii) perhaps to, over time, assign patients to the best treatment. In this current document the assignments initially depend (at most) only on the current value of $\boldsymbol{x}$ and not on the outcomes, and so are mutually independent. Then we consider 'response adaptive' methods.

An element of randomness is required in order that the investigator remain 'blinded' to the subject/treatment pairings; thus we derive functions $\{\rho_i(\cdot)\}_{i=1}^{p}$ and suppose that an assignment will be made to group i with probability $\rho_i(\boldsymbol{x})$.

Possible applications:

1. Clinical trials – determine the better treatment, conditional on the covariates.

2. Treatment regimes – monitor the outcomes sequentially, with an eye to determining the best treatment for a patient with covariates $\boldsymbol{x}$.

3. Response surface exploration, computer experimentation – this is speculative, but suppose one seeks the location $\boldsymbol{x}$ of the minimum of a response surface with multiple local minima. There could be two 'treatment' groups, corresponding to accepting a value of $\boldsymbol{x}$ into the study or not. A value of $\boldsymbol{x}$ resulting in a smaller loss would be accepted with accepted with probability $\rho_i(\boldsymbol{x})$. The idea is that even an unfavourable value of $\boldsymbol{x}$ should sometimes be accepted, in order that one not be trapped in a local minimum.

Assume that $\boldsymbol{x}$ has a density $m(\boldsymbol{x})$ with respect to a measure $\mu(d\boldsymbol{x})$ on a region $\chi \subset \mathbb{R}^d$. Typically some factors will be discrete and others continuous, and then μ will be a product $\mu_1 \times \mu_2$ of d_1-dimensional counting measure and $(d-d_1)$-dimensional Lebesgue measure. Then for functions $\phi(\boldsymbol{x})$,

$$E[\phi(\boldsymbol{x})] = \int_\chi \phi(\boldsymbol{x})\, m(\boldsymbol{x})\, \mu(d\boldsymbol{x}).$$

[1]Department of Mathematical and Statistical Sciences; University of Alberta, Edmonton, Alberta; Canada T6G 2G1. e-mail: doug.wiens@ualberta.ca

The experimenter's (approximate) model is that, given $\boldsymbol{x}$, the (possibly transformed, as in the case of a GLM) response of a subject to treatment $i \in \{1, 2, ..., p\}$ is

$$Y|i, \boldsymbol{x} \approx \alpha_i + \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta}_i + \boldsymbol{g}'(\boldsymbol{x})\boldsymbol{\gamma} + \sigma_i(\boldsymbol{x})\varepsilon, \tag{1}$$

for i.i.d. errors ε with mean zero and variance one. This allows for the elements of $\boldsymbol{f}(\boldsymbol{x}) : q_1 \times 1$ to interact with the treatments and those of $\boldsymbol{g}(\boldsymbol{x}) : q_2 \times 1$ to have an effect common to all treatments. In order that the treatment effects α_i be identifiable it is assumed that the mean of $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta}_i + \boldsymbol{g}'(\boldsymbol{x})\boldsymbol{\gamma}$ has been absorbed into them, i.e. that

$$E\left[\begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}) \\ \boldsymbol{g}(\boldsymbol{x}) \end{pmatrix}\right] = \mathbf{0}_{q\times 1} \ (q \stackrel{def}{=} q_1 + q_2). \tag{2}$$

The regressors may be assumed to have been standardized as well as centred. Set

$$\boldsymbol{P}_{ff} = E[\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}'(\boldsymbol{x})] : q_1 \times q_1, \tag{3a}$$
$$\boldsymbol{P}_{gg} = E[\boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}'(\boldsymbol{x})] : q_2 \times q_2, \tag{3b}$$
$$\boldsymbol{P}_{fg} = E[\boldsymbol{f}(\boldsymbol{x})\boldsymbol{g}'(\boldsymbol{x})] : q_1 \times q_2. \tag{3c}$$

We assume that $\boldsymbol{P}_{ff}$ and $\boldsymbol{P}_{gg}$ are nonsingular, in which case we may also assume that $\boldsymbol{f}$ and $\boldsymbol{g}$ have been normalized in such a way that

$$\boldsymbol{P}_{ff} = \boldsymbol{I}_{q_1}, \boldsymbol{P}_{gg} = \frac{1}{p}\boldsymbol{I}_{q_2}, \boldsymbol{P}_{fg} = \mathbf{0}_{q_1\times q_2}. \tag{4}$$

[**How? – First ensure that $\boldsymbol{P}_{ff} = I_{q_1}, \boldsymbol{P}_{gg} = \frac{1}{p}I_{q_2}$ by pre-multiplying $\boldsymbol{f}$ and g by $\boldsymbol{P}_{ff}^{-1/2}$ and $(p\boldsymbol{P}_{gg})^{-1/2}$ respectively. Then replace $\boldsymbol{g}(\boldsymbol{x})$ by $\tilde{\boldsymbol{g}}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{P}'_{fg}\boldsymbol{f}(\boldsymbol{x})$ and β_i by $\tilde{\beta}_i = \beta_i + P_{fg}\gamma$ in (1), which then becomes $Y|i, x \approx \alpha_i + \boldsymbol{f}'(\boldsymbol{x})\tilde{\beta}_i + \tilde{\boldsymbol{g}}'(\boldsymbol{x})\gamma$, with $\boldsymbol{P}_{f\tilde{g}} = 0$.**]

To formalize the approximate nature of (1), we define the parameters by

$$\left(\{\alpha_i, \boldsymbol{\beta}_i\}_{i=1}^p, \boldsymbol{\gamma}\right) = \arg\min_{\left(\{\mathring{\alpha}_i, \mathring{\boldsymbol{\beta}}_i\}_{i=1}^p, \mathring{\boldsymbol{\gamma}}\right)} \sum_{i=1}^p \int_\chi \left(E_\varepsilon[Y|i, \boldsymbol{x}] - \mathring{\alpha}_i - \boldsymbol{f}'(\boldsymbol{x})\mathring{\boldsymbol{\beta}}_i - \boldsymbol{g}'(\boldsymbol{x})\mathring{\boldsymbol{\gamma}}\right)^2 m(\boldsymbol{x})\mu(d\boldsymbol{x}).$$

Carrying out the minimization, and defining

$$\psi_{n,i}(\boldsymbol{x}) = E[Y|i, \boldsymbol{x}] - \alpha_i - \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta}_i - \boldsymbol{g}'(\boldsymbol{x})\boldsymbol{\gamma},$$

results in the constraints

$$E\left[\begin{pmatrix} 1 \\ \boldsymbol{f}(\boldsymbol{x}) \end{pmatrix}\psi_{n,i}(\boldsymbol{x})\right] = \mathbf{0}, \tag{5a}$$

$$E\left[\boldsymbol{g}(\boldsymbol{x})\left(\sum_{i=1}^p \psi_{n,i}(\boldsymbol{x})\right)\right] = \mathbf{0}. \tag{5b}$$

Thus the 'true' model is specified by

$$Y|i,\boldsymbol{x} = E\left[Y|i,\boldsymbol{x}\right] + \sigma_i\left(\boldsymbol{x}\right)\varepsilon,$$

with

$$E\left[Y|i,\boldsymbol{x}\right] = \alpha_i + \boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{\beta}_i + \boldsymbol{g}'\left(\boldsymbol{x}\right)\boldsymbol{\gamma} + \psi_{n,i}\left(\boldsymbol{x}\right) \text{ and } E\left[Y|i\right] = \alpha_i.$$

The dependence of the ψ's on n is for the asymptotics – in order that bias and variance decrease at the same rate we must impose bounds on the magnitudes of the $\psi_{n,i}$ – and we will in fact assume that there are limit functions

$$\psi_i\left(\cdot\right) = \lim_{n\to\infty}\sqrt{n}\psi_{n,i}\left(\cdot\right).$$

With $\boldsymbol{\sigma}\left(\boldsymbol{x}\right) = \left(\sigma_1\left(\boldsymbol{x}\right),\cdots,\sigma_p\left(\boldsymbol{x}\right)\right)'$ and $\boldsymbol{\sigma}^2\left(\boldsymbol{x}\right) = \left(\sigma_1^2\left(\boldsymbol{x}\right),\cdots,\sigma_p^2\left(\boldsymbol{x}\right)\right)'$ we impose a bound, for given σ_0^2, of either

$$E\left[\left\|\boldsymbol{\sigma}\left(\boldsymbol{x}\right)\right\|^2\right] \le \sigma_0^2, \tag{6}$$

or

$$\sqrt{E\left[\left\|\boldsymbol{\sigma}^2\left(\boldsymbol{x}\right)\right\|^2\right]} \le \sigma_0^2. \tag{7}$$

We use the notation $\boldsymbol{A}\oplus\boldsymbol{B}$ for the direct sum $diag\left(\boldsymbol{A},\boldsymbol{B}\right)$ of matrices, and $\boldsymbol{A}\otimes\boldsymbol{B}$ for the Kronecker product $\left(a_{ij}\boldsymbol{B}\right)$. With $s = p + pq_1 + q_2$ (= the number of regression parameters) we define a $p\times s$ matrix $\boldsymbol{R}$ by

$$\boldsymbol{R}\left(\boldsymbol{x}\right) = \left(\mathbf{I}_p \vdots \mathbf{I}_p\otimes\boldsymbol{f}'\left(\boldsymbol{x}\right) \vdots \mathbf{1}_p\boldsymbol{g}'\left(\boldsymbol{x}\right)\right).$$

The i^{th} row of $\boldsymbol{R}$ is

$$\boldsymbol{r}_i'\left(\boldsymbol{x}\right) = \left(\underbrace{0,\cdots 0,\overset{\overset{i}{\downarrow}}{1},0,\cdots,0}_{1\times p}\vdots\underbrace{\mathbf{0}'\cdots\mathbf{0}'\overset{\overset{i}{\downarrow}}{\boldsymbol{f}'}(\boldsymbol{x})\mathbf{0}'\cdots\mathbf{0}'}_{1\times pq_1}\vdots\underbrace{\boldsymbol{g}'\left(\boldsymbol{x}\right)}_{1\times q_2}\right), \tag{8}$$

Then with $\boldsymbol{\psi}_n\left(\boldsymbol{x}\right) = \left(\psi_{n,1}\left(\boldsymbol{x}\right),\cdots,\psi_{n,p}\left(\boldsymbol{x}\right)\right)'$ and $\boldsymbol{\psi}\left(\boldsymbol{x}\right) = \left(\psi_1\left(\boldsymbol{x}\right),\cdots,\psi_p\left(\boldsymbol{x}\right)\right)' = \lim_{n\to\infty}\sqrt{n}\boldsymbol{\psi}_n\left(\boldsymbol{x}\right)$, the constraints (5) become

$$E\left[\boldsymbol{R}'\left(\boldsymbol{x}\right)\boldsymbol{\psi}_n\left(\boldsymbol{x}\right)\right] = \mathbf{0}_{s\times 1} = E\left[\boldsymbol{R}'\left(\boldsymbol{x}\right)\boldsymbol{\psi}\left(\boldsymbol{x}\right)\right]. \tag{9}$$

We also impose a bound

$$E\left[\left\|\boldsymbol{\psi}\left(\boldsymbol{x}\right)\right\|^2\right] \le \eta^2. \tag{10}$$

The constraints (2) and (3) are now conveniently expressed as

$$E\left[\boldsymbol{R}'\left(\boldsymbol{x}\right)\boldsymbol{R}\left(\boldsymbol{x}\right)\right] = \boldsymbol{I}_s. \tag{11}$$

Example: $p = 2$ treatments, $d = 2$ covariates: $X_1 = \pm 1$ with probability 1/2 each and X_2 has density $\frac{3}{16}\left(1+5x_2^2\right)I\left(-1\le x_2\le 1\right)$ (symmetric with variance 1/2) independently of X_1. Put $\boldsymbol{x} = (X_1, X_2)'$, $f(\boldsymbol{x}) = X_1$, $g(\boldsymbol{x}) = X_2$. Interpretation: X_1 denotes gender and may interact with the treatments, X_2 is (transformed and standardized) blood pressure of the subject and has a common effect. We have

$$E[f(\boldsymbol{x})] = E[X_1] = 0, E[f(\boldsymbol{x}) f'(\boldsymbol{x})] = E\left[X_1^2\right] = 1$$

and

$$E[g(\boldsymbol{x})] = E[X_2] = 0, E[g(\boldsymbol{x}) g'(\boldsymbol{x})] = E\left[X_2^2\right] = 1/2,$$

to satisfy (3). Then

$$\boldsymbol{R}(\boldsymbol{x}) = \begin{pmatrix} 1 & 0 & X_1 & 0 & X_2 \\ 0 & 1 & 0 & X_1 & X_2 \end{pmatrix}.$$

2 Asymptotic MSE

Denote by $\{\boldsymbol{x}_{ij}\}_{j=1}^{n_i}$ and $\{y_{ij}\}_{j=1}^{n_i}$ the covariates and responses associated with the subjects assigned to treatment i, and define vectors

$$\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})' \text{ and } \boldsymbol{z}_i = (\psi_i(\boldsymbol{x}_{i1}), \cdots, \psi_i(\boldsymbol{x}_{in_i}))'.$$

Conditional on the σ-algebra $\mathcal{F}_n$ generated by the first $n = \sum n_i$ arrivals and their treatment assignments, we represent the complete data as

$$\boldsymbol{y}_{n\times 1} = \left(\boldsymbol{y}_1', \cdots, \boldsymbol{y}_p'\right)' = \boldsymbol{V\theta} + \boldsymbol{z} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\varepsilon}.$$

Here $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors arising with $\boldsymbol{y}$, $\boldsymbol{\Sigma}_{n\times n}$ is the diagonal matrix with diagonal elements $\sigma_i^2(\boldsymbol{x}_{ij})$, with these ordered in the same manner as the elements of $\boldsymbol{y}$,

$$\begin{aligned} \boldsymbol{\theta}_{s\times 1} &= \left(\alpha_1, \cdots, \alpha_p, \boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_p', \boldsymbol{\gamma}'\right)', \\ \boldsymbol{z}_{n\times 1} &= \left(\boldsymbol{z}_1', \cdots, \boldsymbol{z}_p'\right)', \end{aligned}$$

and $\boldsymbol{V}$ is an $n \times s$ matrix whose $(i,j)^{th}$ row (these are ordered in the same manner as the elements of $\boldsymbol{y}$) is – recall (8) – given by $\boldsymbol{v}_{ij}' = \boldsymbol{r}_i'(\boldsymbol{x}_{ij})$.

We suppose that the analyst will carry out weighted least squares (wls) estimation, with weights $w_i(\boldsymbol{x}) \in [0,1]$. Let $\boldsymbol{W}$ be the diagonal matrix with diagonal elements $w_i(\boldsymbol{x}_{ij})$, with these ordered in the same manner as the elements of $\boldsymbol{y}$. Then the wls estimate is

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V})^{-1}\boldsymbol{V}'\boldsymbol{W}\boldsymbol{y} = \boldsymbol{\theta} + (\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V})^{-1}\boldsymbol{V}'\boldsymbol{W}\boldsymbol{z} + (\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V})^{-1}\boldsymbol{V}'\boldsymbol{W}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\varepsilon}.$$

With $t_i(\boldsymbol{x}) \overset{def}{=} w_i(\boldsymbol{x})\rho_i(\boldsymbol{x})$, $\boldsymbol{A_t}(\boldsymbol{x}) \overset{def}{=} \oplus_{i=1}^p t_i(\boldsymbol{x})$ and $\boldsymbol{A}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}(\boldsymbol{x}) \overset{def}{=} \oplus_{i=1}^p \sigma_i^2(\boldsymbol{x}) w_i(\boldsymbol{x}) t_i(\boldsymbol{x})$, we define

$$\begin{aligned} \boldsymbol{M_t} &= E\left[\boldsymbol{R}'(\boldsymbol{x})\boldsymbol{A_t}(\boldsymbol{x})\boldsymbol{R}(\boldsymbol{x})\right], \\ \boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}} &= E\left[\boldsymbol{R}'(\boldsymbol{x})\boldsymbol{A}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}\boldsymbol{R}(\boldsymbol{x})\right], \\ \boldsymbol{q}_{\boldsymbol{t},\boldsymbol{\psi}} &= E\left[\boldsymbol{R}'(\boldsymbol{x})\boldsymbol{A_t}(\boldsymbol{x})\boldsymbol{\psi}(\boldsymbol{x})\right]. \end{aligned}$$

Example continued: Here we have

$$\boldsymbol{A_t}(\boldsymbol{x}) = \begin{pmatrix} t_1(\boldsymbol{x}) = w_1(\boldsymbol{x})\rho_1(\boldsymbol{x}) & 0 \\ 0 & t_2(\boldsymbol{x}) = w_2(\boldsymbol{x})\rho_2(\boldsymbol{x}) \end{pmatrix},$$

$$\boldsymbol{A}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}(\boldsymbol{x}) = \begin{pmatrix} \sigma_1^2(\boldsymbol{x})w_1(\boldsymbol{x})t_1(\boldsymbol{x}) & 0 \\ 0 & \sigma_2^2(\boldsymbol{x})w_2(\boldsymbol{x})t_2(\boldsymbol{x}) \end{pmatrix},$$

$$\boldsymbol{M_t} = E\left[\begin{pmatrix} \boldsymbol{A_t}(\boldsymbol{x}) & X_1\boldsymbol{A_t}(\boldsymbol{x}) & \begin{pmatrix} t_1(\boldsymbol{x}) \\ t_2(\boldsymbol{x}) \end{pmatrix} X_2 \\ * & X_1^2\boldsymbol{A_t}(\boldsymbol{x}) & \begin{pmatrix} t_1(\boldsymbol{x}) \\ t_2(\boldsymbol{x}) \end{pmatrix} X_1X_2 \\ * & * & X_2^2(t_1(\boldsymbol{x}) + t_2(\boldsymbol{x})) \end{pmatrix}\right],$$

$$\boldsymbol{q}_{\boldsymbol{t},\boldsymbol{\psi}} = E\left[\begin{pmatrix} \boldsymbol{A_t}(\boldsymbol{x}) \\ X_1\boldsymbol{A_t}(\boldsymbol{x}) \\ X_2\begin{pmatrix} t_1(\boldsymbol{x}) & t_2(\boldsymbol{x}) \end{pmatrix} \end{pmatrix}\boldsymbol{\psi}(\boldsymbol{x})\right],$$

and $\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}$ is like $\boldsymbol{M_t}$ but with all six occurrences of $t_i(\boldsymbol{x})$ replaced by $\sigma_i^2(\boldsymbol{x})w_i(\boldsymbol{x})t_i(\boldsymbol{x})$, $i = 1, 2$.

Theorem 1 *Assume that for $i = 1, ..., p$ the group sizes $n_i \to \infty$ in such a way that $n_i/n \to \rho_{0,i}$, for constants $\rho_{0,i} \in (0, 1)$, as the study size $n \to \infty$. Then the mean squared error matrix, conditional on $\mathcal{F}_n$, satisfies*

$$\text{MSE}\left[\sqrt{n}\hat{\boldsymbol{\theta}}\ |\mathcal{F}_n\right] \overset{a.s.}{\to} \boldsymbol{M}_{\boldsymbol{t}}^{-1}\left(\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}} + \boldsymbol{q}_{\boldsymbol{t},\boldsymbol{\psi}}\boldsymbol{q}'_{\boldsymbol{t},\boldsymbol{\psi}}\right)\boldsymbol{M}_{\boldsymbol{t}}^{-1}. \tag{12}$$

In the following, a special role is played by assignment probabilities and weights satisfying the condition that their product be constant:

$$\text{For } i = 1, ..., p,\ t_i(\boldsymbol{x}) \equiv t_i \text{ (necessarily} = E[w_i(\boldsymbol{x})\rho_i(\boldsymbol{x})]\text{), for all } \boldsymbol{x}. \tag{13}$$

Under (13), and using (2), (9) and (4), the components of the asymptotic MSE simplify to

$$\boldsymbol{M_t} = \oplus_{i=1}^p t_i \oplus \left(\oplus_{i=1}^p t_i \otimes \boldsymbol{I}_{q_1}\right) \oplus \left(\frac{1}{p}\sum_{i=1}^p t_i\right)\boldsymbol{I}_{q_2},$$

$$\mathbf{Q}_{11} = \oplus_{i=1}^p t_i E\left[w_i(\boldsymbol{x})\sigma_i^2(\boldsymbol{x})\right], \tag{14}$$

$$\boldsymbol{q} = \begin{pmatrix} \mathbf{0}_{p\times 1} \\ \mathbf{0}_{pq_1\times 1} \\ E\left[\left(\sum_{i=1}^p t_i\psi_i(\boldsymbol{x})\right)\boldsymbol{g}(\boldsymbol{x})\right] \end{pmatrix}, \tag{15}$$

where in (14) the subscript 11 refers to the upper-left $p \times p$ block. The other blocks of $\mathbf{Q}$ turn out not to be needed at this point.

[In the case that (13) holds since both $w_i(\boldsymbol{x})$ and $\rho_i(\boldsymbol{x})$ are independent of $\boldsymbol{x}$, and if as well the $\sigma_i^2(\boldsymbol{x})$ are independent of $\boldsymbol{x}$, then a special case uses efficient weights $w_i = \left(\sum_{i=1}^p \sigma_i^2\right)/\sigma_i^2$ ($\propto 1/\sigma_i^2$ and normalized by $\sum_{i=1}^p w_i^{-1} = 1$) and $\rho_{0,i} = w_i^{-1}$. This results in asymptotic unbiasedness ($\boldsymbol{q}_{\boldsymbol{t},\boldsymbol{\psi}} = \mathbf{0}_{s\times 1}$) and $\boldsymbol{M_t} = \boldsymbol{I}_s$, $\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}} = \left(\sum_{i=1}^p \sigma_i^2\right)\boldsymbol{I}_s$, whence $\text{MSE}\left[\sqrt{n}\hat{\boldsymbol{\theta}}\ |\mathcal{F}_n\right] \overset{a.s.}{\to} \left(\sum_{i=1}^p \sigma_i^2\right)\boldsymbol{I}_s$.]

3 Contrasts of treatment effects

Suppose that interest is on a complete set of orthogonal contrasts

$$\mathbf{\Pi}_{p-1\times p}\boldsymbol{\alpha} = \left(\mathbf{\Pi}\vdots\mathbf{0}_{p-1\times s-p}\right)\boldsymbol{\theta},$$

where the rows of $\mathbf{\Pi}$ are mutually orthogonal and sum to zero, i.e.

$$\left(\begin{array}{c}\mathbf{1}_p'/\sqrt{p}\\ \mathbf{\Pi}\end{array}\right)$$

is a $p\times p$ orthogonal matrix. Define loss

$$\mathcal{L}_1\left(\boldsymbol{\rho},\boldsymbol{w};\boldsymbol{\psi},\boldsymbol{\sigma}\right) = \lim_{n\to\infty}\det\left\{\mathrm{MSE}\left[\sqrt{n}\left(\left(\mathbf{\Pi}\vdots\mathbf{0}_{p-1\times s-p}\right)\hat{\boldsymbol{\theta}}\right)|\mathcal{F}_n\right]\right\}.$$

The following result extends Theorem 2 of Wiens (2005), where treatment/covariate interactions were not considered. (See also the technical report Wiens (2000).

Theorem 2 *Suppose that the variance functions $\sigma_i^2(\boldsymbol{x})$ are constant: $\sigma_i^2(\boldsymbol{x})\equiv\sigma_i^2$. Then:*
(i) For any such variance functions, any vector $\boldsymbol{\rho}(\cdot) = (\rho_1(\cdot),...,\rho_p(\cdot))'$ of assignment probabilities, any vector $\boldsymbol{w}(\cdot) = (w_1(\cdot),...,w_p(\cdot))'$ of weights and any $\boldsymbol{\psi}$, the loss $\mathcal{L}_1(\boldsymbol{\rho},\boldsymbol{w};\boldsymbol{\psi},\boldsymbol{\sigma})$ exceeds that using weights and assignment probabilities satisfying (13).
(ii) Define $w_i = E[w_i(\boldsymbol{x})]$. The loss $\mathcal{L}_1(\cdot)$ using weights and assignment probabilities satisfying (13) does not depend upon $\boldsymbol{\psi}$, and is

$$\mathcal{L}_1\left(\boldsymbol{t};\boldsymbol{\sigma}\right) = \frac{\sum_{i=1}^p\frac{t_i/w_i}{\sigma_i^2}}{p\prod_{i=1}^p\frac{t_i/w_i}{\sigma_i^2}}. \tag{16}$$

Note: At this point we should try to maximize over (6) or (7), then find minimizing ratios t_i/w_i.

Corollary 1 *Suppose that for $i=1,...,p$ both $w_i(\boldsymbol{x})\equiv w_i$ and $\rho_i(\boldsymbol{x})\equiv\rho_i$ are independent of $\boldsymbol{x}$. Then the loss (16) becomes*

$$\mathcal{L}_1\left(\boldsymbol{\rho},\boldsymbol{\sigma}\right) = \frac{\sum_{i=1}^p\frac{\rho_i}{\sigma_i^2}}{p\prod_{i=1}^p\frac{\rho_i}{\sigma_i^2}}. \tag{17}$$

This does not depend on $\boldsymbol{w}$, and for given $\boldsymbol{\sigma}$ is minimized by assignments described as follows. Assume that the groups have been relabelled, if necessary, so that $\sigma_1^2 = \max_i\sigma_i^2$. Then $\tau_i \stackrel{def}{=} \sigma_1^2/\sigma_i^2\geq 1$. Define a function, for $\kappa\geq 0$, by

$$h(\kappa) = \kappa - 1 - \sum_i\frac{\tau_i - 1}{p-1+\frac{\tau_i}{\kappa}}.$$

Then there is a unique zero $\kappa_*\geq 1$ of $h(\kappa)$, and (17) is minimized by $\boldsymbol{\rho}^* = \left(\rho_1^*,...,\rho_p^*\right)'$ given by $\rho_i^* = 1/\left(p-1+\frac{\tau_i}{\kappa_*}\right)$ if $i>1$, and $\rho_1^* = 1-\sum_{i=2}^p\rho_i^*$.

Remarks: Corollary 1 is proven as Theorem 3 of Wiens (2005). From this corollary it follows that $\kappa_* = \sum_{i=1}^{p} \tau_i \rho_i^*$. When $p = 2$, we find $\rho_i^* = \sigma_i / (\sigma_1 + \sigma_2)$. When $p = 3$ the ρ_i^* are obtained from

$$\kappa_* = \sqrt{\frac{\tau_2\tau_3 + \tau_2 + \tau_3}{3}} \cos\left(\frac{1}{3}\arctan\sqrt{\frac{(\tau_2\tau_3 + \tau_2 + \tau_3)^3}{27\tau_2^2\tau_3^2} - 1}\right).$$

When all σ_i^2 are equal, the corollary gives $\rho_i^* \equiv 1/p$.

4 Minimax prediction

A problem of obvious interest is that of estimating the mean responses to treatments, given the covariates. One will estimate $E[Y|i, \boldsymbol{x}] = \boldsymbol{r}_i'(\boldsymbol{x})\boldsymbol{\theta} + \psi_{n,i}(\boldsymbol{x})$ by $\hat{Y}_i(\boldsymbol{x}) = \boldsymbol{r}_i'(\boldsymbol{x})\hat{\boldsymbol{\theta}}$, with asymptotic *mse*

$$\text{MSE}_i(\boldsymbol{x}) = \lim_{n\to\infty} E\left[\left\{\sqrt{n}\left(\boldsymbol{r}_i'(\boldsymbol{x})\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) - \psi_{n,i}(\boldsymbol{x})\right)\right\}^2 |\mathcal{F}_n\right]$$

and integrated (i.e., expected w.r.t. $\boldsymbol{x}$) *mse*

$$\text{IMSE}_i = E[\text{MSE}_i(\boldsymbol{x})] = \int_{\chi} \text{MSE}_i(\boldsymbol{x})\, m(\boldsymbol{x})\, \mu(d\boldsymbol{x}).$$

Theorem 3 *(i) With notation as above, we have that* $\mathcal{L}_2(\boldsymbol{\rho}, \boldsymbol{w}, \boldsymbol{\sigma}, \boldsymbol{\psi}) \stackrel{def}{=} \sum_{i=1}^{p}\text{IMSE}_i$ *is given by*

$$\mathcal{L}_2(\boldsymbol{\rho}, \boldsymbol{w}, \boldsymbol{\sigma}, \boldsymbol{\psi}) = tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\} + \left\{\left\|\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{q}_{\boldsymbol{t},\boldsymbol{\psi}}\right\|^2 + E\left[\|\boldsymbol{\psi}(\boldsymbol{x})\|^2\right]\right\}.$$

(ii) Define $\boldsymbol{K}_{\boldsymbol{t}} = E\left[\boldsymbol{R}'(\boldsymbol{x})\,\boldsymbol{A}_{\boldsymbol{t}}^2(\boldsymbol{x})\,\boldsymbol{R}(\boldsymbol{x})\right]$. *The maximum loss over* $\boldsymbol{\psi}$, *subject to (9) and (10) is obtained from*

$$\max_{\boldsymbol{\psi}}\left\{\left\|\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{q}_{\boldsymbol{t},\boldsymbol{\psi}}\right\|^2 + E\left[\|\boldsymbol{\psi}(\boldsymbol{x})\|^2\right]\right\} = \eta^2 ch_{\max}\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{K}_{\boldsymbol{t}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\}. \tag{18}$$

Thus with $\nu = \eta^2/(1+\eta^2)$ *the maximum loss* $\max_{\boldsymbol{\psi}}\mathcal{L}_2(\boldsymbol{\rho}, \boldsymbol{w}, \boldsymbol{\sigma}, \boldsymbol{\psi})$ *is* $1+\eta^2$ *times*

$$\mathcal{L}_2' = (1-\nu)\, tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\} + \nu ch_{\max}\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{K}_{\boldsymbol{t}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\}.$$

(iii) Define $\boldsymbol{L}_{\boldsymbol{t}}(\boldsymbol{x}) = \boldsymbol{A}_{\boldsymbol{t}}(\boldsymbol{x})\,\boldsymbol{R}(\boldsymbol{x})\,\boldsymbol{M}_{\boldsymbol{t}}^{-2}\boldsymbol{R}'(\boldsymbol{x})$, *with diagonal elements*

$$\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x}) = t_i(\boldsymbol{x})\,\boldsymbol{r}_i'(\boldsymbol{x})\,\boldsymbol{M}_{\boldsymbol{t}}^{-2}\boldsymbol{r}_i(\boldsymbol{x}).$$

The maximum loss over $\boldsymbol{\sigma}$, *subject to (7), is obtained from*

$$\max_{\boldsymbol{\sigma}} tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\} = \sigma_0^2\sqrt{E\left[\sum_{i=1}^{p} w_i^2(\boldsymbol{x})\left\{\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x})\right\}^2\right]}. \tag{19}$$

Define $\nu = \eta^2 / (\sigma_0^2 + \eta^2)$*. The maximum of* $\mathcal{L}_2(\boldsymbol{\rho}, \boldsymbol{w}, \boldsymbol{\sigma}, \boldsymbol{\psi})$ *over both* $\boldsymbol{\psi}$ *and* $\boldsymbol{\sigma}$*, subject to these constraints, is* $\sigma_0^2 + \eta^2$ *times*

$$\mathcal{L}_2(\boldsymbol{w}, \boldsymbol{t}) \stackrel{def}{=} \max_{\boldsymbol{\psi}, \boldsymbol{\sigma}} \mathcal{L}_2(\boldsymbol{\rho}, \boldsymbol{w}, \boldsymbol{\sigma}, \boldsymbol{\psi}) / \left(\sigma_0^2 + \eta^2\right)$$
$$= (1-\nu)\sqrt{E\left[\sum_{i=1}^{p} w_i^2(\boldsymbol{x})\left\{\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x})\right\}^2\right]} + \nu ch_{\max}\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{K}_{\boldsymbol{t}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\}. \qquad (20)$$

Example continued: The matrix $\boldsymbol{K}_{\boldsymbol{t}}$ has the same structure as $\boldsymbol{M}_{\boldsymbol{t}}$, but each $t_i(\boldsymbol{x})$ is replaced by its square.

Note: I see two ways to proceed from here.

1. We could consider only the loss maximized over $\boldsymbol{\psi}$. For this we could proceed sequentially, estimating variances as we go along then using efficient weights – see §4.1. Or we could choose weights and variance functions and minimize $\mathcal{L}_2'$ numerically, over parametric classes $\{\rho_i(\boldsymbol{x}; \boldsymbol{\phi})\}$.

   Example continued: I computed

   $$\mathcal{L}_2' = (1-\nu)\, tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\} + \nu ch_{\max}\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{K}_{\boldsymbol{t}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\}$$

   for $\sigma_i(\boldsymbol{x}) = i^2(12 + x_1 + 10x_2)$ and both with weights $w_i(\boldsymbol{x}) = 1/\sigma_i^2(\boldsymbol{x})$ (in which case $\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}} = \boldsymbol{M}_{\boldsymbol{t}}$ and $t_i(\boldsymbol{x}) = \rho_i(\boldsymbol{x})/\sigma_i^2(\boldsymbol{x})$) and $w_i(\boldsymbol{x}) \equiv 1$ (OLS). I looked at several parametric families $\{\rho_i(\boldsymbol{x}; \boldsymbol{\phi})\}$, each of the form $\rho_1(\boldsymbol{x}; \boldsymbol{\phi}) = 1/\left(1 + e^{S(\boldsymbol{x};\boldsymbol{\phi})}\right)$, $\rho_2(\boldsymbol{x}; \boldsymbol{\phi}) = 1 - \rho_1(\boldsymbol{x}; \boldsymbol{\phi})$: (recall $\boldsymbol{x} = (\pm 1, x_2)$ and $i = 1, 2$), and one five-parameter family:

   (a) $S(\boldsymbol{x}; \boldsymbol{\phi}) = \phi_1 + \phi_2 x_1 + \phi_3 x_2$,

   (b) $S(\boldsymbol{x}; \boldsymbol{\phi}) = \phi_1 + \phi_2 x_1 + \phi_3 x_2 + \phi_4 x_2^2$,

   (c) $S(\boldsymbol{x}; \boldsymbol{\phi}) = \phi_1 + \phi_2 x_1 + \phi_3 x_2 + \phi_4 x_2^2 + \phi_5 x_1$.

   Recall $\boldsymbol{x} = (\pm 1, x_2)$. Thus the expectations over x_1 are simple to compute - they are just averages of functions evaluated at $x_1 = 1$ and $x_1 = -1$. Then the integrations over x_2 were done using Simpson's Rule. The MATLAB minimizer `fminsearch` was used to minimize over $\boldsymbol{\phi}$. I prepared plots of $\rho_1(\boldsymbol{x} = (1, x_2); \boldsymbol{\phi})$ and $\rho_1(\boldsymbol{x} = (-1, x_2); \boldsymbol{\phi})$ vs. x_2 – Figures 1 and 2.

2. Minimize $\mathcal{L}_2(\boldsymbol{w}, \boldsymbol{t})$ by minimizing $E\left[\sum_{i=1}^{p} w_i^2(\boldsymbol{x})\left\{\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x})\right\}^2\right]$ over $w_i(\boldsymbol{x}) \in [0,1]$ subject to $E\left[\sum_{i=1}^{p} w_i(\boldsymbol{x})\right] = 1$, thus obtaining minimax weights $w_{*,i}(\boldsymbol{x}; \boldsymbol{t})$, and then minimizing $\mathcal{L}_2(\boldsymbol{w}_*, \boldsymbol{t})$ over $\boldsymbol{t}$ subject to

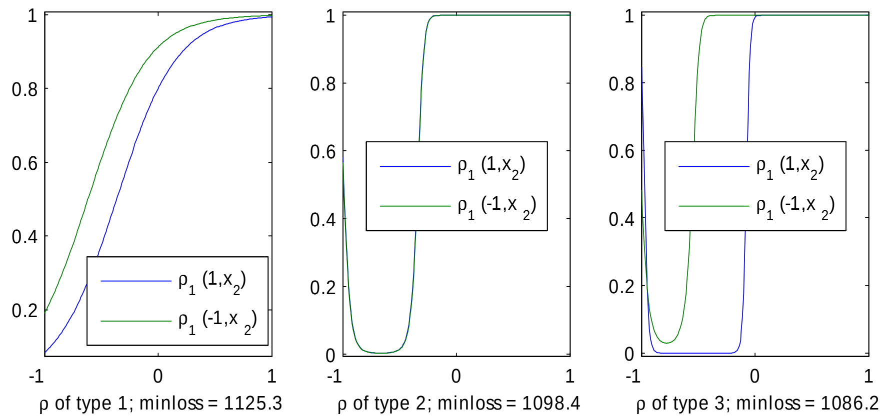

Figure 1: Efficient weights $w_i\left(\boldsymbol{x}\right) \propto 1/\sigma_i^2\left(\boldsymbol{x}\right)$; $\nu = .5$.

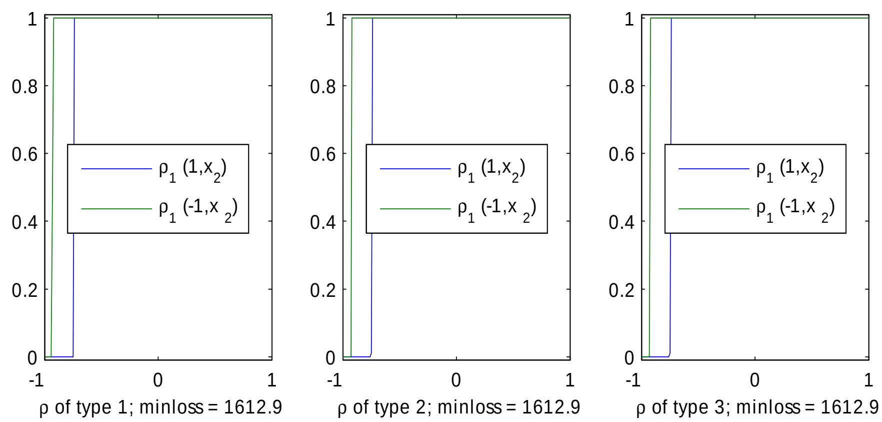

Figure 2: Constant weights, $\nu = .5$.

$\left(\sum_{i=1}^p \rho_i\left(\boldsymbol{x}\right) =\right) \sum_{i=1}^p \frac{t_i(\boldsymbol{x})}{w_i^*(\boldsymbol{x};\boldsymbol{t})} = 1$. A problem is that for fixed $\boldsymbol{x}$, the minimizing weights will be unrealistic – they will be

$$w_i\left(\boldsymbol{x};\boldsymbol{t}\right) = \begin{cases} 1, & \text{if } \left\{\boldsymbol{L}_{\boldsymbol{t},ii}\left(\boldsymbol{x}\right)\right\}^2 = \min_i \left\{\boldsymbol{L}_{\boldsymbol{t},ii}\left(\boldsymbol{x}\right)\right\}^2, \\ 0, & \text{otherwise,} \end{cases}$$

with $\boldsymbol{M}_{\boldsymbol{t}}$ and $\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}$ singular. But note that these weights satisfy $w_i\left(\boldsymbol{x}\right) = w_i^2\left(\boldsymbol{x}\right)$, so that $E\left[\sum_{i=1}^p w_i^2\left(\boldsymbol{x}\right)\right] = 1$, which is the maximum possible. So I propose constraining $E\left[\sum_{i=1}^p w_i^2\left(\boldsymbol{x}\right)\right]$ as well:

$$E\left[\sum_{i=1}^p w_i^2\left(\boldsymbol{x}\right)\right] = w_0^2 \in \left[p^{-1}, 1\right].$$

It is sufficient to find weights $\in [0,1]$ satisfying the constraints and minimizing

$$\begin{aligned} &E\left[\sum_{i=1}^{p} w_i^2(\boldsymbol{x})\left\{\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x})\right\}^2 + \lambda_1 \sum_{i=1}^{p} w_i^2(\boldsymbol{x}) - 2\lambda_2 \sum_{i=1}^{p} w_i(\boldsymbol{x})\right] \\ &= \sum_{i=1}^{p} E\left[w_i^2(\boldsymbol{x})\left(\left\{\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x})\right\}^2 + \lambda_1\right) - 2\lambda_2 w_i(\boldsymbol{x})\right] \end{aligned}$$

for Lagrange multipliers λ_1 and λ_2. For this we minimize the integrands pointwise. Put $y = w_i(\boldsymbol{x})$, $l_i(\boldsymbol{x};\boldsymbol{t},\lambda_1) = \left[\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x})\right]^2 + \lambda_1$ and consider the problem of minimizing the quadratic polynomial $\phi(y) = l_i(\boldsymbol{x};\boldsymbol{t},\lambda_1)\, y^2 - 2\lambda_2 y$ over $y \in [0,1]$. The critical point is $\lambda_2 / l_i(\boldsymbol{x};\lambda_1)$. If this is positive, then $\phi(y)$ is a minimum at this point, and so its minimum in $[0,1]$ is at $\min(\lambda_2/l_i(\boldsymbol{x};\lambda_1), 1)$. If the critical point is negative, then $\phi(y)$ is a maximum at this point and the minimum in $[0,1]$ is at 1. Thus minimizing weights are

$$w_{*,i}(\boldsymbol{x};\boldsymbol{t},\boldsymbol{\lambda}) = \begin{cases} \lambda_2/l_i(\boldsymbol{x};\boldsymbol{t},\lambda_1), & \text{if } 0 < \lambda_2/l_i(\boldsymbol{x};\boldsymbol{t},\lambda_1) < 1, \\ 1, & \text{otherwise.} \end{cases}$$

In order that these weights satisfy the constraints, the multipliers are to be determined from

$$E\left[\sum_{i=1}^{p} w_{*,i}(\boldsymbol{x};\boldsymbol{t},\boldsymbol{\lambda})\right] = 1, \tag{21a}$$

$$E\left[\sum_{i=1}^{p} w_{*,i}^2(\boldsymbol{x};\boldsymbol{t},\boldsymbol{\lambda})\right] = w_0^2. \tag{21b}$$

This now looks like a substantial numerical problem. We are to minimize

$$\mathcal{L}_2(\boldsymbol{w}_*,\boldsymbol{t}) = (1-\nu)\sqrt{E\left[\sum_{i=1}^{p} w_{*,i}^2(\boldsymbol{x};\boldsymbol{t},\boldsymbol{\lambda})\left\{\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x})\right\}^2\right]} + \nu ch_{\max}\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{K}_{\boldsymbol{t}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\}$$

over $\{t_i(\boldsymbol{x})\}_{i=1}^{p}$ subject to

$$\sum_{i=1}^{p} \frac{t_i(\boldsymbol{x})}{w_i^*(\boldsymbol{x};\boldsymbol{t},\boldsymbol{\lambda})} = 1,$$

with $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{t})$ determined from (21).

4.1 Sequentially minimizing the maximum (over $\boldsymbol{\psi}$) loss

See Note 1 above. A possibility is to calculate an estimate of the maximum (over $\boldsymbol{\psi}$) loss after the first n assignments have been made, with n_i to group i, and to then minimize the (updated) estimate over the p possibilities for the $n+1^{th}$ assignment. In the description below I use variances and weights depending only on the treatment group, since estimating variances which depend as well on the covariates might require too much data. But my description is easily modified to take into account other variance structures.

If the first n covariates are $\{\boldsymbol{x}_j\}_{j=1}^n$ and if the sample variance estimates in the p treatment groups are $\{\hat{\sigma}_{i:n}^2\}_{i=1}^p$ then we set $\hat{w}_i = 1/\hat{\sigma}_{i:n}^2$ and estimate

$$\boldsymbol{A_t}\left(\boldsymbol{x}\right) = \oplus_{i=1}^p w_i \rho_i\left(\boldsymbol{x}\right)$$

by

$$\hat{\boldsymbol{A}}_{\boldsymbol{t}}\left(\boldsymbol{x}\right) = \oplus_{i=1}^p \hat{w}_i I_i\left(\boldsymbol{x}\right),$$

where

$$I_i\left(\boldsymbol{x}_j\right) = I(\text{covariates } \boldsymbol{x}_j \text{ resulted in an assignment to group } i).$$

The idea here is that

$$E\left[I_i\left(\boldsymbol{x}_j\right)\right] = P(\text{covariates } \boldsymbol{x}_j \text{ resulted in an assignment to group } i) = \rho_i\left(\boldsymbol{x}_j\right).$$

Similarly, estimate

$$\boldsymbol{A}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}\left(\boldsymbol{x}\right) = \oplus_{i=1}^p w_i^2 \sigma_i^2 \rho_i\left(\boldsymbol{x}\right)$$

by

$$\hat{\boldsymbol{A}}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}\left(\boldsymbol{x}\right) = \oplus_{i=1}^p \hat{w}_i^2 \hat{\sigma}_{i:n}^2 I_i\left(\boldsymbol{x}\right) = \hat{\boldsymbol{A}}_{\boldsymbol{t}}\left(\boldsymbol{x}\right). \tag{22}$$

Then $\boldsymbol{M_t}$ and $\boldsymbol{K_t}$ are estimated by

$$\hat{\boldsymbol{M}}_{\boldsymbol{t}} = \hat{\boldsymbol{M}}_{\boldsymbol{t}}^{(n)} = \frac{1}{n}\sum_{j=1}^n \boldsymbol{R}'\left(\boldsymbol{x}_j\right)\hat{\boldsymbol{A}}_{\boldsymbol{t}}\left(\boldsymbol{x}_j\right)\boldsymbol{R}\left(\boldsymbol{x}_j\right),$$

$$\hat{\boldsymbol{K}}_{\boldsymbol{t}} = \hat{\boldsymbol{K}}_{\boldsymbol{t}}^{(n)} = \frac{1}{n}\sum_{j=1}^n \boldsymbol{R}'\left(\boldsymbol{x}_j\right)\hat{\boldsymbol{A}}_{\boldsymbol{t}}^2\left(\boldsymbol{x}_j\right)\boldsymbol{R}\left(\boldsymbol{x}_j\right),$$

and $\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}^2\boldsymbol{w}}$ has the same estimate as $\boldsymbol{M_t}$, by virtue of (22). The loss after n assignments is thus estimated by

$$\hat{L}_n = tr\left\{\hat{\boldsymbol{M}}_{\boldsymbol{t}}^{-1}\right\} + \eta^2 ch_{\max}\left\{\hat{\boldsymbol{M}}_{\boldsymbol{t}}^{-1}\hat{\boldsymbol{K}}_{\boldsymbol{t}}\hat{\boldsymbol{M}}_{\boldsymbol{t}}^{-1}\right\}. \tag{23}$$

Now if the $n+1^{th}$ arrival presents with covariates $\boldsymbol{x}_{n+1}$, and if the assignment is to group i, then all of these estimates can be updated, yielding possible estimates $\hat{L}_{n+1}\left(i\right)$, $i = 1, ..., p$. Here $\hat{L}_{n+1}\left(i\right)$ is computed as at (23), but with

$$\hat{\boldsymbol{A}}_{\boldsymbol{t}}\left(\boldsymbol{x}_{n+1}\right) = \hat{w}_i diag(0, .., 0, \overset{\overset{i}{\downarrow}}{1}, 0, ..., 0),$$

so that

$$\hat{\boldsymbol{M}}_{\boldsymbol{t}}^{(n+1)} = \frac{n}{n+1}\hat{\boldsymbol{M}}_{\boldsymbol{t}}^{(n)} + \frac{\hat{w}_i}{n+1}\boldsymbol{r}_i\left(\boldsymbol{x}_{n+1}\right)\boldsymbol{r}_i'\left(\boldsymbol{x}_{n+1}\right),$$

$$\hat{\boldsymbol{K}}_{\boldsymbol{t}}^{(n+1)} = \frac{n}{n+1}\hat{\boldsymbol{K}}_{\boldsymbol{t}}^{(n)} + \frac{\hat{w}_i^2}{n+1}\boldsymbol{r}_i\left(\boldsymbol{x}_{n+1}\right)\boldsymbol{r}_i'\left(\boldsymbol{x}_{n+1}\right).$$

The actual assignment is made to group

$$i^* = \arg\min_i \hat{L}_{n+1}(i).$$

The procedure described above has asymptotic optimality properties. In order to state the result we let $\mathcal{M}$ denote the collection of all design probability density functions with respect to $\mu(\boldsymbol{x})$ on a space $\mathcal{X}$. For a design $m \in \mathcal{M}$ and true standard deviations $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_p)'$, we have

$$L(m, \boldsymbol{\sigma}) = tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}(m, \boldsymbol{\sigma})\right\} + \eta^2 ch_{\max}\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}(m, \boldsymbol{\sigma})\boldsymbol{K}_{\boldsymbol{t}}(m, \sigma)\boldsymbol{M}_{\boldsymbol{t}}^{-1}(m, \boldsymbol{\sigma})\right\}, \tag{24}$$

where $\boldsymbol{M}_{\boldsymbol{t}}(m, \boldsymbol{\sigma})$ and $\boldsymbol{K}_{\boldsymbol{t}}(m, \boldsymbol{\sigma})$ denote the $\boldsymbol{M}_{\boldsymbol{t}}$ and $\boldsymbol{K}_{\boldsymbol{t}}$ determined by the design m and standard deviation $\boldsymbol{\sigma}$, respectively. Given $N \geq n$, let $\hat{m}_N$ denote the sequential optimal design constructed as described above, and $\hat{\boldsymbol{\sigma}}_N = (\hat{\sigma}_{1:N}, \cdots, \hat{\sigma}_{p:N})'$ the estimated standard deviation. We have

$$\hat{L}_N(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N) = tr\left\{\hat{\boldsymbol{M}}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N)\right\} + \eta^2 ch_{\max}\left\{\hat{\boldsymbol{M}}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N)\hat{\boldsymbol{K}}_{\boldsymbol{t}N}(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N)\hat{\boldsymbol{M}}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N)\right\},$$

where $\hat{\boldsymbol{M}}_{\boldsymbol{t}N}(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N)$ and $\boldsymbol{K}_{\boldsymbol{t}N}(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N)$ denote the $\boldsymbol{M}_{\boldsymbol{t}}$ and $\boldsymbol{K}_{\boldsymbol{t}}$ determined by the design $\hat{m}_N$ and standard deviation $\hat{\boldsymbol{\sigma}}_N$, respectively.

To ensure the consistency of $\hat{\boldsymbol{\sigma}}_N$ [**I think these have to be regression-based variance estimates, not merely the sample variances of the responses, if we hope to prove them consistent.**] and $\hat{L}_N(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N)$, we need the following two assumptions. Let $\varepsilon_{ij} = (Y_{ij} - E[Y_{ij}|i, \boldsymbol{x}_j])/\sigma_i$. We assume that ε_{ij}'s are independently and identically distributed [**but they aren't - they depend on $\boldsymbol{x}_j$**] and satisfy the moment condition

$$E[|\varepsilon_{ij}|^{4+\delta}] \overset{def}{=} \kappa < \infty, \text{ for some } \delta > 0.$$

Further for the initial number of subjects n, we assume that

$$\lim_{N\to\infty} (\log N)^2/n = 0, \quad \lim_{N\to\infty} n/N = 0.$$

Theorem 4 *Under the two aforementioned assumptions, and as $N \to \infty$,*
(i) $\hat{\sigma}_{i:N} \overset{Pr}{\to} \sigma_i$, for all $i = 1, \cdots, p$; and
(ii) $\hat{L}_N(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N) \overset{Pr}{\to} \min_{m\in\mathcal{M}} L(m, \boldsymbol{\sigma})$.

Example here (continuation). I first simulated N values of x_1 (by randomly permuting $N/2$ 1's and $N/2$ -1's) and of x_2. To simulate x_2 I compute $F^{-1}(U)$, where $F(x_2) = (5x_2^3 + 3x_2 + 8)/16$ $(-1 \leq x_2 \leq 1)$ is the d.f. of X_2 and U is uniformly distributed on $(0,1)$. Then values of Y are simulated, using normally distributed random errors with $\sigma_i = i$, $\alpha = (0,1)'$, $\beta = (1,0)'$ and $\gamma = 1$. Thus in treatment group 1, $E[Y] = X_1 + X_2 + \psi_{n,1}(\mathbf{x})$ and in treatment group 2, $E[Y] = 1 + X_2 + \psi_{n,2}(\mathbf{x})$. To simulate the $\psi_{n,i}$ I compute the

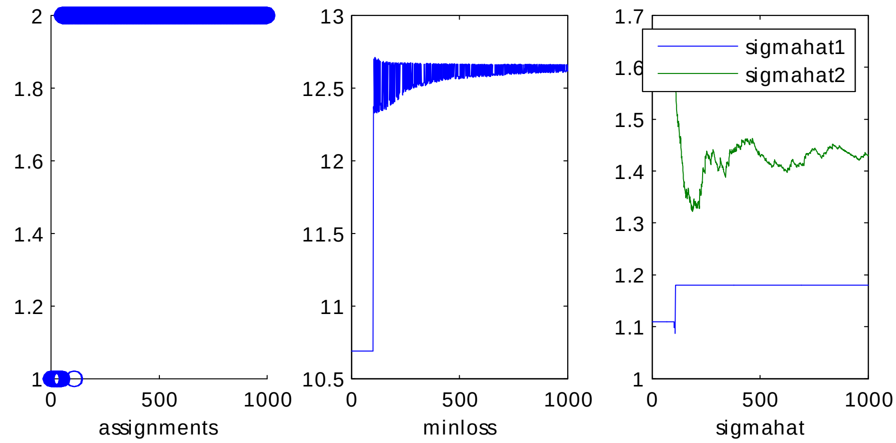

Figure 3: Output for a simulation as described in the continuation of the Example with $N = 1000$ and an intital sample of size 50. True $\sigma_i = \sqrt{i}$ for $i = 1, 2$. Minimum loss using a ρ as in Figure 1. is 7.7.

model matrix $\mathbf{F} = \left[1 \vdots X_1 \vdots X_2\right]$, using all N rows, find the $N \times (N-3)$ matrix $\mathbf{Q}_2$ whose columns form an orthogonal basis for the orthogonal complement of $\mathrm{col}\,(\mathbf{F})$, and then set $\psi_{n,i} = \mathbf{Q}_2\mathbf{c}_i / \left(\sqrt{N}\,\|\mathbf{c}_i\|\right)$, where $\mathbf{c}_i$ is a vector of $N-3$ standard normals. The results are terrible - see Figure 3 for the case that the errors in each group *are* homoscedastic; Figure 4 when they are as used in producing Figures 1 and 2. **What's more, this procedure is deterministic - there is no randomness in the assignments to treatments**.

5 Robust CARA (Covariate-adjusted, response-adaptive) design

Remarks:

1. The *response adaptive* methods of Rosenberger and Sverdlov make the next assignment based on, e.g., the vector of $\hat{\alpha}$'s after the first n assignments. The methods of Atkinson & Biswas, and of Rosenberger et al. mentioned there (§8) look like they might be candidates for robustification.

2. Along the lines of Pronzato (preprint), we could (initially, take $p = 2$ – a treatment and a control, say) try to optimize the probability that assignments are made to the best treatment. This probability could perhaps be represented in the form $F_0\left(\phi\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)$ for an assumed distribution function F_0 and a function $\phi\left(\boldsymbol{x};\boldsymbol{\theta}\right)$ (perhaps $= E\left[Y|\boldsymbol{x}\right]$?); robustness could be brought in by taking neighbourhoods of F_0 (as in Li and Wiens 2011) and ϕ.

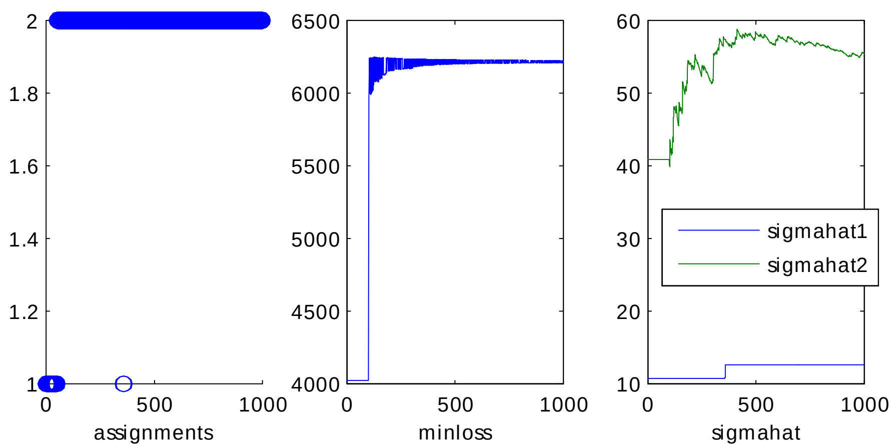

Figure 4: Output for a simulation as described in the continuation of the Example with $N = 1000$ and an intital sample of size 50. True $\sigma_i(x)$ as in Figures 1 and 2. Compare the minimum loss with that using a ρ as in Figure 1.

3. In our notation, the approach of Pronzato seems to be to choose $\rho(\boldsymbol{x})$ in order to minimize a convex combination of a function of the MSE matrix and the 'regret' - the ethical cost of assigning a patient to an inferior treatment. This also looks like a promising area in which to robustify – see $\mathcal{L}_\pi(\boldsymbol{\theta}; F)$ below.

4. I expect that problems such as we are addressing here can also be cast in a machine learning framework – this would certainly increase the possible audience. What those guys call 'active learning' is what we call experimental design.

5. What else?

Here is a possible way to begin. Assume that $p = 2$, so that there is a treatment and a control, say. Then $\boldsymbol{\theta}' = (\alpha_1, \alpha_2, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\gamma}')$. Use a subscript n to indicate evaluation after n treatment allocations have been made and the responses observed. Define

$$\Delta(\boldsymbol{x}; \boldsymbol{\theta}) = E[Y|1, \boldsymbol{x}] - E[Y|2, \boldsymbol{x}] = (\alpha_1 - \alpha_2) + \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2).$$

Suppose that, for some d.f. F, we adopt a rule that we assign to treatment 1 with probability

$$\rho_1^{(n+1)}(\boldsymbol{x}) = F\left(\Delta\left(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_n\right)\right)$$

and otherwise to treatment 2: $\rho_2^{(n+1)}(\boldsymbol{x}) = 1 - \rho_1^{(n+1)}(\boldsymbol{x}) = \bar{F}\left(\Delta\left(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_n\right)\right)$. (This assumes that large values of Y are preferred, i.e. indicative of a successful outcome. If $F = \delta_0$ is point mass at 0, then the assignment is to treatment 1 iff $\Delta\left(\boldsymbol{x}_{n+1}; \hat{\boldsymbol{\theta}}_n\right) \geq 0$. But then there is no randomization, and so clinicians would not be blinded. So we might want F to be in a neighbourhood of a proper d.f. like the Normal.)

Some asymptotics would have to be added in here but I expect we could establish that the theory of §2 continues to hold, with $\rho_1(\boldsymbol{x}) = F(\Delta(\boldsymbol{x};\boldsymbol{\theta}))$, $\rho_2(\boldsymbol{x}) = \bar{F}(\Delta(\boldsymbol{x};\boldsymbol{\theta}))$. (i.e. the estimate is consistent and we can replace the (random) assignment probabilities by their limits in probability.) Treatment 1 is inferior if $\Delta(\boldsymbol{x};\boldsymbol{\theta}) < 0$, and treatment 2 is inferior if $\Delta(\boldsymbol{x};\boldsymbol{\theta}) > 0$. Since $\rho_1^{(n+1)}(\boldsymbol{x}) \to F(\Delta(\boldsymbol{x};\boldsymbol{\theta}))$ and $\rho_2^{(n+1)}(\boldsymbol{x}) \to \bar{F}(\Delta(\boldsymbol{x};\boldsymbol{\theta}))$ the asymptotic probability of an incorrect assignment is then

$$P(\text{error}) \to E\left[F(\Delta(\boldsymbol{x};\boldsymbol{\theta}))\, I(\Delta(\boldsymbol{x};\boldsymbol{\theta}) < 0)\right] + E\left[\bar{F}(\Delta(\boldsymbol{x};\boldsymbol{\theta}))\, I(\Delta(\boldsymbol{x};\boldsymbol{\theta}) > 0)\right] \stackrel{def}{=} P_F(\boldsymbol{\theta}). \tag{25}$$

Now a possible continuation is to consider a convex combination of the asymptotic error probability (25) and the maximum IMSE from some version of Theorem 3:

$$\mathcal{L}_\pi(\boldsymbol{\theta}; F) = (1-\pi)\, P_F(\boldsymbol{\theta}) + \pi \mathcal{L}(\boldsymbol{t}).$$

This could be evaluated at $\rho_1(\boldsymbol{x}) = F(\Delta(\boldsymbol{x};\boldsymbol{\theta}))$, $\rho_2(\boldsymbol{x}) = \bar{F}(\Delta(\boldsymbol{x};\boldsymbol{\theta}))$. Then we could seek an optimizing (minimizing) F. (Perhaps integrate out $\boldsymbol{\theta}$ w.r.t. some prior?) Or leave F fixed and establish asymptotic optimality of a sequential procedure.

Appendix: Derivations

Proof of Theorem 1: The (conditional) bias and covariance of $\hat{\boldsymbol{\theta}}$ are, respectively,

$$\text{BIAS}\left[\hat{\boldsymbol{\theta}}\,|\mathcal{F}_n\right] = \left(\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n}\right)^{-1}\frac{1}{n}\boldsymbol{V}'\boldsymbol{W}\boldsymbol{z},$$
$$\text{COV}\left[\hat{\boldsymbol{\theta}}\,|\mathcal{F}_n\right] = \frac{\sigma_0^2}{n}\left(\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n}\right)^{-1}\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{V}}{n}\left(\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n}\right)^{-1},$$

so that the conditional mean squared error of $\sqrt{n}\hat{\boldsymbol{\theta}}$ is

$$\text{MSE}\left[\sqrt{n}\hat{\boldsymbol{\theta}}\,|\mathcal{F}_n\right] = E\left[\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right)^2\right]$$
$$= \left(\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n}\right)^{-1}\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{z}}{\sqrt{n}}\frac{\boldsymbol{z}'\boldsymbol{W}\boldsymbol{V}}{\sqrt{n}}\left(\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n}\right)^{-1} + \sigma_0^2\left(\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n}\right)^{-1}\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{V}}{n}\left(\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n}\right)^{-1}. \tag{A.1}$$

In terms of the rows $\boldsymbol{r}_i'$ of $\boldsymbol{R}$, these terms are

$$\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n} = \sum_{i=1}^{p}\frac{n_i}{n}\cdot\frac{1}{n_i}\sum_{j=1}^{n_i} w_i\left(\boldsymbol{x}_{ij}\right)\boldsymbol{r}_i\left(\boldsymbol{x}_{ij}\right)\boldsymbol{r}_i'\left(\boldsymbol{x}_{ij}\right),$$
$$\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{V}}{n} = \sum_{i=1}^{p}\frac{n_i}{n}\cdot\frac{1}{n_i}\sum_{j=1}^{n_i} w_i^2\left(\boldsymbol{x}_{ij}\right)\sigma_i^2\left(\boldsymbol{x}_{ij}\right)\boldsymbol{r}_i\left(\boldsymbol{x}_{ij}\right)\boldsymbol{r}_i'\left(\boldsymbol{x}_{ij}\right),$$
$$\frac{1}{\sqrt{n}}\boldsymbol{V}'\boldsymbol{W}\boldsymbol{z} = \sum_{i=1}^{p}\frac{n_i}{n}\cdot\frac{1}{n_i}\sum_{j=1}^{n_i}\left\{w_i\left(\boldsymbol{x}_{ij}\right)\boldsymbol{r}_i\left(\boldsymbol{x}_{ij}\right)\boldsymbol{\cdot}\left(\sqrt{n}\psi_{n,i}\left(\boldsymbol{x}_{ij}\right)\right)\right\}.$$

As each $n_i\to\infty$, by the Strong Law of Large Numbers we have that for functions $\phi_i\left(\boldsymbol{x}\right)$,

$$\frac{n_i}{n}\cdot\frac{1}{n_i}\sum_{j=1}^{n_i}\phi_i\left(\boldsymbol{x}_{ij}\right)\overset{a.s.}{\to}P\left(\text{group } i\right)\cdot E\left[\phi_i\left(\boldsymbol{x}\right)|i\right] = \int_{\chi}\phi_i\left(\boldsymbol{x}\right)\rho_i\left(\boldsymbol{x}\right)m\left(\boldsymbol{x}\right)\mu\left(d\boldsymbol{x}\right).$$

From this observation it follows that

$$\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{V}}{n}\overset{a.s.}{\to}\boldsymbol{M}_{\boldsymbol{t}},$$
$$\frac{\boldsymbol{V}'\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{V}}{n}\overset{a.s.}{\to}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}},$$
$$\frac{1}{\sqrt{n}}\boldsymbol{V}'\boldsymbol{W}\boldsymbol{z}\overset{a.s.}{\to}\boldsymbol{q}_{\boldsymbol{t},\boldsymbol{\psi}};$$

these in (A.1) yield (12). □

Proof of Theorem 2: (i) Define $\boldsymbol{U}_1 = \boldsymbol{U}(\boldsymbol{\rho}, \boldsymbol{w}, \boldsymbol{\sigma}) = \left[\left(\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right)_{11}\right]^{-1}$, and let $\boldsymbol{U}_0 = \left(\oplus_{i=1}^{p}\frac{\sigma_i^2}{t_i}\int_{\chi} w_i(\boldsymbol{x})\, m(\boldsymbol{x})\,\mu(d\boldsymbol{x})\right)^{-1}$ be the evaluation of $\boldsymbol{U}_1$ under (13). By (12),

$$\begin{aligned}\mathcal{L}_1(\boldsymbol{\rho}, \boldsymbol{w}; \boldsymbol{\psi}, \boldsymbol{\sigma}) &= \det\left\{\boldsymbol{\Pi}\boldsymbol{U}_1^{-1}\boldsymbol{\Pi}' + \boldsymbol{\Pi}\left(\left(\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{q}_{\boldsymbol{t},\psi}\right)_1\left(\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{q}_{\boldsymbol{t},\psi}\right)_1'\right)\boldsymbol{\Pi}'\right\} \\ &= \left|\boldsymbol{\Pi}\boldsymbol{U}_1^{-1}\boldsymbol{\Pi}'\right| \cdot \left\{1 + \left(\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{q}_{\boldsymbol{t},\psi}\right)_1'\boldsymbol{\Pi}'\left(\boldsymbol{\Pi}\boldsymbol{U}_1^{-1}\boldsymbol{\Pi}'\right)^{-1}\boldsymbol{\Pi}\left(\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{q}_{\boldsymbol{t},\psi}\right)_1\right\}.\end{aligned}$$

(The subscript 1 refers to the leading $p\times 1$ subvector.) In particular, $\mathcal{L}_1(\boldsymbol{\rho}, \boldsymbol{w}; \boldsymbol{\psi}, \boldsymbol{\sigma}) \geq \left|\boldsymbol{\Pi}\boldsymbol{U}_1^{-1}\boldsymbol{\Pi}'\right|$. But under (13) we have, using (15), that $\left(\boldsymbol{M}_{\rho_0,\boldsymbol{w}}^{-1}\boldsymbol{q}_{\rho_0,\boldsymbol{w},\psi}\right)_1 = \boldsymbol{0}$, whence $\mathcal{L}_1(\boldsymbol{t}; \boldsymbol{\sigma}) = \left|\boldsymbol{\Pi}\boldsymbol{U}_0^{-1}\boldsymbol{\Pi}'\right|$, and it suffices to show that $\left|\boldsymbol{\Pi}\boldsymbol{U}_1^{-1}\boldsymbol{\Pi}'\right| \geq \left|\boldsymbol{\Pi}\boldsymbol{U}_0^{-1}\boldsymbol{\Pi}'\right|$; this in turn will follow if we can establish that

$$\boldsymbol{U}_0 \succeq \boldsymbol{U}_1, \tag{A.2}$$

where '$\succeq$' denotes the ordering by positive semidefiniteness. To show (A.2) we partition the relevant matrices as

$$\boldsymbol{M}_{\boldsymbol{t}} = \begin{pmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12}\\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22}\end{pmatrix},\ \boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}}^{-1} = \begin{pmatrix}\boldsymbol{Q}^{11} & \boldsymbol{Q}^{12}\\ \boldsymbol{Q}^{21} & \boldsymbol{Q}^{22}\end{pmatrix},\ \boldsymbol{M}_{\boldsymbol{t}}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}}^{-1}\boldsymbol{M}_{\boldsymbol{t}} = \begin{pmatrix}\boldsymbol{J}_{11} & \boldsymbol{J}_{12}\\ \boldsymbol{J}_{21} & \boldsymbol{J}_{22}\end{pmatrix},$$

whence $\boldsymbol{U}_1 = \boldsymbol{J}_{11} - \boldsymbol{J}_{12}\boldsymbol{J}_{22}^{-1}\boldsymbol{J}_{21}$, and it suffices to show that

$$\boldsymbol{J}_{11} \preceq \boldsymbol{U}_0. \tag{A.3}$$

We calculate (using identities in Corollaries 1.4.1, 1.4.2 of Khatri and Srivastava (1979)) that

$$\begin{aligned}\boldsymbol{J}_{11} &= \left\{\begin{pmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12}\\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22}\end{pmatrix}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}}^{-1}\begin{pmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12}\\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22}\end{pmatrix}\right\}_{11} \\ &= \begin{pmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12}\end{pmatrix}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}}^{-1}\begin{pmatrix}\boldsymbol{M}_{11}\\ \boldsymbol{M}_{21}\end{pmatrix} \\ &= \begin{pmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12}\end{pmatrix}\left\{\begin{pmatrix}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} + \begin{pmatrix}-\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\\ \boldsymbol{I}\end{pmatrix}\boldsymbol{Q}^{22}\begin{pmatrix}-\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{I}\end{pmatrix}\right\}\begin{pmatrix}\boldsymbol{M}_{11}\\ \boldsymbol{M}_{21}\end{pmatrix} \\ &= \boldsymbol{M}_{11}\boldsymbol{Q}_{11}^{-1}\boldsymbol{M}_{11} + \left(\boldsymbol{M}_{12} - \boldsymbol{M}_{11}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\right)\boldsymbol{Q}^{22}\left(\boldsymbol{M}_{12} - \boldsymbol{M}_{11}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\right)' \\ &\preceq \boldsymbol{M}_{11}\boldsymbol{Q}_{11}^{-1}\boldsymbol{M}_{11} = \boldsymbol{U}_0,\end{aligned}$$

where the final equality follows from the assumption of constant variance functions applied to (14). This proves (A.3).

(ii) By Lemma 1 of Wiens (2005), $\left|\boldsymbol{\Pi}\boldsymbol{U}_0^{-1}\boldsymbol{\Pi}'\right| = \boldsymbol{1}_p'\boldsymbol{U}_0\boldsymbol{1}_p p\,|\boldsymbol{U}_0|$, and (16) follows. □

Proof of Theorem 3: (i) Using Theorem 1 we have

$$\begin{aligned}\text{MSE}_i(\boldsymbol{x}) &= \lim_{n\to\infty} E\left[\left\{\sqrt{n}\left(\boldsymbol{r}_i'(\boldsymbol{x})\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) - \psi_{n,i}(\boldsymbol{x})\right)\right\}^2\right] \\ &= \boldsymbol{r}_i'(\boldsymbol{x})\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\left(\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}} + \boldsymbol{q}_{\boldsymbol{t},\psi}\boldsymbol{q}_{\boldsymbol{t},\psi}'\right)\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\}\boldsymbol{r}_i(\boldsymbol{x}) \\ &\quad - 2\lim_{n\to\infty}\left\{E\left[\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)|\mathcal{F}_n\right]'\boldsymbol{r}_i(\boldsymbol{x})\,\psi_{n,i}(\boldsymbol{x})\right\} + \psi_i^2(\boldsymbol{x}),\end{aligned}$$

and

$$\begin{aligned}\mathrm{IMSE}_i &= E\left[\mathrm{MSE}_i\left(\boldsymbol{x}\right)m\left(\boldsymbol{x}\right)\right]\\ &= tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\left(\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}}+\boldsymbol{q}_{\boldsymbol{t},\psi}\boldsymbol{q}_{\boldsymbol{t},\psi}'\right)\boldsymbol{M}_{\boldsymbol{t}}^{-1}\cdot E\left[\boldsymbol{r}_i\left(\boldsymbol{x}\right)\boldsymbol{r}_i'\left(\boldsymbol{x}\right)\right]\right\}\\ &\quad -2\lim_{n\to\infty}\left\{E\left[\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)|\mathcal{F}_n\right]'E\left[\boldsymbol{r}_i\left(\boldsymbol{x}\right)\psi_{n,i}\left(\boldsymbol{x}\right)\right]\right\}+E\left[\psi_i^2\left(\boldsymbol{x}\right)\right].\end{aligned}$$

Using (5a),

$$\begin{aligned}\sum_{i=1}^{p}\mathrm{IMSE}_i &= tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\left(\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}}+\boldsymbol{q}_{\boldsymbol{t},\psi}\boldsymbol{q}_{\boldsymbol{t},\psi}'\right)\boldsymbol{M}_{\boldsymbol{t}}^{-1}E\left[\boldsymbol{R}'\left(\boldsymbol{x}\right)\boldsymbol{R}\left(\boldsymbol{x}\right)\right]\right\}\\ &\quad -2\lim_{n\to\infty}\left\{E\left[\sqrt{n}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)|\mathcal{F}_n\right]'E\left[\boldsymbol{g}\left(\boldsymbol{x}\right)\psi_n\left(\boldsymbol{x}\right)\right]\right\}+E\left[\left\|\boldsymbol{\psi}\left(\boldsymbol{x}\right)\right\|^2\right],\end{aligned}$$

and the result follows from (5b), and (11).

(ii) With

$$\mathcal{B}\left(\boldsymbol{\psi}\right)\overset{def}{=}tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\left(\boldsymbol{q}_{\boldsymbol{t},\psi}\boldsymbol{q}_{\boldsymbol{t},\psi}'\right)\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\}=\left\|E\left[\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{R}'\left(\boldsymbol{x}\right)\boldsymbol{A}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\boldsymbol{\psi}\left(\boldsymbol{x}\right)\right]\right\|^2,$$

we first show that, subject to (9) and (10),

$$\max_{\boldsymbol{\psi}}\left\{\mathcal{B}\left(\boldsymbol{\psi}\right)+E\left[\left\|\boldsymbol{\psi}\left(\boldsymbol{x}\right)\right\|^2\right]\right\}=\eta^2 ch_{\max}\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{K}_{\boldsymbol{t}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\}. \tag{A.4}$$

Since $\mathcal{B}\left(\boldsymbol{\psi}\right)$ increases if $\boldsymbol{\psi}$ is multiplied by a constant exceeding unity, we may assume equality in (10).

Denote by Ψ the class of functions $\boldsymbol{\psi}\left(\boldsymbol{x}\right)$, $\boldsymbol{x}\in\chi$ constrained by (9) and (10). Define

$$\boldsymbol{\Phi}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)=\boldsymbol{A}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\boldsymbol{R}\left(\boldsymbol{x}\right)-\boldsymbol{R}\left(\boldsymbol{x}\right)\boldsymbol{M}_{\boldsymbol{t}}:p\times s,$$

and assume that $\boldsymbol{\rho},\boldsymbol{w}$ are such that $E\left[\boldsymbol{\Phi}_{\boldsymbol{t}}'\left(\boldsymbol{x}\right)\boldsymbol{\Phi}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\right]$ is nonsingular. (If not, take a perturbation – our final result does not require the nonsingularity of this matrix.) It follows from the definition of $\boldsymbol{M}_{\boldsymbol{t}}$, together with (11), that

$$E\left[\boldsymbol{\Phi}_{\boldsymbol{t}}'\left(\boldsymbol{x}\right)\boldsymbol{\Phi}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\right]=E\left[\boldsymbol{R}\left(\boldsymbol{x}\right)\boldsymbol{A}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\boldsymbol{\Phi}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\right]=\boldsymbol{K}_{\boldsymbol{t}}-\boldsymbol{M}_{\boldsymbol{t}}^2.$$

Define

$$\begin{aligned}\boldsymbol{\Theta}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)&=\boldsymbol{\Phi}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\left[E\left[\boldsymbol{\Phi}_{\boldsymbol{t}}'\left(\boldsymbol{x}\right)\boldsymbol{\Phi}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\right]\right]^{-1/2}\\ &=\boldsymbol{\Phi}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\left[\boldsymbol{K}_{\boldsymbol{t}}-\boldsymbol{M}_{\boldsymbol{t}}^2\right]^{-1/2}:p\times s,\end{aligned}$$

and consider the class $\boldsymbol{\Psi}_0=\left\{\boldsymbol{\psi}_{\boldsymbol{\beta}}\left(\boldsymbol{x}\right)=\eta\boldsymbol{\Theta}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\boldsymbol{\beta}\,|\,\left\|\boldsymbol{\beta}_{s\times1}\right\|=1\right\}$. Note that

(1) $E\left[\boldsymbol{\Theta}_{\boldsymbol{t}}'\left(\boldsymbol{x}\right)\boldsymbol{\Theta}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\right]=\boldsymbol{I}_s$,

(2) $E\left[\boldsymbol{R}'\left(\boldsymbol{x}\right)\boldsymbol{\Theta}_{\boldsymbol{t}}\left(\boldsymbol{x}\right)\right]=\boldsymbol{0}_{s\times s}$.

By (1) and (2), $\boldsymbol{\Psi}_0 \subset \Psi$ and all members of $\boldsymbol{\Psi}_0$ attain equality in (10). We claim that for any $\boldsymbol{\psi} \in \Psi$ there is $\boldsymbol{\psi}_{\boldsymbol{\beta}} \in \boldsymbol{\Psi}_0$ with $\mathcal{B}\left(\boldsymbol{\psi}_{\boldsymbol{\beta}}\right) \geq \mathcal{B}(\boldsymbol{\psi})$, so that

$$\sup_{\Psi} \mathcal{B}(\boldsymbol{\psi}) = \sup_{\boldsymbol{\beta}} \mathcal{B}\left(\boldsymbol{\psi}_{\boldsymbol{\beta}}\right). \tag{A.5}$$

For this, let $\boldsymbol{\psi} \in \Psi$ be arbitrary and define

$$\begin{aligned}
\boldsymbol{\alpha}_{\psi} &= E\left[\boldsymbol{M}_{\boldsymbol{t}}^{-1} \boldsymbol{R}'(\boldsymbol{x})\, \boldsymbol{A}_{\boldsymbol{t}}(\boldsymbol{x})\, \boldsymbol{\psi}(\boldsymbol{x})\right], \\
\boldsymbol{\beta}_{\psi} &= \frac{\left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \boldsymbol{M}_{\boldsymbol{t}}^{-1} \boldsymbol{\alpha}_{\psi}}{\left\| \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \boldsymbol{M}_{\boldsymbol{t}}^{-1} \boldsymbol{\alpha}_{\psi} \right\|}, \\
\boldsymbol{\psi}_* &= \eta \boldsymbol{\Theta}_{\boldsymbol{t}}(\boldsymbol{x})\, \boldsymbol{\beta}_{\boldsymbol{\psi}}.
\end{aligned}$$

Then $\boldsymbol{\psi}_* \in \boldsymbol{\Psi}_0$. Since $\mathcal{B}(\boldsymbol{\psi}) = \|\boldsymbol{\alpha}_{\psi}\|^2$, (A.5) will follow from

$$\left\|\boldsymbol{\alpha}_{\boldsymbol{\psi}_*}\right\|^2 \geq \|\boldsymbol{\alpha}_{\psi}\|^2. \tag{A.6}$$

First, from the Cauchy-Schwarz inequality and the identities above we obtain

$$\|\boldsymbol{\alpha}_{\psi}\|^2 \left\|\boldsymbol{\alpha}_{\psi_*}\right\|^2 \geq \left(\boldsymbol{\alpha}_{\psi}' \boldsymbol{\alpha}_{\psi_*}\right)^2 = \eta^2 \left\| \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \boldsymbol{M}_{\boldsymbol{t}}^{-1} \boldsymbol{\alpha}_{\psi} \right\|^2. \tag{A.7}$$

Similarly,

$$\begin{aligned}
\eta^2 &\geq \left\{ E\left[\|\boldsymbol{\psi}(\boldsymbol{x})\|^2\right] \cdot E\left[\|\boldsymbol{\psi}_*(\boldsymbol{x})\|^2\right] \right\}^{1/2} \\
&\geq \left| E\left[\boldsymbol{\psi}'(\boldsymbol{x})\, \boldsymbol{\psi}_*(\boldsymbol{x})\right] \right| \\
&= \eta \frac{\|\boldsymbol{\alpha}_{\psi}\|^2}{\left\| \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \boldsymbol{M}_{\boldsymbol{t}}^{-1} \boldsymbol{\alpha}_{\psi} \right\|},
\end{aligned}$$

so that

$$\left\| \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \boldsymbol{M}_{\boldsymbol{t}}^{-1} \boldsymbol{\alpha}_{\psi} \right\| \geq \frac{\|\boldsymbol{\alpha}_{\psi}\|^2}{\eta}. \tag{A.8}$$

From (A.7) and (A.8),

$$\|\boldsymbol{\alpha}_{\psi}\|^2 \left\|\boldsymbol{\alpha}_{\psi_*}\right\|^2 \geq \|\boldsymbol{\alpha}_{\psi}\|^4,$$

yielding (A.6) and hence (A.5).

We must now maximize

$$\mathcal{B}\left(\boldsymbol{\psi}_{\boldsymbol{\beta}}\right) = \eta^2 \boldsymbol{\beta}' \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \boldsymbol{M}_{\boldsymbol{t}}^{-2} \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \boldsymbol{\beta}$$

over $\|\boldsymbol{\beta}\| = 1$, obtaining

$$\begin{aligned}
\max \mathcal{B}(\boldsymbol{\psi}) &= \eta^2 ch_{\max} \left\{ \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \boldsymbol{M}_{\boldsymbol{t}}^{-2} \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right]^{1/2} \right\} \\
&= \eta^2 ch_{\max} \left\{ \boldsymbol{M}_{\boldsymbol{t}}^{-1} \left[\boldsymbol{K}_{\boldsymbol{t}} - \boldsymbol{M}_{\boldsymbol{t}}^2\right] \boldsymbol{M}_{\boldsymbol{t}}^{-1} \right\} \\
&= \eta^2 ch_{\max} \boldsymbol{M}_{\boldsymbol{t}}^{-1} \boldsymbol{K}_{\boldsymbol{t}} \boldsymbol{M}_{\boldsymbol{t}}^{-1} - \eta^2,
\end{aligned}$$

from which (A.4) and then (18) follow.

It remains to establish (19). Denote by $\boldsymbol{d}(\boldsymbol{x})$ the p-vector with (non-negative) elements

$$\begin{aligned} d_i(\boldsymbol{x}) &= \left(\left[\oplus_{i=1}^{p}\rho_i(\boldsymbol{x})\,w_i^2(\boldsymbol{x})\right]\boldsymbol{R}(\boldsymbol{x})\,\boldsymbol{M}_{\boldsymbol{t}}^{-2}\boldsymbol{R}'(\boldsymbol{x})\right)_{ii} \\ &= w_i(\boldsymbol{x})\left[\boldsymbol{A}_{\boldsymbol{t}}(\boldsymbol{x})\,\boldsymbol{R}(\boldsymbol{x})\,\boldsymbol{M}_{\boldsymbol{t}}^{-2}\boldsymbol{R}'(\boldsymbol{x})\right]_{ii} \\ &= w_i(\boldsymbol{x})\,\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x}). \end{aligned}$$

Then using (7) and the Cauchy-Schwarz inequality,

$$\begin{aligned} tr\left\{\boldsymbol{M}_{\boldsymbol{t}}^{-1}\boldsymbol{Q}_{\boldsymbol{t},\boldsymbol{\sigma}}\boldsymbol{M}_{\boldsymbol{t}}^{-1}\right\} &= \int_{\chi}\boldsymbol{d}'(\boldsymbol{x})\,\boldsymbol{\sigma}^2(\boldsymbol{x})\,m(\boldsymbol{x})\,\mu(d\boldsymbol{x}) \le \sigma_0^2\sqrt{E\left[\|\boldsymbol{d}(\boldsymbol{x})\|^2\right]} \\ &= \sigma_0^2\sqrt{E\left[\sum_{i=1}^{p}w_i^2(\boldsymbol{x})\left\{\boldsymbol{L}_{\boldsymbol{t},ii}(\boldsymbol{x})\right\}^2\right]}, \end{aligned}$$

and thus bound is attained by

$$\boldsymbol{\sigma}_*^2(\boldsymbol{x}) = \sigma_0^2\frac{\boldsymbol{d}(\boldsymbol{x})}{\sqrt{E\left[\|\boldsymbol{d}(\boldsymbol{x})\|^2\right]}}.$$

Now (20) is immediate. □

Proof of Theorem 4: (i) The proof follows the same argument of that in the proof of Theorem 4 (i) in Wiens and Li (2014). [**See my earlier note - Pengfei and I just took the sample variances of the responses. Now I think we need to estimate the regression parameters and use the mse of the residuals to estimate the variances.**] We will not repeat it here. In addition, we have

$$\hat{\sigma}_{i:n} \xrightarrow{Pr} \sigma_i, \quad \text{for all} \;\; i = 1, \cdots, p, \;\; \text{as} \;\; N \to \infty. \tag{A.9}$$

(ii) To show this, we need verify the following three claims. As $N \to \infty$,

(C1) $\min_{m\in\mathcal{M}} L(m, \hat{\boldsymbol{\sigma}}_n) \xrightarrow{Pr} \min_{m\in\mathcal{M}} L(m, \boldsymbol{\sigma})$,

(C2) $\hat{L}_N(\hat{m}_N, \hat{\boldsymbol{\sigma}}_n) - \min_{m\in\mathcal{M}} L(m, \hat{\boldsymbol{\sigma}}_n) \xrightarrow{Pr} 0$, and

(C3) $\hat{L}_N(\hat{m}_N, \hat{\boldsymbol{\sigma}}_N) - \hat{L}_N(\hat{m}_N, \hat{\boldsymbol{\sigma}}_n) \xrightarrow{Pr} 0$.

We start with claim **(C1)**. Let $m_n^* = \arg\min_{m\in\mathcal{M}} L(m, \hat{\boldsymbol{\sigma}}_n)$ and $m^* = \arg\min_{m\in\mathcal{M}} L(m, \boldsymbol{\sigma})$. From the definition of m_n^*, we have $L(m_n^*, \hat{\boldsymbol{\sigma}}_n) \le L(m^*, \hat{\boldsymbol{\sigma}}_n)$. Due to (A.9) and the continuity of $L(m, \boldsymbol{\sigma})$ with respect to $\boldsymbol{\sigma}$, the following holds

$$L(m_n^*, \hat{\boldsymbol{\sigma}}_n) \le L(m^*, \hat{\boldsymbol{\sigma}}_n) \xrightarrow{Pr} L(m^*, \boldsymbol{\sigma}) = \min_{m\in\mathcal{M}} L(m, \boldsymbol{\sigma}). \tag{A.10}$$

The lower limit of $L(m_n^*, \hat{\boldsymbol{\sigma}}_n)$ can be found by the following argument. The continuity of $L(m_n^*, \boldsymbol{\sigma})$ ensures that $L(m_n^*, \hat{\boldsymbol{\sigma}}_n) = L(m_n^*, \boldsymbol{\sigma}) + o(\|\hat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}\|)$. From the definition of m^*, we have $L(m_n^*, \boldsymbol{\sigma}) \geq L(m^*, \boldsymbol{\sigma})$. Therefore, thanks to (A.9),

$$L(m_n^*, \hat{\boldsymbol{\sigma}}_n) = L(m_n^*, \boldsymbol{\sigma}) + o(\|\hat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}\|) \geq L(m^*, \boldsymbol{\sigma}) + o(\|\hat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}\|) \xrightarrow{Pr} L(m^*, \boldsymbol{\sigma}) = \min_{m \in \mathcal{M}} L(m, \boldsymbol{\sigma}). \tag{A.11}$$

Combining (A.10) and (A.11), we have $L(m_n^*, \hat{\boldsymbol{\sigma}}_n) \xrightarrow{Pr} L(m^*, \boldsymbol{\sigma})$, that is claim **(C1)** holds.

Now we turn to claim **(C2)**. To follow the proof of proof of Theorem 3 in Wiens and Li (2014), we need to verify two statements: (a) $L(m_\lambda, \boldsymbol{\sigma})$ is a convex function of λ, where $m_\lambda = (1-\lambda)m_0 + \lambda m_1$, $m_0 = \arg\min_{m \in \mathcal{M}} L(m, \boldsymbol{\sigma})$ and $m_1 \in \mathcal{M}$ for $\lambda \in [0, 1]$; (b) $\hat{L}_{N+1}(\hat{m}_{N+1}, \boldsymbol{\sigma}) = \frac{N+1}{N}\left(\hat{L}_N(\hat{m}_N, \boldsymbol{\sigma}) + o_p(1)\right)$. Note here $\boldsymbol{\sigma}$ could be the true standard deviation or any consistent estimate. The trace of inverse matrix is convex implies that convexity of the first term in (24). The second term is convex due to the fact that both the largest eigenvalue and matrix inverse operators are convex. This verifies statement (a). Now as

$$\boldsymbol{M}_{\boldsymbol{t}(N+1)}(\hat{m}_{N+1}, \boldsymbol{\sigma}) = \frac{N}{N+1}\boldsymbol{M}_{\boldsymbol{t}N}(\hat{m}_N, \boldsymbol{\sigma}) + \frac{w_i}{N+1}\boldsymbol{r}_i(\boldsymbol{x}_{N+1})\boldsymbol{r}_i'(\boldsymbol{x}_{N+1}),$$

its inverse can be written as, according to Binomial inverse theorem,

$$\begin{aligned}
\boldsymbol{M}_{\boldsymbol{t}(N+1)}^{-1}(\hat{m}_{N+1}, \boldsymbol{\sigma}) &= \left(\frac{N}{N+1}\boldsymbol{M}_{\boldsymbol{t}N}(\hat{m}_N, \boldsymbol{\sigma}) + \frac{w_i}{N+1}\boldsymbol{r}_i(\boldsymbol{x}_{N+1})\boldsymbol{r}_i'(\boldsymbol{x}_{N+1})\right)^{-1} \\
&= \frac{N+1}{N}\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma}) - \frac{w_i(N+1)}{N^2}\frac{\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})\boldsymbol{r}_i(\boldsymbol{x}_{N+1})\boldsymbol{r}_i'(\boldsymbol{x}_{N+1})\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})}{1 + \frac{w_i}{N}\boldsymbol{r}_i'(\boldsymbol{x}_{N+1})\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})\boldsymbol{r}_i(\boldsymbol{x}_{N+1})} \\
&= \frac{N+1}{N}\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma}) + \frac{1}{N}\Delta\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma}).
\end{aligned}$$

The eigenvalues of $M_{\boldsymbol{t}}$ are bounded away from 0 and ∞, so do $\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})$ and $\boldsymbol{M}_{\boldsymbol{t}N}^{-2}(\hat{m}_N, \boldsymbol{\sigma})$. As a consequence, $\Delta\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})$ has bounded eigenvalues and its trace is bounded. Therefore, we have

$$tr\left(\boldsymbol{M}_{\boldsymbol{t}(N+1)}^{-1}(\hat{m}_{N+1}, \boldsymbol{\sigma})\right) = \frac{N+1}{N} tr\left(\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})\right) + O_p(\frac{1}{N}). \tag{A.12}$$

Meanwhile

$$\boldsymbol{M}_{\boldsymbol{t}(N+1)}^{-2}(\hat{m}_{N+1}, \boldsymbol{\sigma}) = \frac{(N+1)^2}{N^2}\boldsymbol{M}_{\boldsymbol{t}N}^{-2}(\hat{m}_N, \boldsymbol{\sigma}) + \frac{1}{N}\Delta\boldsymbol{M}_{\boldsymbol{t}N}^{-2}(\hat{m}_N, \boldsymbol{\sigma}), \tag{A.13}$$

where

$$\begin{aligned}
\Delta\boldsymbol{M}_{\boldsymbol{t}N}^{-2}(\hat{m}_N, \boldsymbol{\sigma}) =& \frac{N+1}{N}\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})\Delta\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma}) \\
&+ \frac{N+1}{N}\Delta\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma}) + \frac{1}{N}\left(\Delta\boldsymbol{M}_{\boldsymbol{t}N}^{-1}(\hat{m}_N, \boldsymbol{\sigma})\right)^2
\end{aligned}$$

and it has all eigenvalues bounded.

$$\begin{aligned}\boldsymbol{K}_{\boldsymbol{t}(N+1)}\left(\hat{m}_{N+1},\boldsymbol{\sigma}\right) =& \frac{N}{N+1}\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)+\frac{w_i^2}{N+1}\boldsymbol{r}_i\left(\boldsymbol{x}_{N+1}\right)\boldsymbol{r}_i'\left(\boldsymbol{x}_{N+1}\right),\\ =& \frac{N}{N+1}\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)+\frac{1}{N+1}\Delta\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right),\end{aligned}\tag{A.14}$$

where the eigenvalues of $\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)$ are bounded. Combining (A.13) and (A.14), it is obtained

$$\boldsymbol{M}^{-2}_{\boldsymbol{t}(N+1)}\left(\hat{m}_{N+1},\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}(N+1)}\left(\hat{m}_{N+1},\boldsymbol{\sigma}\right)=\frac{N+1}{N}\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)+\frac{1}{N}\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right),\tag{A.15}$$

where

$$\begin{aligned}\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right)=&\frac{N+1}{N}\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\Delta\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)+\frac{N}{N+1}\Delta\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\\&+\frac{1}{N+1}\Delta\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\Delta\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right).\end{aligned}$$

Similarly, all eigenvalues of $\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right)$ are bounded. From the Weyl inequalities, we have

$$\begin{aligned}&ch_{max}\left(\frac{N+1}{N}\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)+\frac{1}{N}\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)\\&\leq\frac{N+1}{N}ch_{max}\left(\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)+\frac{1}{N}ch_{max}\left(\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)\\&=\frac{N+1}{N}ch_{max}\left(\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)+O_p\left(\frac{1}{N}\right).\end{aligned}\tag{A.16}$$

The Weyl inequalities also implies

$$\begin{aligned}&ch_{max}\left(\frac{N+1}{N}\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)+\frac{1}{N}\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)\\&\geq\max\left\{\frac{N+1}{N}ch_k\left(\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)+\frac{1}{N}ch_l\left(\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)\right\}\\&=\frac{N+1}{N}ch_{max}\left(\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)+O_p\left(\frac{1}{N}\right),\end{aligned}\tag{A.17}$$

where ch_k denotes the k-th largest eigenvalue and $k+l=p+pq_1+q_2+1$. The last equality holds due to that $\frac{1}{N}ch_l\left(\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)\rightarrow 0$ as $N\rightarrow\infty$. Putting (A.16) and (A.17) together, we have

$$\begin{aligned}&ch_{max}\left(\frac{N+1}{N}\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)+\frac{1}{N}\Delta_N\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)\\&=\frac{N+1}{N}ch_{max}\left(\boldsymbol{M}^{-2}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_N,\boldsymbol{\sigma}\right)\right)+O_p\left(\frac{1}{N}\right).\end{aligned}$$

Therefore,

$$
\begin{aligned}
& ch_{max}\left(\boldsymbol{M}_{\boldsymbol{t}(N+1)}^{-1}\left(\hat{m}_{N+1}, \boldsymbol{\sigma}\right) \boldsymbol{K}_{\boldsymbol{t}(N+1)}\left(\hat{m}_{N+1}, \boldsymbol{\sigma}\right) \boldsymbol{M}_{\boldsymbol{t}(N+1)}^{-1}\left(\hat{m}_{N+1}, \boldsymbol{\sigma}\right)\right) \\
& = ch_{max}\left(\boldsymbol{M}_{\boldsymbol{t}(N+1)}^{-2}\left(\hat{m}_{N+1}, \boldsymbol{\sigma}\right) \boldsymbol{K}_{\boldsymbol{t}(N+1)}\left(\hat{m}_{N+1}, \boldsymbol{\sigma}\right)\right) \\
& = \frac{N+1}{N} ch_{max}\left(\boldsymbol{M}_{\boldsymbol{t}N}^{-2}\left(\hat{m}_{N}, \boldsymbol{\sigma}\right) \boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_{N}, \boldsymbol{\sigma}\right)\right) + O_p\left(\frac{1}{N}\right) \\
& = \frac{N+1}{N} ch_{max}\left(\boldsymbol{M}_{\boldsymbol{t}N}^{-1}\left(\hat{m}_{N}, \boldsymbol{\sigma}\right) \boldsymbol{K}_{\boldsymbol{t}N}\left(\hat{m}_{N}, \boldsymbol{\sigma}\right) \boldsymbol{M}_{\boldsymbol{t}N}^{-1}\left(\hat{m}_{N}, \boldsymbol{\sigma}\right)\right) + O_p\left(\frac{1}{N}\right). \qquad \text{(A.18)}
\end{aligned}
$$

Combining (A.12) and (A.18), we obtain

$$
\hat{L}_{N+1}\left(\hat{m}_{N+1}, \boldsymbol{\sigma}\right) = \frac{N+1}{N} \hat{L}_{N}\left(\hat{m}_{N}, \boldsymbol{\sigma}\right) + O_p\left(\frac{1}{N}\right),
$$

which is statement (b).

The claim **(C3)** follows from the continuity of $\hat{L}_N\left(\hat{m}_N, \boldsymbol{\sigma}\right)$ with respect to $\boldsymbol{\sigma}$ and the consistency of $\hat{\boldsymbol{\sigma}}_n$ and $\hat{\boldsymbol{\sigma}}_N$. □

Acknowledgements

This research is supported by the Natural Sciences and Engineering Research Council of Canada.

References

Li, P., and Wiens, D. P. (2011), "Robustness of Design for Dose-Response Studies," *Journal of the Royal Statistical Society (Series B)*, 17, 215-238.

Srivastava, M. S., and Khatri, C. G. (1979), *An Introduction to Multivariate Statistics.* North Holland.

Wiens, D. P. (2000); "Designs for Clinical Trials," U*niversity of Alberta Statistics Centre Technical Report Series*; #00.06.

Wiens, D. P. (2005), "Robust Allocation Schemes for Clinical Trials With Prognostic Factors," *Journal of Statistical Planning and Inference*, 127, 323-340.

Wiens, D. P., and Li, P. (2014), "V-optimal Designs for Heteroscedastic Regression," *Journal of Statistical Planning and Inference*, 145, 125-138.