Sequential, robust design strategies

International Conference On Robust Statistics

May, 2002

Doug Wiens

University of Alberta

preprints, etc.: www.stat.ualberta.ca

Approximate regression models

- Experimenter fits a response $\hat{Y}(\mathbf{x}) = f\left(\mathbf{x}; \hat{\boldsymbol{\theta}}\right)$ by regression, when in fact

$$E[Y|\mathbf{x}] \approx f\left(\mathbf{x}; \boldsymbol{\theta}\right).$$

- The points $\mathbf{x}_i$ at which Y will be observed are to be chosen with an eye to protection against a misspecified response function.

- Best fitting parameter is

$$\boldsymbol{\theta}_0 = \arg\min \int_S \left\{E([Y|\mathbf{x}] - f\left(\mathbf{x}; \boldsymbol{\theta}\right)\right\}^2 d\mathbf{x}$$

for $\mathbf{x} \in \mathcal{S}$ ("design space").

- Put $g(\mathbf{x}) = E\left[Y|\mathbf{x}\right] - f\left(\mathbf{x}; \boldsymbol{\theta}_0\right)$; then (additive errors)

$$Y(\mathbf{x}) = f\left(\mathbf{x}; \boldsymbol{\theta}_0\right) + g(\mathbf{x}) + \varepsilon.$$

PROBLEM: Choose a design ξ ($=$ a measure placing mass n^{-1} at selected points $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{S}$) so as to minimise loss due to:

- random variation; depends only on ξ
- bias (of $\hat{Y}(\mathbf{x})$ as estimate of $E[Y|\mathbf{x}]$; depends on (g, ξ))

Loss: Integrated MSE of the predictions

$$\begin{aligned}\mathcal{L}(g, \xi) &= \int_{\mathcal{S}} E\left[\left\{\hat{Y}(\mathbf{x}) - E(Y|\mathbf{x})\right\}^2\right] d\mathbf{x} \\ &= \int_{\mathcal{S}} VAR\left[\hat{Y}(\mathbf{x})\right] d\mathbf{x} \\ &\quad + \int_{\mathcal{S}} \left\{E\left[f\left(\mathbf{x}; \hat{\boldsymbol{\theta}}\right) - f\left(\mathbf{x}; \boldsymbol{\theta}_0\right) - g(\mathbf{x})\right]\right\}^2 d\mathbf{x}\end{aligned}$$

- Find $\xi_0 = \arg\min \mathcal{L}(g, \xi)$ after
(i) maximising over g $(= E[Y|\mathbf{x}] - f(\mathbf{x}; \boldsymbol{\theta}_0))$; or
(ii) estimating g.

- Sequential strategy may be called for, in either case

- $\hat{\boldsymbol{\theta}}$ can be LSE, or M-estimate (with σ^2 replaced by, e.g., $\sigma^2 E\left[\psi^2\right] / \left(E\left[\psi'\right]\right)^2$).

NONLINEAR REGRESSION (with Sanjoy Sinha):

Fit $E[Y|\mathbf{x}] = f(\mathbf{x}; \boldsymbol{\theta}_0)$ when in fact this is only approximate, e.g.

$$f(x; \boldsymbol{\theta}_0) = \theta_0 e^{-\theta_1 x} \text{ but } E[Y|\mathbf{x}] = \frac{\theta_0 x}{\theta_1 + x}.$$

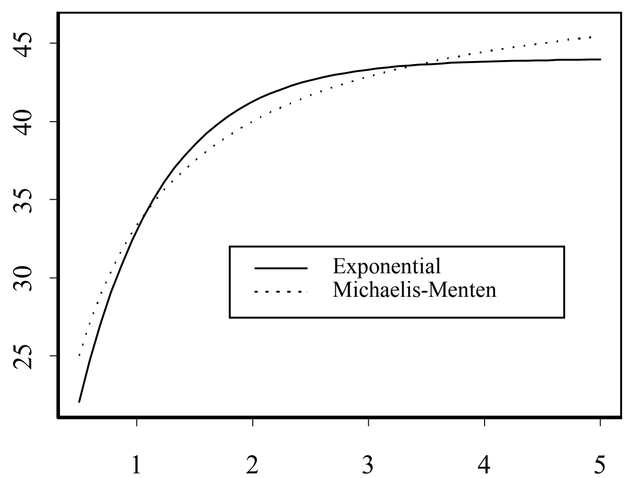

Figure 1: $E[Y|x]$ is Michaelis-Menten with $\boldsymbol{\theta} = (50, .5)^T$; best-fitting exponential is $f(x; \boldsymbol{\theta}_0)$ with $\boldsymbol{\theta}_0 = (44, 1.39)^T$. ($\boldsymbol{\theta}_0 = \arg\min \int_{.5}^{5} \{E([Y|\mathbf{x}] - f(\mathbf{x}; \boldsymbol{\theta})\}^2 \, d\mathbf{x}$.)

$$g(\mathbf{x};\boldsymbol{\theta}_0) = E\left[Y|\mathbf{x}\right] - f\left(\mathbf{x};\boldsymbol{\theta}_0\right)$$

Asymptotic MSE matrix is $\mathsf{MSE}_N(\boldsymbol{\theta}_0) =$

$$\mathbf{M}_N^{-1}(\boldsymbol{\theta}_0)\left\{\mathbf{Q}_N(\boldsymbol{\theta}_0) + \mathbf{b}_N(\boldsymbol{\theta}_0)\mathbf{b}_N^T(\boldsymbol{\theta}_0)\right\}\mathbf{M}_N^{-1}(\boldsymbol{\theta}_0),$$

where $\mathbf{z}(\mathbf{x};\boldsymbol{\theta}) = \partial f(\mathbf{x};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and where

$$\begin{aligned}
\mathbf{M}_N(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta})\mathbf{z}^T(\mathbf{x}_i;\boldsymbol{\theta}), \\
\mathbf{Q}_N(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta})\sigma^2(\mathbf{x}_i)\mathbf{z}^T(\mathbf{x}_i;\boldsymbol{\theta}), \\
\mathbf{b}_N(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta})g(\mathbf{x}_i;\boldsymbol{\theta}).
\end{aligned}$$

Loss is IMSE:

$$\begin{aligned}\mathcal{L}(g,\xi) &= \int_{\mathcal{S}} E\left[\left\{\hat{Y}(\mathbf{x}) - E(Y|\mathbf{x})\right\}^2\right] d\mathbf{x} \\ &\approx tr\left[\mathsf{MSE}_N(\boldsymbol{\theta}_0)\cdot \mathbf{A}(\boldsymbol{\theta}_0)\right] + \int_{\mathcal{S}} g^2(\mathbf{x};\boldsymbol{\theta}_0) d\mathbf{x},\end{aligned}$$

where $\mathbf{A}(\boldsymbol{\theta}) = \int_{\mathcal{S}} \mathbf{z}(\mathbf{x};\boldsymbol{\theta})\mathbf{z}^T(\mathbf{x};\boldsymbol{\theta}) d\mathbf{x}$.

Sequential approach. Given $\{\mathbf{x}_i, Y_i\}_{i=1}^N$:
(i) Compute $\hat{\boldsymbol{\theta}}_N$ and estimates of $g(\mathbf{x})$, $\sigma^2(\mathbf{x})$.
(ii) Using these estimates, estimate $\Delta_{N+1}(\mathbf{x}) =$ increase in $\mathcal{L}$ if the next design point is $\mathbf{x}$.
(iii) Choose $\mathbf{x}_{N+1} = \arg\min \Delta_{N+1}(\mathbf{x})$.

Estimate $g(\mathbf{x})$ by smoothing the residuals (cubic spline in 1-dimensional; generalised additive model for higher dimensions).

Asymptotic results hold for sequentially chosen design points - Sinha and Wiens (2002).

CLINICAL TRIALS: Subjects are assigned to one of p treatment groups. Covariates $\mathbf{x}$ are measured and treatment assignments made, according to a random mechanism.

Optimal assignment probabilities

$$\text{Pr}\,(\text{treatment } i|\mathbf{x}) = \rho_i(\mathbf{x})$$

are to be determined.

Post treatment response to treatment is

$$Y = \theta_i + \mathbf{z}^T(\mathbf{x})\boldsymbol{\phi} + g_i(\mathbf{x}) + \sigma_i\varepsilon$$

for regressors $\mathbf{z}(\mathbf{x})$, error variances σ_i, response errors $g_i(\mathbf{x})$.

$$\text{Design } \xi = \left\{\rho_1, ..., \rho_p\right\}.$$

Let $\mathbf{W}_{p-1\times p}$ have rows which are mutually orthogonal and orthogonal to $\mathbf{1}$. We estimate a complete set $\mathbf{W}\boldsymbol{\theta}$ of contrasts of the treatment effects $\{\theta_i\}_{i=1}^{p}$.

Loss is

$$\mathcal{L}\left(\rho_1, ..., \rho_p\right) = \lim_{n\to\infty} \left| nMSE\left(\mathbf{W}\hat{\boldsymbol{\theta}}\right)\right|.$$

- Heckman (1987) - similar approach; different neighbourhood structure. Under realistic conditions *constant* assignment probabilities were found to be optimal.

It turns out that *constant probabilities*

$$\rho_i(\mathbf{x}) \equiv r_i$$

minimize the COV part of MSE.

Optimal $\{r_i\}_{i=1}^{p}$ are those which

$$\text{minimise } \frac{\sum \left(r_i/\sigma_i^2\right)}{\prod \left(r_i/\sigma_i^2\right)},$$

subject to $\{r_i\}_{i=1}^{p}$ being a probability distribution.

When $p = 2$,

$$r_i = \frac{\sigma_i}{\sigma_1 + \sigma_2}.$$

Sequential assignments. Adjust the (asymptotically) variance minimising $\{r_i\}_{i=1}^{p}$, while also minimising variance and bias in finite samples.

Suppose there are L levels of the (grouped) covariates $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(L)}$. If n assignments have been made, and the $(n+1)^{th}$ subject arrives with covariates $\mathbf{x}_*$, then assign to treatment k with probability

$$P\left(k|\mathbf{x}_*\right) \propto \hat{r}_k d_k^* b_k^*,$$

where:

(i) $\hat{r}_k$ is the optimal r, with the σ_i estimated.

(ii) d_k^* measures the reduction in $\left|COV\left(\mathbf{W}\hat{\boldsymbol{\theta}}\right)\right|$ resulting from an assignment to treatment k.

(iii) b_k^* is inversely proportional to the (finite sample) bias^2 of $\hat{\boldsymbol{\theta}}$, resulting from an assignment to treatment k.

$$P\left(k|\mathbf{x}_*\right) \propto \hat{r}_k d_k^* b_k^*$$

Similar to Atkinson (1982) who takes $P\left(k|\mathbf{x}_*\right) \propto d_k^*$ (assuming no bias, and that all σ_i^2 are equal).

Computation of b_k^* requires $\hat{g}_1(\mathbf{x}), ..., \hat{g}_p(\mathbf{x})$; an *ad hoc* estimate is the adjusted residual

$$\hat{g}_i(\mathbf{x}^{(l)}) = sign\left(\tilde{e}_{i,l}\right)\left(\tilde{e}_{i,l}^2 + \frac{\hat{\sigma}_i^2}{n_{i,l}}\right)^{1/2},$$

where $n_{i,l} = \#$ of assignments of $\mathbf{x}^{(l)}$ to group i; $\tilde{e}_{i,l} =$ median of corresponding residuals.

SPATIAL STUDIES

- Observe $Y(\mathbf{t}) = X(\mathbf{t}) + \varepsilon(\mathbf{t})$ at locations $\mathbf{t} \in \mathcal{T} \subset \mathbb{R}^d$.

- $X(\mathbf{t})$ random: $X(\mathbf{t}) = E[X(\mathbf{t})] + \delta(\mathbf{t})$.

- $E[X(\mathbf{t})] \approx \mathbf{z}^T(\mathbf{t})\boldsymbol{\theta}$ for regressors $\mathbf{z}(\mathbf{t})$

- $VAR[\varepsilon(\mathbf{t})] = f(\mathbf{t})$ only approximately known (assumed constant?)

- $COV[\delta(\mathbf{t}), \delta(\mathbf{t}')] = g(\mathbf{t}, \mathbf{t}')$ only approximately known (assumed isotropic?)

- Choose n locations from $\mathcal{T}$ (with N sites) so as to minimise the MSE of the predictions, maximised over neighbourhoods of the assumed f, g and regression model.

NEXT:

- Sequential choice of sites?

- Simulated annealing?

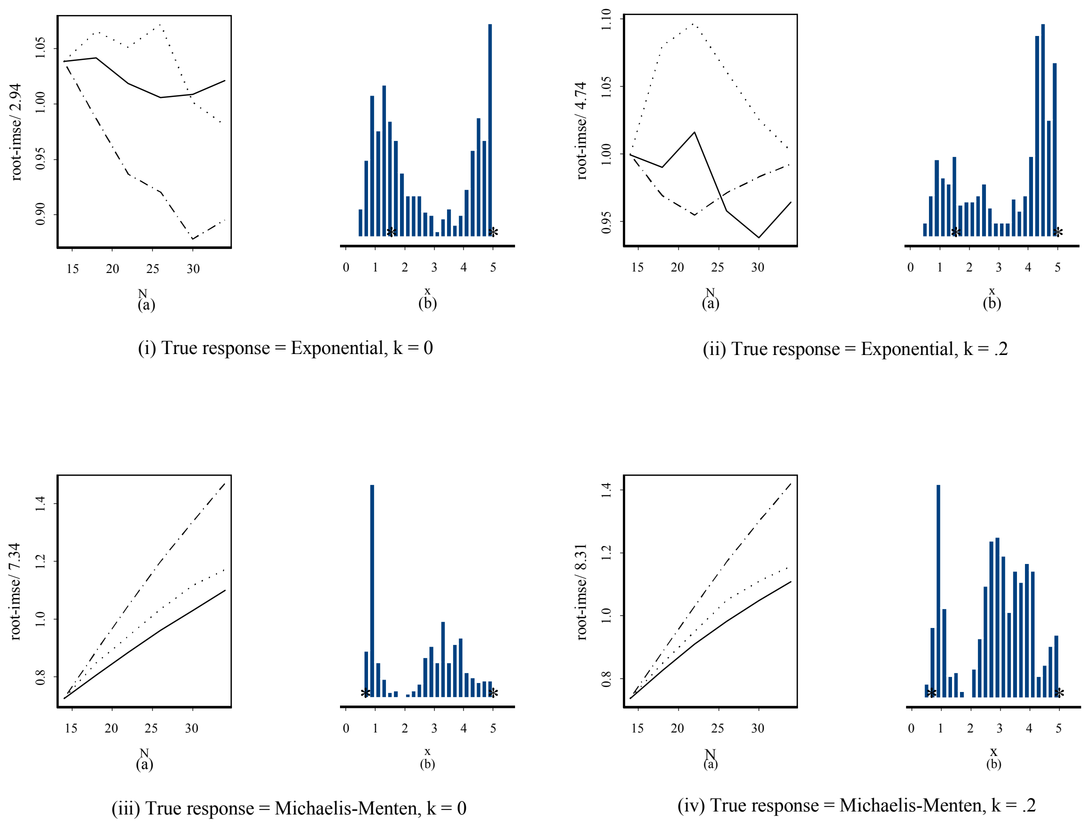

Figure 2: Fitted response is exponential, true response is either exponential or Michaelis-Menten; $n_0 = 10$ equally spaced sites chosen initially, with $r_0 = 3$ replicates at each. Then $n_1 = 6$ additional sites chosen sequentially, with $r_1 = 4$ replicates at each. (a) Average (over 100 sample paths) values of $(N \cdot IMSE)^{1/2}$ for sequential (——), uniform $(\cdots)$ and D-optimal $(- \cdot - \cdot -)$ designs. Variance function is $\sigma^2(x) = 1 + .2(x - .5)^2$. (b) Probability histogram of all points chosen by the 100 sequential designs; asterisks are at the average sites of the D-optimal designs.

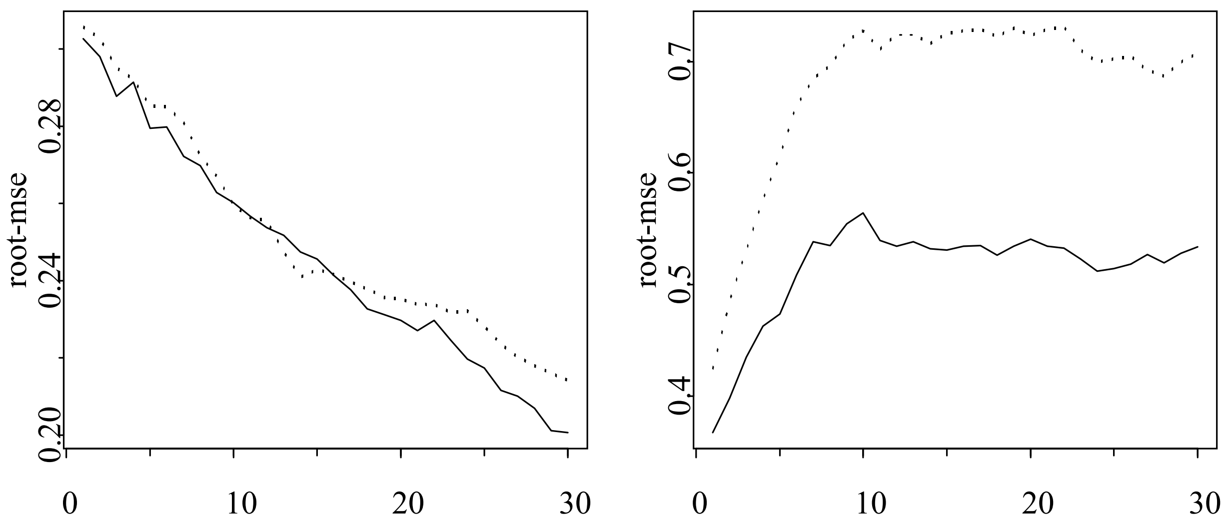

Figure 3: Root-mse of estimated treatment effects versus new subjects; average of 200 simulated runs. Two treatments, two covariates X_1, X_2. Heteroscedastic errors: $\sigma_1^2 = 1$, $\sigma_2^2 = 1/4$. Dotted line is Atkinson's method modified for heteroscedasticity: $P(k|\mathbf{x}_*) \propto \hat{r}_k d_k^*$; solid line is the robust method. Left: $g_1(\mathbf{x}) = g_2(\mathbf{x}) \equiv 0$ (fitted model correct). Right: $g_i(\mathbf{x}) \propto (-1)^i x_1 x_2$.

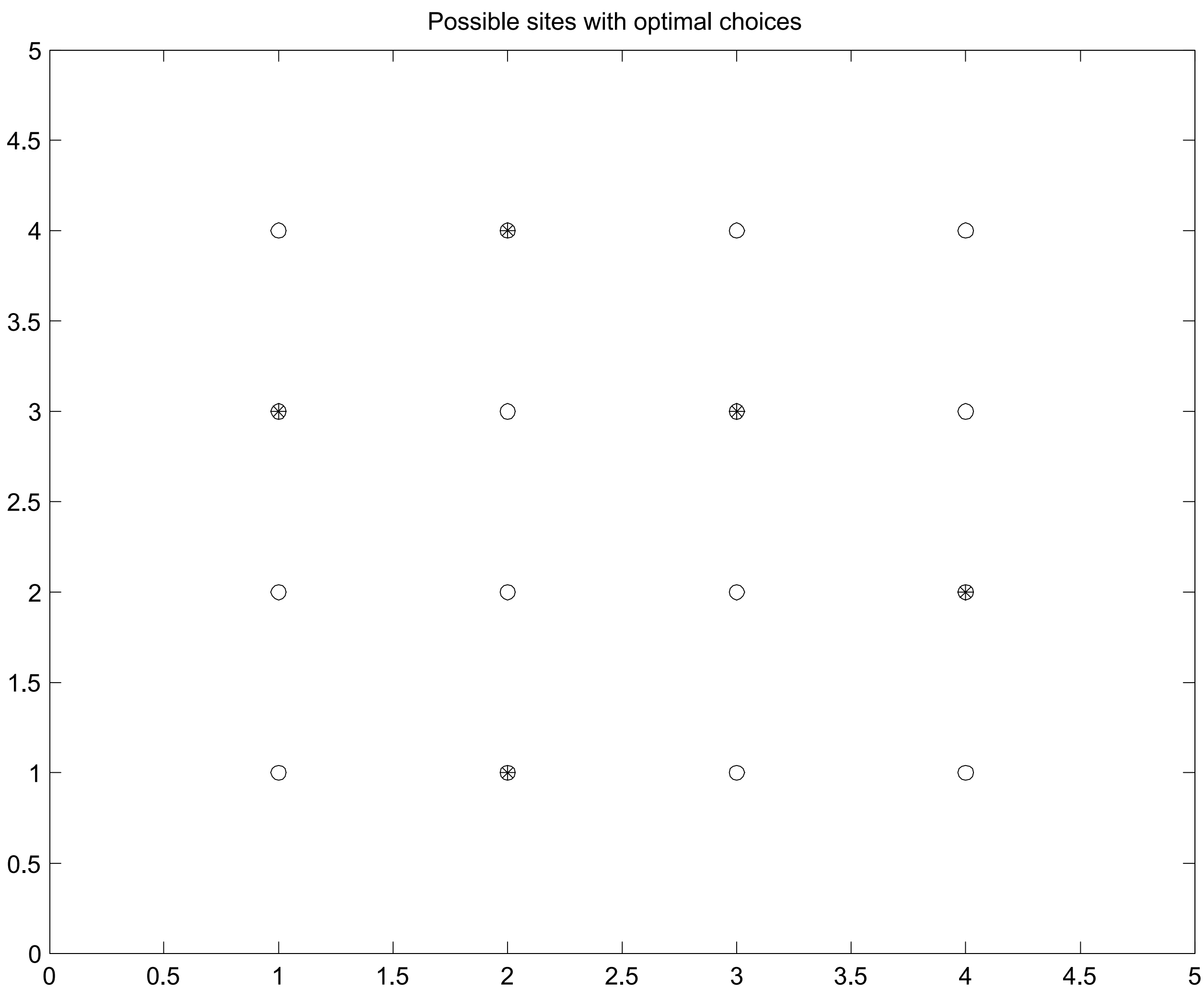

Figure 4: 4×4 grid of possible locations; 5 sites chosen to minimise trace of MSE matrix. Fitted model exact: constant measurement errors, isotropic covariance function $\exp\left(-.2\,\|\mathbf{t}-\mathbf{t}'\|\right)$, regressors $\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^T$.

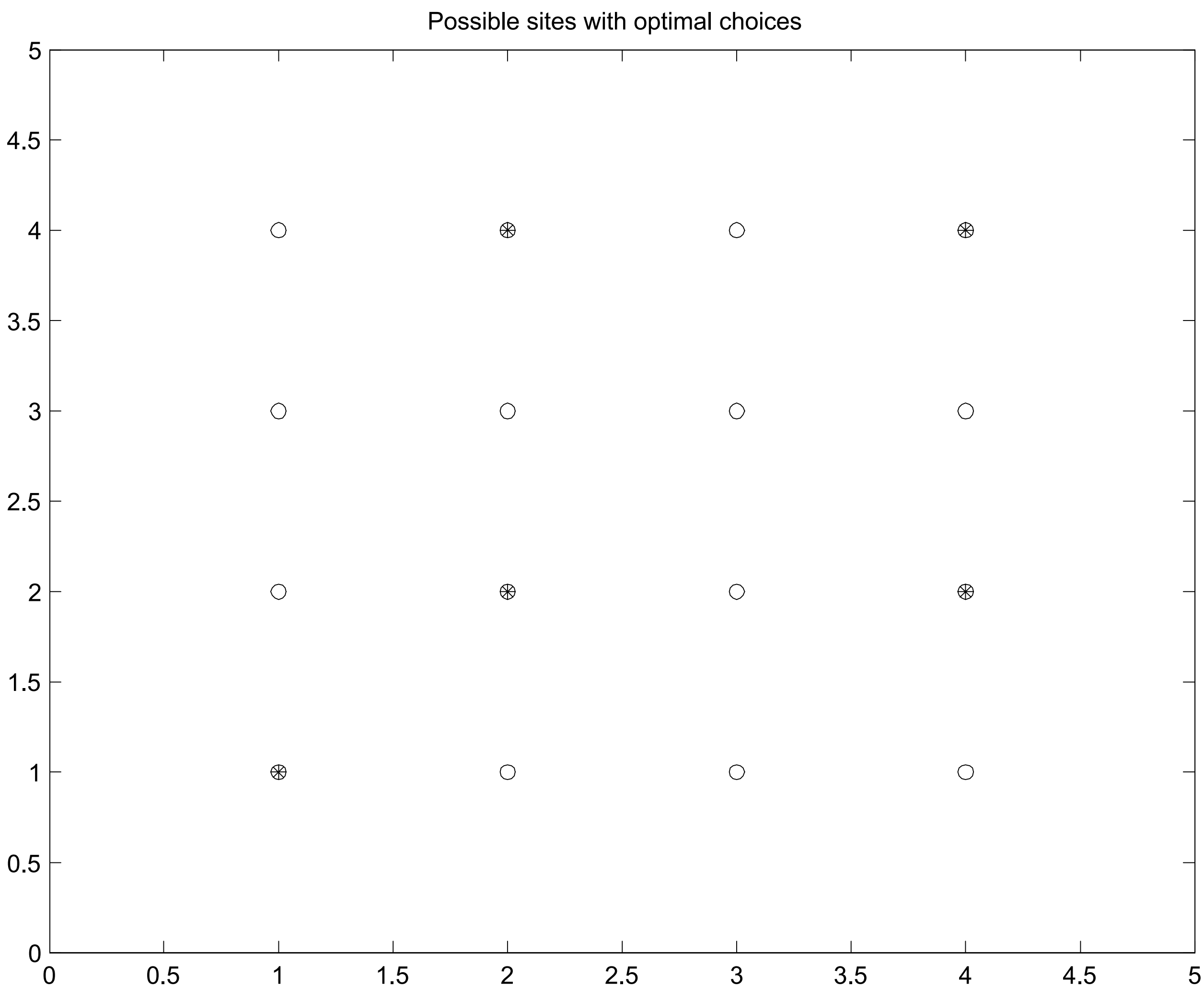

Figure 5: Same fitted model, but loss is maximised over neighbourhoods of the model, then minimised over choices of locations.