Model selection in presence of errors in factor levels

Eva Riccomagno[1,2]

Grazia Vicario[1]

[1]Dipartimento di matematica – Politecnico di Torino, Italy
[2]Department of statistics – The University of Warwick, UK

ICODOE, Memphis TN, May 12-15, 2005

Preliminary

Let D be a design in d factors.

Let $\mathcal{L}(D) = \{f : D \longrightarrow R\}$

Then $\mathcal{L}(D) \sim R[x_1, \ldots, x_d]/\operatorname{Ideal}(D)$ as vector spaces and as rings, even if there are replicated runs in the design.

If the design points are distinct, then vector space bases of $R[x_1, \ldots, x_d]/\operatorname{Ideal}(D)$ correspond to saturated models identified by the design. Moreover if there are n runs in the design, then there are n terms in the saturated identifiable models (Pistone and Wynn, 1996).

Fan of a design, (sub)fan optimal designs: Hugo Maruri-Aguilar

Set-up

A set of control factors $\{x_1, \ldots, x_p\}$

A set of noise variables $\{Z_1, \ldots, Z_q\}$

under the hypothesis $\underline{Z} \sim N_q(\underline{0}, V_{\underline{Z}})$ where $V_{\underline{Z}} = \text{diag}(\sigma_i^2 : i = 1, \ldots, q)$

A finite set of n (distinct) points in $p + q$ dimensions

$$D = \{d_1, \ldots, d_n\}$$

A set of unidimensional responses $\{y_1, \ldots, y_n\}$ at the design points

Identifiable models

A set of saturated polynomial hierarchical identifiable models

$$m_1(\underline{x}, \underline{Z}), \ldots, m_L(\underline{x}, \underline{Z})$$

where for $j = 1, \ldots, L$

$$m_j(\underline{x}, \underline{Z}) = f_j(\underline{x}) + g_j(\underline{Z}) + h_j(\underline{x}, \underline{Z})$$

with $g_j(\underline{0}) = 0 = h_j(\underline{x}, \underline{0}) = h_j(\underline{0}, \underline{Z})$

For m submodel of m_j

$$Y(\underline{x}, \underline{Z}) = m(\underline{x}, \underline{Z}) + \epsilon$$

with $\epsilon \sim N_1(0, \sigma^2)$

Notes

If there are n distinct design points in D then each m_j has n terms

The (full) list of models of the above type could be obtained by algebraic statistics methods

Examples

A. Polynomial regression models

$$Y = f(\underline{x}) + \epsilon$$

B. Random effect models

$$Y = f(\underline{0}) + g(\underline{Z}) + \epsilon$$

C. Error in variable models

$$Y = l(\underline{x} + \underline{Z}) + \epsilon$$

D. Response surface models, in the notation of Myers, Khuri, Vining (1992)

$$Y = \beta_0 + g(\underline{x})^T \underline{\beta} + \underline{\delta}^T \underline{Z} + g(\underline{x})^T \Lambda \underline{Z} + \epsilon$$

with $f(\underline{x}) = \beta_0 + g(\underline{x})^T \underline{\beta}$ and $g(\underline{Z}) = \underline{\delta}^T \underline{Z}$ and $g(\underline{x}, \underline{Z}) = g(\underline{x})^T \Lambda \underline{Z}$

Saturated model selection

$$m_1, \ldots, m_L$$

A. R. and Wynn (1997)

$$H_f = \left[\frac{\partial^2 f(\underline{x})}{\partial x_i \partial x_j} \right]_{i,j=1,\ldots,p}$$

$$\rho_f = \int_{\chi} \text{trace}(H_f^T H_f)(\underline{x})\, d\underline{x}$$

a measure of the curvature of the (fixed effect) model

$$\min_j \rho_j = \rho_*$$
$$\text{argmin}_j\, \rho_j = m_*$$

B. Process mean

$$E(m_j(\underline{0},\underline{Z})) = f_j(\underline{0}) + E(g_j(\underline{Z}))$$

is a known polynomial function of $\sigma_1^2,\ldots,\sigma_q^2$.

If σ_i^2 are known or estimated, then

$$m_* = \text{argmin}_j\, E(g_j(\underline{Z}))$$

If σ_i^2 are unknown and $\sigma_i^2 \in I_i$, then

$$\rho_* = \min_j \max_{\sigma_i \in I_i, i=1,\ldots,q} E(g_j(\underline{Z}))$$

(supposed finite)

D. Myers et al. (1992) for $Y = m(\underline{x},\underline{Z}) + \epsilon$ usually not saturated

$$R(\underline{x}) = \lambda V(x) + (1-\lambda)E((\widehat{Y}(\underline{x},\underline{Z}) - t)^2)$$

with $\lambda \in [0,1]$.

$$m_j(\underline{x},\underline{Z}) = f_j(\underline{x}) + g_j(\underline{Z}) + h_j(\underline{x},\underline{Z})$$

$$\Phi_1(m_j) = \int_{\mathcal{X}} \text{trace}(H_f^T H_f)(\underline{x}) \quad d\underline{x}$$

$$\Phi_2(m_j) = \sum_{\underline{x}\in Design} MSE_{f_j}(m_j(\underline{x},\underline{Z}))$$

$$= \sum_{\underline{x}\in Design} E\left((g_j(\underline{Z}) + h_j(\underline{x},\underline{Z}))^2\right)$$

is a positive function of $\sigma_1^2,\ldots,\sigma_q^2$

If σ_i^2 are known or estimated, then $\Phi_2(m_j)$ is a known positive number

If σ_i^2 are unknown, then

$$\Phi_2(m_j) := \max_{\sigma_i\in I_i, i=1,\ldots,q} \Phi_2(m_j)$$

(supposed finite)

Choose the model that minimises curvature and MSE simultaneously

"Daniel's plot:" $\big(\Phi_1(m_j), \Phi_2(m_j)\big)$ for $i = 1, \ldots, L$

"Single number criterion:" $\mathrm{argmin}_j \big(a\Phi_1(m_j) + \Phi_2(m_j)\big)$ with $a \in R_{>0}$.

Submodel selection

$$m_*(\underline{x}, \underline{Z}) = f_*(\underline{x}) + g_*(\underline{Z}) + h_*(\underline{x}, \underline{Z})$$

A. Bates et al. (2003) for $g_* = h_* = 0$

$$\text{argmin} \left(\lambda \frac{\rho_m}{\rho_*} + (1-\lambda) \frac{RMSE(m)}{SS_*} \right)$$

over all m hierarchical submodels of m_* and for $\lambda \in [0,1]$ and

$$RMSE(m) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}$$

and

$$\rho_m = \Phi_1(m) = \int_{\chi} \text{trace}(H_f^T H_f)(\underline{x})\, d\underline{x}$$

For

$$\Phi_3(m) = SS_m = \sum_i (\hat{y}_i - \bar{\hat{y}}_i)^2$$

$$\psi(m) = \lambda \frac{\rho_m}{\rho_*} + (1-\lambda)\frac{\Phi_3(m)}{\Phi_3(m_*)}$$

and get submodels with maximal $\psi(m)$

If the computation of an internal minimum is preferred, then

$$\Phi_4(m) = SS_{res} = \sum_i (y_i - \hat{y}_i)^2$$

$$\psi_1(m) = \lambda \frac{\rho_m}{\rho_*} + (1-\lambda)\frac{\Phi_4(m)}{\Phi_3(m_*)}$$

and get submodels with minimal $\psi_1(m)$

Notes

Choice of λ

Plot of the number of terms in the submodel vs. $\psi(m)$

Coloured plot of ρ_m/ρ_* vs. SS_m/SS_* (increasing in ρ_m/ρ_*)

indication of the relationship between model curvature, fitting quality and number of terms

of possible values of λ

of need of substitution of functions of ρ_m/ρ_* and SS_m/SS_* in $\psi(m)$

Extend ideas of Myers et al. to find a design to minimise submodel variance

Exploit polynomial nature of the problem to get explicit representation of mean and variance process and to search in the class of hierarchical submodels

Change distributional assumptions and use finitely generated cumulants to obtain explicit representation of mean and variance process

Case study