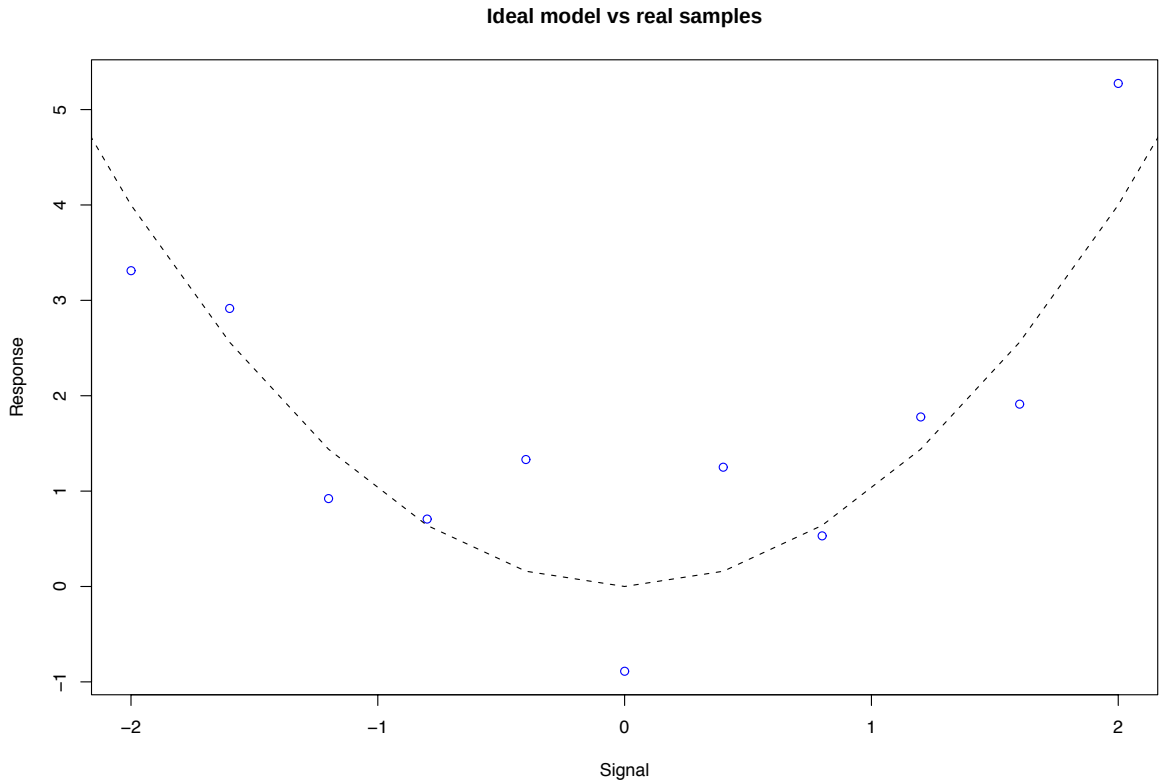
Model Structure Selection: Main Concepts

Sven Laur
swen@math.ut.ee

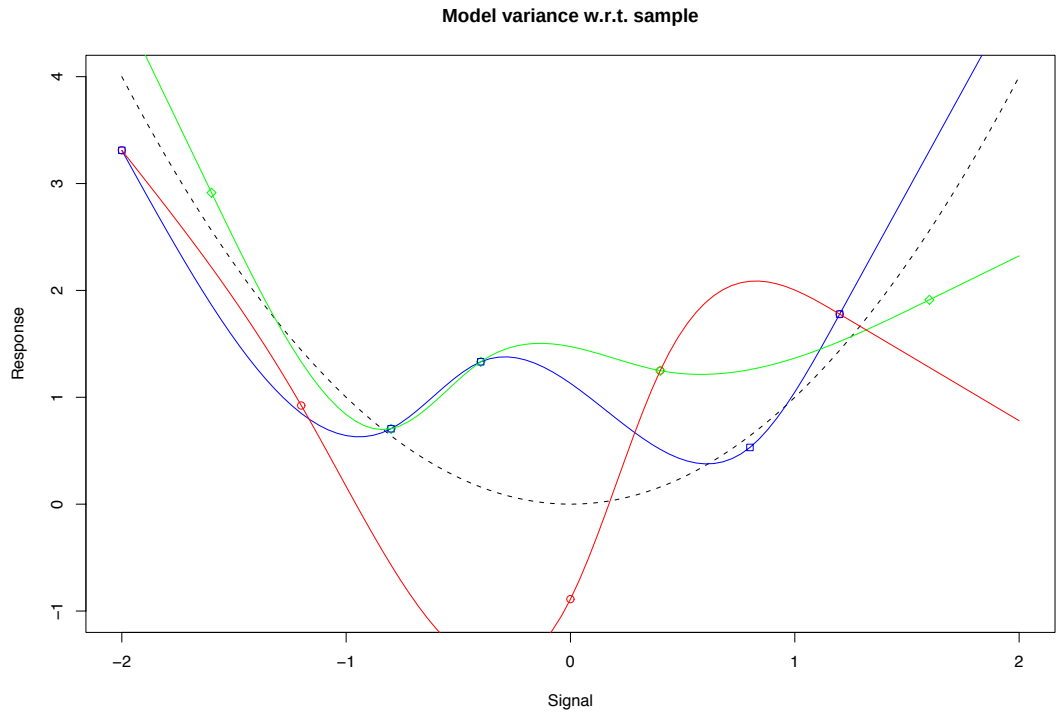
September 28, 2005

Sampling as a stochastic process



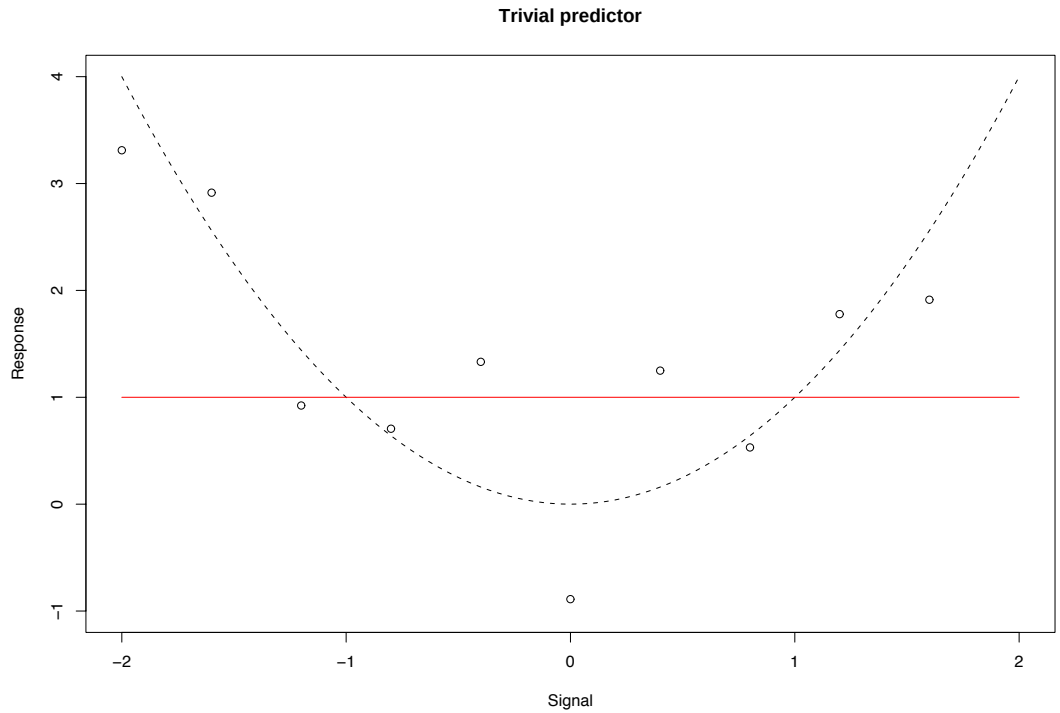
An inherent noise component disturbs the response to the signal.

High variance of complex model structures



Complex models have "many parameters" to adjust and thus the obtained predictor has high variance. Simple models are more tolerant to noise.

Super-stable models



If the choice of model does not depend on a sample, then there is no variation. However, such models have a large bias—an approximation error.

The aim of model order selection

The best "average" performance is achieved when a reasonable compromise between model flexibility and complexity is achieved.

Overfitting happens when the model complexity is too high and statistical fluctuations in a sample have dominant effect on the model parameters.

Regularisation. Adding additional penalty term to the optimisation goal that penalises structures with high complexity stabilises learning method:

- Ridge regression—increasing the weight of main diagonal.
- MDL—adding implicit term that approximates the model variance.
- BIC, AIC, SLT, etc.

There is no universal algorithm to solve bias variance dilemma, as the variance depends on a specific model class and a signal-noise ratio.

Formal specification of learning algorithm

We collect previous input and output values $\boldsymbol{u_{t-1}}, \ldots, \boldsymbol{u_{t-r}}, \boldsymbol{y_{t-1}}, \ldots, \boldsymbol{y_{t-s}}$ and map them into a *feature space* and obtain regression vectors $\boldsymbol{\varphi_i} \in \mathbb{R}^n$.

Assume that feature space is large enough to fit a true signal-response model $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}^d$ such that

$$\boldsymbol{y_i} = f_0(\boldsymbol{\varphi_i}) + \boldsymbol{e_i}$$

where the error term $\boldsymbol{e_i}$ is independent identically distributed (white noise).

Let $g : \mathbb{R}^{n+k} \rightarrow \mathbb{R}^d$ be a parametrised model structure and $\mathcal{S}$ be a sample

$$\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = g_{\mathcal{S}}(\boldsymbol{\varphi_i}) = g(\boldsymbol{\varphi_i}, \widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{S}}})$$

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{S}}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{(\boldsymbol{\varphi_i}, \boldsymbol{y_i}) \in \mathcal{S}} \left\| \boldsymbol{y_i} - g(\boldsymbol{\varphi_i}, \boldsymbol{\theta}) \right\|^2$$

Model quality. Validation error

Validation error is usually defined

$$\bar{V}(f) = \mathbf{E}(\|\boldsymbol{y} - f(\boldsymbol{\varphi})\|^2) \underset{a.s}{=} \lim_N \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{y} - f(\boldsymbol{\varphi_i})\|^2$$

Even the true signal-response model makes errors

$$\bar{V}(f_0) = \lim_N \frac{1}{N} \sum_{i=1}^{N} \|\underbrace{\boldsymbol{y} - f_0(\boldsymbol{\varphi_i})}_{\boldsymbol{e_i}}\|^2 = \mathbf{Var}(\boldsymbol{e}) = \lambda.$$

Other models must have a validation error larger than the noise variance λ.

Components of validation error

For each model structure $g : \mathbb{R}^{n+k} \to \mathbb{R}^d$, there exist an optimal parameter set $\boldsymbol{\theta}_*$ such that the validation error $\bar{V}(g(\cdot, \boldsymbol{\theta}_*))$ is minimal.

We can infer the sample optimum $\boldsymbol{\theta}_{\boldsymbol{\mathcal{S}}}$ and not the true optimum $\boldsymbol{\theta}_*$

$$\mathbf{E}(\bar{V}(g_{\mathcal{S}})) = \mathop{\mathbf{E}}_{\mathcal{S},\boldsymbol{\varphi}}(\|\boldsymbol{y} - g_{\mathcal{S}}(\boldsymbol{\varphi})\|^2) = \mathbf{E}(\|\boldsymbol{y} \pm f_0(\boldsymbol{\varphi}) \pm g(\boldsymbol{\varphi}, \boldsymbol{\theta}_*) + g(\boldsymbol{\varphi}, \boldsymbol{\theta}_{\boldsymbol{\mathcal{S}}})\|^2)$$

$$\lesssim \underbrace{\mathbf{E}(\|\boldsymbol{y} - f_0(\boldsymbol{\varphi})\|^2)}_{noise\ \lambda} + \underbrace{\mathbf{E}(\|f_0(\boldsymbol{\varphi}) - g(\boldsymbol{\varphi}, \boldsymbol{\theta}_*)\|^2)}_{bias} + \underbrace{\mathbf{E}(\|g(\boldsymbol{\varphi}, \boldsymbol{\theta}_*) - g(\boldsymbol{\varphi}, \boldsymbol{\theta}_{\boldsymbol{\mathcal{S}}})\|^2)}_{variance}$$

- Noise term is unavoidable unless we do not add additional inputs.
- Bias term is essentially an approximation error.
- Variance term is caused by statistical fluctuations.

Asymptotic behaviour

Regression method is "consistent" if $\boldsymbol{\theta}_{\boldsymbol{\mathcal{S}}} \xrightarrow[a.s.]{} \boldsymbol{\theta}_*$ as $|\mathcal{S}| \to \infty$. For over-parametrised models $\mathbf{E}(\|g(\boldsymbol{\varphi}, \boldsymbol{\theta}_*) - g(\boldsymbol{\varphi}, \boldsymbol{\theta}_{\boldsymbol{\mathcal{S}}})\|^2) \xrightarrow[a.s.]{} 0$ is more appropriate.

In any case

$$\mathbf{E}(\bar{V}(g_{\mathcal{S}})) \xrightarrow[a.s.]{} \underbrace{\mathbf{E}(\|\boldsymbol{y} - f_0(\boldsymbol{\varphi})\|^2)}_{noise\ \lambda} + \underbrace{\mathbf{E}(\|f_0(\boldsymbol{\varphi}) - g(\boldsymbol{\varphi}, \boldsymbol{\theta}_*)\|^2)}_{bias}$$

For a sufficiently smooth model class, one can approximate

$$\underbrace{\mathbf{E}(\|g(\boldsymbol{\varphi}, \boldsymbol{\theta}_*) - g(\boldsymbol{\varphi}, \boldsymbol{\theta}_{\boldsymbol{\mathcal{S}}})\|^2)}_{variance} \approx c \cdot \frac{m\lambda}{N}$$

where $c > 0$ is a fixed constant and m is "the number of parameters" in the model, i.e. the variance error has order $\mathcal{O}(1/N)$.

Penalised regression

Note that the first term decreases and the second term increases

$$\mathbf{E}(\bar{V}(g_{\mathcal{S}})) \approx \mathbf{E}(\|\boldsymbol{y} - g(\boldsymbol{\varphi}, \boldsymbol{\theta}_*)\|^2) + c \cdot \frac{m\lambda}{N}$$

if we use more complex models.

One can approximate the effect of variance by adding a penalty term

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{S}}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{(\boldsymbol{\varphi_i}, \boldsymbol{y_i}) \in \mathcal{S}} \|\boldsymbol{y_i} - g(\boldsymbol{\varphi_i}, \boldsymbol{\theta})\|^2 + \delta \|\boldsymbol{\theta}\|^2$$

where $\delta \propto c \cdot \lambda$. Unfortunately, the exact ratio between δ and $c \cdot \lambda$ cannot be derived from theory.

For a linear models, the penalty term also increases the computational stability, as it increases the weight of the main diagonal of system matrix.

Computational stability of predictions

Fix any plausible input signal $\boldsymbol{u}(t),\ t > 0$. Then any prediction rule

$$\boldsymbol{y(t)} = F(\boldsymbol{u(t-1)}, \ldots, \boldsymbol{u(t-r)}, \boldsymbol{y(t-1)}, \ldots, \boldsymbol{y(t-s)})$$

defines a path $X(t)$ of corresponding tuples $(\boldsymbol{y(t)}, \ldots, \boldsymbol{y(t-s+1)})$.

Iterative prediction method is stable, if two paths starting from two close points $X_0(t_0)$ and $X_1(t_0)$ do not diverge

$$\forall\varepsilon\forall t_0\ \exists\delta : \|X_0(t_0) - X_1(t_0)\| \leq \delta \Rightarrow \|X_0(t) - X_1(t)\| \leq \varepsilon,\ t > t_0.$$

We can consider also uniform, asymptotic, exponential and BIBO stability. In a way they describe more the long-term behaviour of a model than statistical properties of short term predictions.

Why is computational stability important?

- Obviously, unstable predictors are not suitable for long-term prediction.
- If controller "uses" unstable model then it cannot "look" far ahead and we get probably sub-optimal controller strategy.
- Computationally stable models allow to make more accurate simulations of the modelled system provided that the models are precise enough.
- Stable systems, especially BIBO stable systems, are more easy to control, since bounded input difference implies bounded output difference.

Optimal size of regressor. General idea

Assume that you have mapped data points to regression vectors $\boldsymbol{\varphi_i}$.

More coordinates than necessary. If the "true" model uses only a subset $\boldsymbol{\psi}$ of coordinates of $\boldsymbol{\varphi}$ and the "true" model is Lipschitz continuous then the Lipschitz quotients

$$q_{ij} = \frac{\|\boldsymbol{y_i} - \boldsymbol{y_j}\|}{\|\boldsymbol{\varphi_i} - \boldsymbol{\varphi_j}\|} = \underbrace{\frac{\|\boldsymbol{\psi_i} - \boldsymbol{\psi_j}\|}{\|\boldsymbol{\varphi_i} - \boldsymbol{\varphi_j}\|}}_{small} \cdot \underbrace{\frac{\|\boldsymbol{y_i} - \boldsymbol{y_j}\|}{\|\boldsymbol{\psi_i} - \boldsymbol{\psi_j}\|}}_{bounded}$$

are relatively small. Obviously, with high probability $\|\boldsymbol{\psi}\| \ll \|\boldsymbol{\varphi}\|$ since there are other nonzero coordinates in redundant directions.

Optimal size of regressor. General idea

Missing coordinates. If the "true" model uses coordinates not represented in feature space then additional noise emerges. Moreover $\boldsymbol{\varphi} \longrightarrow \boldsymbol{\varphi_i}$ then $\boldsymbol{y} \not\to \boldsymbol{y_i}$ and thus with high probability our data sample contains pairs with high Lipschitz quotients.

General idea. To determine the optimal size of regressor, we have to somehow compare appropriate aggregate statistics of Lipschitz quotients.

Bonus. As the method does not explicitly use the parametrisation of the feature space it is independent of parametrisation, i.e. we do not have to care how the coordinate axis are placed.

Heuristic choice of lag space

1. Fix a candidate for a lag space, i.e. a collection of (delayed) signals

$$\boldsymbol{y(t-1)}, \ldots, \boldsymbol{y(t-n)}, \boldsymbol{u(t-d)}, \ldots, \boldsymbol{u(t-d-m)}$$

   and compute Lipschitz quotients q_{ij}.

2. Choose $p = 0.01N \sim 0.02N$ and select p largest quotients q_k and compute

$$\bar{q}^{(n)} = \left(\prod_{k=1}^{p} \sqrt{n} q_k \right)^{1/p}$$

3. Repeat the procedure for different lag structures (usually $n = m$).

4. Make the corresponding plot and choose the "knee-point"— a regime shift.