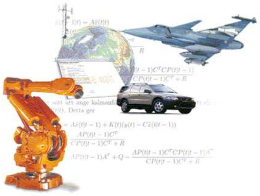
Identification of Linear and Nonlinear Dynamical Systems

Theme 1: Curve Fitting

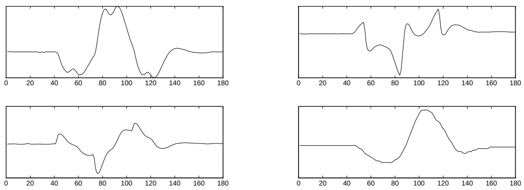


Lennart Ljung

Division of Automatic Control
Linköping University
Sweden

The Problem

Data from Gripen



Pitch rate, Canard,
Elevator, Leading Edge Flap

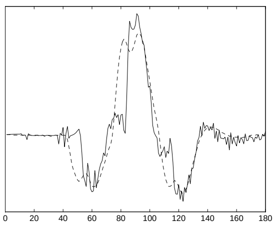
Questions

- How do the control surface angles affect the pitch rate?
- Aerodynamical derivatives?
- How to use the information in flight data?

Aircraft Dynamics: From input 1

$y(t)$ pitch rate at time t. $u_1(t)$ canard angle at time t. Try

$$y(t) = -a_1 y(t-T) - a_2 y(t-2T) - a_3 y(t-3T) - a_4 y(t-4T)$$
$$+ b_1 u_1(t-T) + b_2 u_1(t-2T) + b_3 u_1(t-3T) + b_4 u_1(t-4T) \quad T = 1/60s.$$

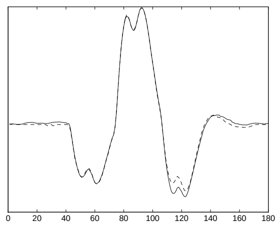


Dashed line: Actual Pitch rate. Solid line: 10 step ahead predicted pitch rate, based on the fourth order model from canard angle only.

Using all inputs

u_1 canard angle; u_2 Elevator angle; u_3 Leading edge flap;

$$y(t) = -a_1 y(t-T) - a_2 y(t-2T) - a_3 y(t-3T) - a_4 y(t-4T) + b_1^1 u_1(t-T) + \ldots$$
$$+ b_4^1 u_1(t-4T) + \ldots + b_1^3 u_3(t-T) + \ldots + b_4^3 u_3(t-4T) \quad T = 1/60s.$$



System Identification: Issues

- Select a class of candidate models
- Select a member in this class using the observed data
- Evaluate the quality of the obtained model
- Design the experiment so that the model will be “good”.

Course Outline: Themes

1. The basic questions and (statistical) tools illustrated for a simple curve fitting problem.
2. Linear models: The model structures, Special techniques for linear models. Time and frequency domain data.
3. Software session with hands-on experience
4. Nonlinear models: Parameterizations, problems and techniques.
5. Some practical issues in system identification: Experiment design and data quality.

Curve Fitting - Outline

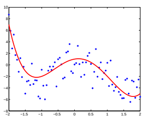
- Corrupted observations of function values
- Model function parameterizations
- Least squares fit and variants
- Example of fit depending on model size
- Statistical asymptotic analysis of parametric methods
- Bias - Variance trade off
- Nonparametric methods

Background Terms

- Random (stochastic) variable, Expectation, Variance
- Independent random variables, random processes
- Law of Large Numbers, Central Limit Theorem

Curve Fitting

Most basic ideas from system identification, choice of model structures and model sizes are brought out by considering the basic curve fitting problem from elementary statistics.



Unknown function $g_0(x)$. For a sequence of x-values (regressors) $\{x_1, x_2, \ldots, x_N\}$ (that may or may not be chosen by the user) observe the corresponding function values with some noise:

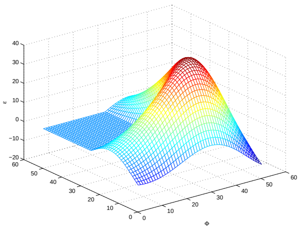
$$y(k) = g_0(x_k) + e(k)$$

Construct an estimate $\hat{g}_N(x)$ from $\{y(1), x_1, y(2), x_2, \ldots, y(N), x_N\}$

.

Curve Fitting II: Several Regressors

"Surface fitting":



- The floor is formed by the regressors x, and the upright wall is the function value $y = g_0(x)$.

The Curve-fitting problem

$$y(k) = g_0(x_k) + e(k)$$

Construct an estimate $\hat{g}_N(x)$ from $\{y(1), x_1, y(2), x_2, \ldots, y(N), x_N\}$
The error $\hat{g}_N(x) - g_0(x)$ should be "as small as possible"
Approaches:

- Parametric: Construct $\hat{g}_N(x)$ by searching over a parameterized set of candidate functions.
- Non-parametric: Construct $\hat{g}_N(x)$ by smoothing over (carefully chosen subsets of) $y(k)$

Curve Fitting - Outline

- Corrupted observations of function values
- Model function parameterizations
- Least squares fit and variants
- Example of fit depending on model size
- Statistical asymptotic analysis of parametric methods
- Bias - Variance trade off
- Nonparametric methods

Parametric Approach

Search for the function g_0 in a parameterized family of functions:

$$g(x, \theta) = \sum_{k=1}^{n} \alpha_k f_k(x, \tilde{\theta}_k), \quad \theta = \{\alpha_k, \tilde{\theta}_k,\ k = 1, \ldots, n\}$$

- Grey box/Black box
- Local/Global basis functions

Examples: $g(x, \theta) = \theta_1 + \theta_2 x + \ldots + \theta_n x^{n-1}$

$$g(x, \theta) = \frac{\theta_1 + \theta_2 x + \ldots + \theta_n x^{n-1}}{1 + \theta_{n+1} x + \ldots + \theta_{n+m-1} x^{m-1}}$$

$$g(x, \theta) = \theta_0 + \sum_{k=1}^{n} \theta_{2k-1} \cos(k\pi x) + \theta_{2k} \sin(k\pi x)$$

$$g(x, \theta) = \sum_{k=1}^{n} \alpha_k U((x - \gamma_k)/\beta_k), \quad U(x) \text{ unit pulse}$$

The Choices in the Parametric case

Type of function family (Basis functions $f_k(x, \theta)$)

Size of model (n or dim θ)

The parameter values

The Choices in the Parametric case

3 Type of function family (Basis functions $f_k(x, \theta)$)
2 Size of model (n or dim θ)
1 The parameter values

Curve Fitting - Outline

- Corrupted observations of function values
- Model function parameterizations
- Least squares fit and variants
- Example of fit depending on model size
- Statistical asymptotic analysis of parametric methods
- Bias - Variance trade off
- Nonparametric methods

Parametric Curve Fitting: Choice of parameters

$$y(t) = g_0(x_t) + e(t)$$

Least Squares:

$$\hat{\theta}_N = \arg\min_\theta V_N(\theta)$$

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} |y(t) - g(x_t, \theta)|^2$$

.

Parametric Curve Fitting: Choice of parameters

$$y(t) = g_0(x_t) + e(t)$$

Weighted Least Squares:

$$\hat{\theta}_N = \arg\min_\theta V_N(\theta)$$

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} |y(t) - g(x_t, \theta)|^2/\lambda_t$$

λ_t Proportional to 'reliability' of t:th measurement $\sim Ee^2(t)$

Parametric Curve Fitting: Choice of parameters

$$y(t) = g_0(x_t) + e(t)$$

Weighted Least Squares:

$$\hat{\theta}_N = \arg\min_\theta V_N(\theta)$$

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} L(x_t)|y(t) - g(x_t, \theta)|^2/\lambda_t$$

λ_t Proportional to 'reliability' of t:th measurement $\sim Ee^2(t)$

A extra weighting $L(x_t)$ could also reflect the 'relevance' of the point x_t.

('Focus in fit')

Parametric Curve Fitting: Choice of parameters

$$y(t) = g_0(x_t) + e(t)$$

(Regularized) Least squares:

$$\hat{\theta}_N = \arg\min_\theta V_N(\theta) + \delta|\theta|^2$$

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} |y(t) - g(x_t, \theta)|^2$$

$\delta|\theta|^2$ penalizes excessive model flexibility. Could come in various forms.

Why the Least Squares Criterion?

- Gauss!
- Maximum Likelihood:

$$\hat{\theta}_N = \arg\min_\theta \frac{1}{N}\sum_{t=1}^{N} \ell(y(t) - g(x_t, \theta), t)$$

$\ell(z, t) = -\log p(z, t)$, $p(z, t)$ is the probability density function (pdf) of $e(t)$

Gaussian distribution $p(z, t) \sim e^{-z^2/2\lambda_t}$ gives a quadratic criterion!

- Other choices
  - $\min_\theta \max_t |y(t) - g(x_t, \theta)|$ ("unknown-but-bounded")
  - $\min \sum |y(t) - g(x_t, \theta)|_\epsilon$ (ϵ-insensitive ℓ_1 norm, "Support vector machines")

Linear Least squares

Note that if the parameterization $g(x, \theta)$ is linear in θ, the basic criterion becomes quadratic in θ, and the minimum can be found analytically:

$$g(x, \theta) = \varphi(x)^T\theta$$

$$V_N(\theta) = \sum (y(t) - \varphi(x_t)^T\theta)^2 = \|Y - \Phi\theta\|^2$$

$$Y = \text{col } y(t), \Phi = \text{col } \varphi(x_t)^T$$

$$\hat{\theta}_N = (\Phi^T\Phi)^{-1}\Phi^T Y$$

Choices in Parametric Methods

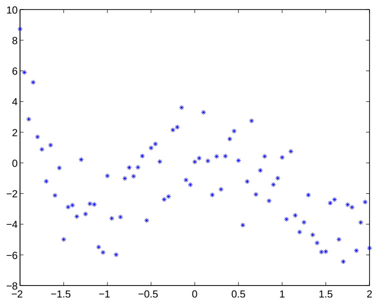
So, the choice of parameters within a parameterized model is not that difficult: Fit to the observed data, by one criterion or another.
The choice of model size and model parameterization is a more interesting issue.

Curve Fitting - Outline

- Corrupted observations of function values
- Model function parameterizations
- Least squares fit and variants
- Example of fit depending on model size
- Statistical asymptotic analysis of parametric methods
- Bias - Variance trade off
- Nonparametric methods

Choice of Model Size: Example

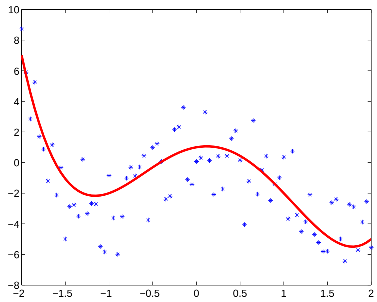
Observed data



Fit polynomials of different orders.

Choice of Model Size: Example

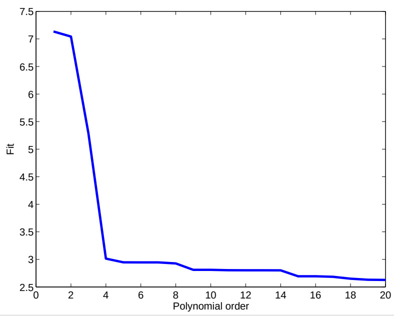
Example: Observed data with true curve



Fit polynomials of different orders.

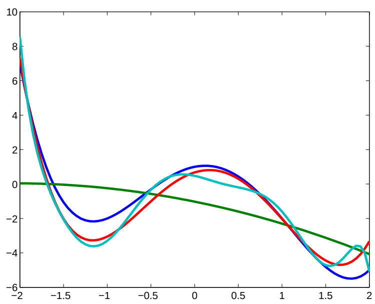
The Model Fit

The value of the criterion as a function of polynomial order.



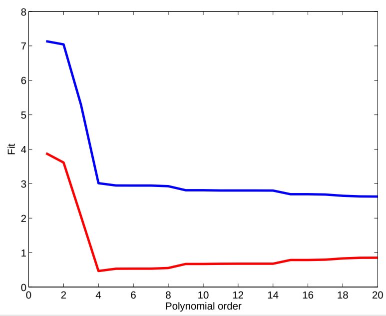
Model Curves

Blue: True curve. Green: 2nd order. Red: 4th order. Cyan: 10th order.



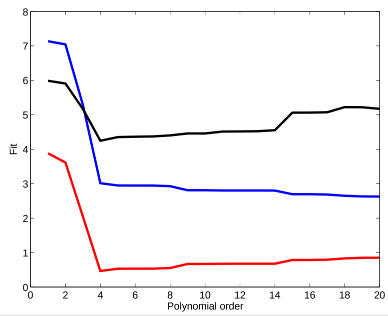
The Model Fit

The value of the criterion as a function of polynomial order. The fit between the true curve and the model curve.



The Model Fit

The value of the criterion as a function of polynomial order. The fit between the true curve and the model curve. The value of the criterion evaluated on a fresh data set.



Choice of Size and Structure

This is a more difficult choice, and we need to understand how the model error $\hat{g}_N(x) - g_0(x)$ depends on our choices.
Players:

- The fit for a certain data set Z: $V_N(\theta, Z) = \frac{1}{N}\sum(y(t) - g(x_t, \theta))^2$
- Estimation (training) data Z_e. Validation (generalization) data Z_v.
- The empirical fit: $V_N(\hat{\theta}_N, Z_e) = \min_\theta V_N(\theta, Z_e)$ (blue curve)
- The validation fit $V_N(\hat{\theta}_N, Z_v)$ (black curve)
- The curve fit $H(x, \theta) = |g_0(x) - g(x, \theta)|^2$
  - For given x_t-sequence $H_N(\theta) = \frac{1}{N}\sum H(x_t, \theta)$. $H_N(\hat{\theta}_N)$ was the red curve.
- The expected (typical) value of $H_N(\hat{\theta}_N)$ would be a suitable goodness measure for the chosen parameterization.

Testing the Model Parameterizations

- Test by simulation: Monte-Carlo.
  - Do not get fooled by the empirical fit $V_N(\hat{\theta}_N, Z_e)$
  - Need to understand how the empirical fit relates to $H_N(\hat{\theta}_N)$
- Compute by calculations: Analysis
  - "Analytical Monte-Carlo": Assume certain properties of x_k and $e(k)$, the compute (if possible) the error.

Curve Fitting - Outline

- Corrupted observations of function values
- Model function parameterizations
- Least squares fit and variants
- Example of fit depending on model size
- Statistical asymptotic analysis of parametric methods
- Bias - Variance trade off
- Nonparametric methods

Basic Tools for Analysis

For a stationary stochastic process $e(\cdot)$ under mild conditions

- Law of large numbers (LLN)
  - $\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^N e(t) = Ee(t)$
- Central limit theorem (CLT)
  If $e(t)$ has zero mean:
  - $\frac{1}{\sqrt{N}}\sum_{t=1}^N e(t)$ converges in distribution to the normal (Gaussian) distribution with zero mean and variance $\overline{\lambda} = \lim \frac{1}{N}\sum_{t,s=1}^N Ee(t)e(s)$.
    "$\frac{1}{\sqrt{N}}\sum_{t=1}^N e(t) \to N(0, \overline{\lambda})$"

Asymptotic Analysis: Probabilistic Setup

"Analytical Monte-Carlo Experiment": For a given $g_0(\cdot)$ and a given sequence x_t collect the data

$$y(t) = g_0(x_t) + e(t), \quad Ee(t)^2 = \lambda$$

where the stochastic process $e(\cdot)$ obeys the LLN and CLT and has variance λ. Use a parameterization $g(x, \theta)$. Form the estimate

$$\hat{\theta}_N = \arg\min \frac{1}{N}\sum_{t=1}^N (y(t) - g(x_t, \theta))^2$$

$$\hat{g}_N(x) = g(x, \hat{\theta}_N)$$

Then $\hat{\theta}_N$ and $\hat{g}_N(x)$ are random variables with properties inherited from e. What can be said about their distributions?

Asymptotic Analysis: Basic Facts – BIAS

Except for very simple parameterizations $g(x, \theta)$, the distribution of $\hat{\theta}_N$ cannot be calculated (mainly due to "arg min").
However its asymptotic distribution as $N \to \infty$ can be established:

- $\overline{E}$ = averaging over x_k: $\overline{E}f(x) = \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^N f(x_k)$
- $H(\theta) = \lim_{N\to\infty} H_N(\theta) = \overline{E}|g_0(x_t) - g(x_t, \theta)|^2$
- Best possible model in parameterization: $\theta^* = \arg\min H(\theta)$
- If $H(\theta^*) = 0$ we have a perfect curve fit, otherwise there be some bias in the curve fit.
- Main Result: $\lim_{N\to\infty} \hat{\theta}_N = \theta^*$

Proof: Formal Calculations

$$V_N(\theta) = \frac{1}{N}\sum(g_0(x_t) + e(t) - g(x_t, \theta))^2$$
$$= \frac{1}{N}\sum(g_0(x_t) - g(x_t, \theta))^2 + \frac{1}{N}\sum e^2(t) + \frac{2}{N}\sum(g_0(x_t) - g(x_t, \theta))e(t)$$

LLN: $\frac{1}{N}\sum(g_0(x_t) - g(x_t, \theta))e(t) \to 0$ (uniformly in θ!)

so $V_N(\theta) \to H(\theta)$ as $N \to \infty$

Asymptotic Analysis: Basic Facts – VARIANCE

Suppose the limit model is correct: $g(x, \theta^*) \approx g_0(x)$ and e white noise with variance λ:

- The asymptotic distribution of $\sqrt{N}(\hat{\theta}_N - \theta^*)$ is normal with zero mean and covariance matrix $P = \lambda[\overline{E}\psi(t)\psi^T(t)]^{-1}, \quad \psi(t) = \frac{d}{d\theta}g(x_t, \theta^*)$
- "Cov $\hat{\theta}_N \sim \frac{\lambda}{N}[\overline{E}\psi(t)\psi^T(t)]^{-1}$"

Proof: Formal Calculations

$$0 = V_N'(\hat\theta_N) = V_N'(\theta^*) + V_N''(\theta^*)(\hat\theta_N - \theta^*)$$
$$(\hat\theta_N - \theta^*) = -[V_N''(\theta^*)]^{-1} V_N'(\theta^*)$$
$$V_N'(\theta) = \frac{2}{N}\sum (y(t) - g(x_t,\theta))g'(x_t,\theta)$$
$$V_N'(\theta^*) = \frac{2}{N}\sum e(t)\psi(t)$$
LLN: $V_N''(\theta^*) = \frac{2}{N}\sum \psi(t)\psi^T(t) + \frac{2}{N}\sum e(t)g''(x_t,\theta^*) \to 2\overline{E}\psi\psi^T$

CLT: $\frac{1}{\sqrt{N}}\sum e(t)\psi(t) \to N(0, \lambda\overline{E}\psi\psi^T)$

$$\sqrt{N}(\hat\theta_N - \theta^*) \to N(0, \lambda[\overline{E}\psi\psi^T]^{-1})$$

A Remarkable Observation

Recall the curve fit $H(x,\theta) = |g_0(x) - g(x,\theta)|^2$, $H(\theta) = \lim_{N\to\infty} \frac{1}{N}\sum H(x_t,\theta)$
(For the x-sequence of the estimation data.)
$H(\hat\theta_N)$ is a random variable, since the estimate depends on the e-sequence, and

$$EH(\hat\theta_N) = H(\theta^*) + \lambda\frac{d}{N}$$

where d is the number of estimated parameters independently of the parameterization!
(Proof: ....)

Proof: Formal Calculations

$g_0(x) = g(x,\theta^*)$ (assumption)

$$H(\hat\theta_N) = H(\theta^*) + H'(\theta^*)(\hat\theta_N - \theta^*) + \frac{1}{2}(\hat\theta_N - \theta^*)^T H''(\theta^*)(\hat\theta_N - \theta^*)$$

$H'(\theta^*) = 0$ (θ^* minimizes $H(\theta)$)

$$H'(\theta) = \frac{2}{N}\sum (g_0(x_t) - g(x_t,\theta))g'(x_t,\theta)^T$$
$$H''(\theta^*) = \frac{2}{N}\sum g'(x_t,\theta^*)g'(x_t,\theta^*)^T = 2\overline{E}\psi(t)\psi^T(t)$$
$$EH(\hat\theta_N) = H(\theta^*) + E\frac{1}{2}(\hat\theta_N - \theta^*)^T H''(\theta^*)(\hat\theta_N - \theta^*)$$
$$E\text{tr}\,[\frac{1}{2}(\hat\theta_N - \theta^*)^T H''(\theta^*)(\hat\theta_N - \theta^*)] = E\text{tr}\,[\frac{1}{2}H''(\theta^*)(\hat\theta_N - \theta^*)(\hat\theta_N - \theta^*)^T]$$
$$= \text{tr}\,\frac{1}{2}H''(\theta^*)\text{Cov}\,\hat\theta_N = \frac{\lambda}{N}\text{tr}\,\left[(\overline{E}\psi\psi^T)(\overline{E}\psi\psi^T)^{-1}\right] = d\frac{\lambda}{N}$$

The Regularized Case

The variance is reduced by regularization, at the price of some bias.
In the previous result, the number of parameters d is replaced by d_{eff}:
Effective dimension of $\theta \approx$ Number of eigenvalues of the Hessian of $\bar{V}$ that are larger than δ (the regularization parameter). Note: $d_{eff} \le d = \dim\theta$

Curve Fitting - Outline

- Corrupted observations of function values
- Model function parameterizations
- Least squares fit and variants
- Example of fit depending on model size
- Statistical asymptotic analysis of parametric methods
- Bias - Variance trade off
- Nonparametric methods

Choice of Size: Basic Trade-off

- $H(\theta) = \lim \frac{1}{N}\sum_t |g_0(x_t) - g(x_t,\theta)|^2$,
  $EH(\hat\theta_N) \approx H(\theta^*) + \frac{\lambda}{N}d$
- A good model size is one that minimizes this expression
- $H(\theta^*)$ is the best possible fit that can be achieved within the parameterization. A smaller value of this means less bias. Thus, more parameters gives a more flexible model parameterization and hence less bias.
- More parameters lead however to higher variance.
- The model size is thus a bias – variance trade-off.
- Note that this balance is usually reached with a non-zero $H(\theta^*)$, that is, it is normal to accept bias. Also a larger size model can be used when more data are available (larger N).
- If a regularized criterion is used, the size of the regularization parameter δ can also be used to control the flexibility of the parametrization.

Choice of Type (basis functions)

Generally speaking, the parameterization should be such that useful flexibility is achieved with as few parameters as possible:
$\Rightarrow$
Grey box models

- Tunable or Non-tunable Basis functions:
  $g(x,\theta) = \sum_{k=1}^n \alpha_k f_k(x,\theta)$
  - $+$ More flexible structure $=$ Less parameters
  - $-$ More work to minimize (non-tunable = Linear Least Squares)
- Use (number of parameters) d or (regularization parameter) δ as a size-tuning knob
  - When no natural ordering of structures: Easier to use δ.

Curve Fitting - Outline

- Corrupted observations of function values
- Model function parameterizations
- Least squares fit and variants
- Example of fit depending on model size
- Statistical asymptotic analysis of parametric methods
- Bias - Variance trade off
- Nonparametric methods

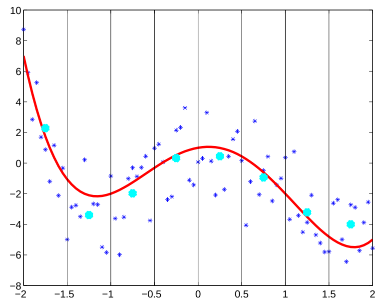
A simple idea is to locally smooth the noisy observations of the function values:

$$\hat{g}_N(x) = \sum_{k=1}^{N} C(x, x_k) y(k)$$

$$\sum_{k=1}^{N} C(x, x_k) = 1 \ \forall x$$

Often $C(x, x_k) = \tilde{c}(x - x_k)/\lambda_k$ and $\tilde{c}(r) = 0$ for $|r| > \beta$, $\beta =$ the "bandwidth"

These are known as "kernel methods" in statistics.
If $C(x, x_t)$ is chosen so that it is non-zero ($= 1/k$) only for k observed values x_t around x, this is the k-nearest neighbor method.



$C(x, x_k) = U((x - x_k)/\beta)$; $U(\cdot)$ the unit pulse. β = 0.25.

Cyan dots: Computed for $x = -1.75 : 0.5 : 1.75$

$$\varepsilon_N(x) = \hat{g}_N(x) - g_0(x) = \sum_{k=1}^{N} C(x, x_k) y(k) - g_0(x) =$$

$$\sum_{k=1}^{N} C(x, x_k)(e(k) + [g_0(x_k) - g_0(x)])$$

$$E\varepsilon_N(x) = \sum_{k=1}^{N} C(x, x_k)[g_0(x_k) - g_0(x)]$$

Var $\varepsilon_N(x) = \sum_{k=1}^{N} C^2(x, x_k)\lambda$ (for white e with variance λ)

Think of $C(x, x_k) = U((x - x_k)/\beta)$ where U is the unit pulse:

$$C(x, x_k) = \begin{cases} \frac{1}{N_k} & \text{if } |x - x_k| \le \beta \\ 0 & \text{else} \end{cases}$$

$N_k =$ number of x_k in the bin $|x - x_k| \le \beta$

MSE: $H(x) = \sum_{k=1}^{N} C^2(x, x_k)\lambda + \left[\sum_{k=1}^{N} C(x, x_k)[g_0(x_k) - g_0(x)]\right]^2 \approx$

$\frac{1}{N_k}\lambda +$ variation of $g_0(x)$ over $|x - x_k| \le \beta$

Trade-off: Want β to be small for small bias. Want β to be large for small variance. The best choice depends on the nature of g_0.

Consider the parametric method using unit pulses $U(x)$:

$$g(x, \theta) = \sum_{k=1}^{n} \theta_k U((x - \gamma_k)/\beta) \quad \beta \text{ and } \gamma_k \text{ given } \gamma_k - \gamma_{k-1} = \beta$$

$$\sum_{t=1}^{N} (y(t) - g(x_t, \theta))^2 = \sum_{k=1}^{n} \sum_{t:|x_t - \gamma_k| < \beta} (y(t) - g(x_t, \theta))^2 =$$

$$\sum_{k=1}^{n} \sum_{t:|x_t - \gamma_k| < \beta} (y(t) - \theta_k)^2 \Rightarrow \hat{\theta}_k = \frac{1}{N_k} \sum_{t:|x_t - \gamma_k| < \beta} y(t)$$

This means that $\hat{g}(\gamma_k) = \hat{\theta}_k$.
If we use a nonparametric method to estimate g at $x = \gamma_k$ with $C(x, x_k) = \frac{1}{N_k} U((x - x_k)/\beta)$ we obtain the same estimate.

- We have used the simple case of curve-fitting to illustrate basic issues, frameworks and techniques for linear and nonlinear system identification
- Parametric – Nonparametric methods
- Choice of model parametrization, model size and parameter values.
- Parameter values easy: Some version of least squares fit.
- Basic asymptotic properties: $\hat{\theta}_N \to \theta^*$, best possible approximation available in the parameterization (for the used x_t-sequence)
- $\sqrt{N}(\hat{\theta}_N - \theta^*) \sim N(0, P)$, $P = \lambda[E\psi(t)\psi^T(t)]^{-1}$ (Normal distribution)
- Choice of parametric model structure guided by bias-variance trade off (number of parameters)
- Choice of nonparametric method guided by bias-variance trade off (band-with of the kernel)