Validation

Rami Rautkorpi
rami.rautkorpi@hut.fi
12.10.2005

Overview

- How to validate?
- Correlations in the residuals
- Average generalization error
- Visualization of the predictions

How to validate?

- Testing the model in practice
  - *Is difficult*
- Simple tests for particular properties of the model
- Can we say this model is *good*... or just *better than the other one*?
- A *test* or *validation* set of data is needed

Correlations in the residuals 1

- Is the model capable of extracting all information from the training data?
- Correlations between *the residuals* and *all linear and nonlinear combinations of past data*
- Sampled correlation functions, Billings et al.
  - Converge to a Gaussian distribution with zero mean and variance 1/N *if the true system has been identified*

Correlations in the residuals 2

$$\hat{r}_{\varepsilon\varepsilon}(\tau)=\frac{\sum_{t=1}^{N-\tau}(\varepsilon(t,\hat{\vartheta})-\bar{\varepsilon})(\varepsilon(t-\tau,\hat{\vartheta})-\bar{\varepsilon})}{\sum_{t=1}^{N}(\varepsilon(t,\hat{\vartheta})-\bar{\varepsilon})^2}=\begin{cases}1, \tau=0\\0, \tau\neq 0\end{cases}$$

$$\hat{r}_{u\varepsilon}(\tau)=\frac{\sum_{t=1}^{N-\tau}(u(t)-\bar{u})(\varepsilon(t-\tau,\hat{\vartheta})-\bar{\varepsilon})}{\sqrt{\sum_{t=1}^{N}(u(t)-\bar{u})^2}\sum_{t=1}^{N}(\varepsilon(t,\hat{\vartheta})-\bar{\varepsilon})^2}=0, \forall\tau$$

- Etc.
- Check that functions are zero within 95% confidence interval $(-1.96/\sqrt{N}< \ \hat{r} \ <1.96/\sqrt{N})$

Average generalization error

- Techniques for estimating the average generalization error
  - Akaike's Final Prediction Error (FPE)
  - Linear-Unlearning-Leave-One-Out (LULOO)
- Useful for validation
- Primarily for model structure selection

Akaike's Final Prediction Error 1

$$\hat{V}_M = \frac{1}{2}\sigma_e^2 (1 + \frac{p}{N}) \qquad \sigma_e^2 = 2\frac{N}{N-p} V_N(\hat{\theta}, Z^N)$$

$$\hat{V}_M = \frac{N+p}{N-p} V_N(\hat{\theta}, Z^N)$$

$$\hat{V}_M = \frac{1}{2}\left[\sigma_e^2 (1 + \frac{p_1}{N}) + \gamma\right] \qquad p_1 = \boldsymbol{tr}\left\{R[R+D]^{-1} \; R[R+D]^{-1}\right\}$$

$$\gamma = \frac{1}{N^2}\theta_0^T D \left[R + \frac{1}{N} D\right]^{-1} \left[R + \frac{1}{N} D\right]^{-1} D \theta_0 \qquad D = \alpha I$$

Akaike's Final Prediction Error 2

$$p_2 = \boldsymbol{tr}\left\{ R\left(R + \frac{1}{N} D \right)^{-1} \right\} = \sum_{i=1}^{p} \frac{\delta_i}{\delta_i + \frac{\alpha}{N}} \simeq p_1$$

$$\hat{V}_M = \frac{N + p_1}{N + p_1 - 2 p_2} V_N(\hat{\theta}, Z^N) \;\simeq\; \frac{N + p_1}{N - p_1} V_N(\hat{\theta}, Z^N)$$

- γ has been discarded
  - ➔ FPE is too small
- Derivation assumes that *the true system is contained in the model structure*

Leave-One-Out

- Training with the entire data set *except for one input-output pair* $\{\varphi(t), y(t)\}$
- Prediction error is evaluated for each t

$$\hat{V}_M = \frac{1}{2N} \sum_{t=1}^{N} \left[y(t) - \hat{y}(t|\hat{\theta}_t) \right]^2 = \frac{1}{2N} \sum_{t=1}^{N} \varepsilon^2(t, \hat{\theta}_t)$$

- N networks are trained to their minimum
  - Realistic only for small data sets
  - Shortcut: Minimum from the entire data set is used as a starting point

Linear-Unlearning-Leave-One-Out 1

- The reduced data set without input-output pair number t: $Z_t^{N-1} = Z^N \setminus \{\varphi(t), y(t)\}$
- Average generalization error estimate is derived from the regularized criterion:

$$W_{N-1}(\theta, Z_t^{N-1}) = W_N(\theta, Z^N) - \frac{1}{2N}[y(t) - \hat{y}(t|\theta)]^2$$

$$H_t = W''_{N-1}(\hat{\theta}, Z_t^{N-1})$$

$$\hat{V}_M = \frac{1}{2N} \sum_{t=1}^{N} \varepsilon^2(t, \hat{\theta}) \left(1 + \frac{2}{N} \psi^T(t, \hat{\theta}) H_t^{-1} \psi(t, \hat{\theta}) \right)$$

Linear-Unlearning-Leave-One-Out 2

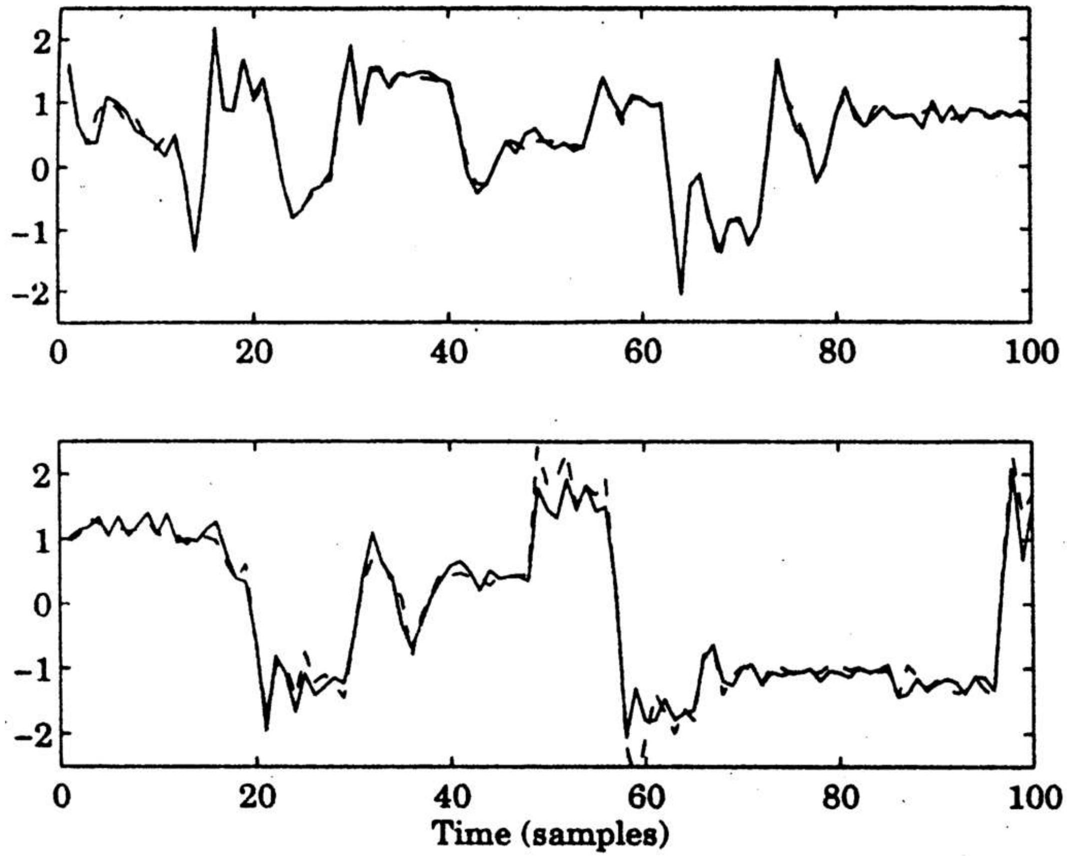
- Calculating the full Hessian is difficult, calculating it N times is even more so
- A simpler estimate is derived using the Gauss-Newton approximation of the Hessian

$$\hat{V}_M = \frac{1}{2N} \sum_{t=1}^{N} \left[\varepsilon^2(t, \hat{\theta}) \frac{N + \psi^T(t, \hat{\theta}) H^{-1} \psi(t, \hat{\theta})}{N - \psi^T(t, \hat{\theta}) H^{-1} \psi(t, \hat{\theta})} \right]$$

- “Example-based FPE estimate”

Visualization of the predictions

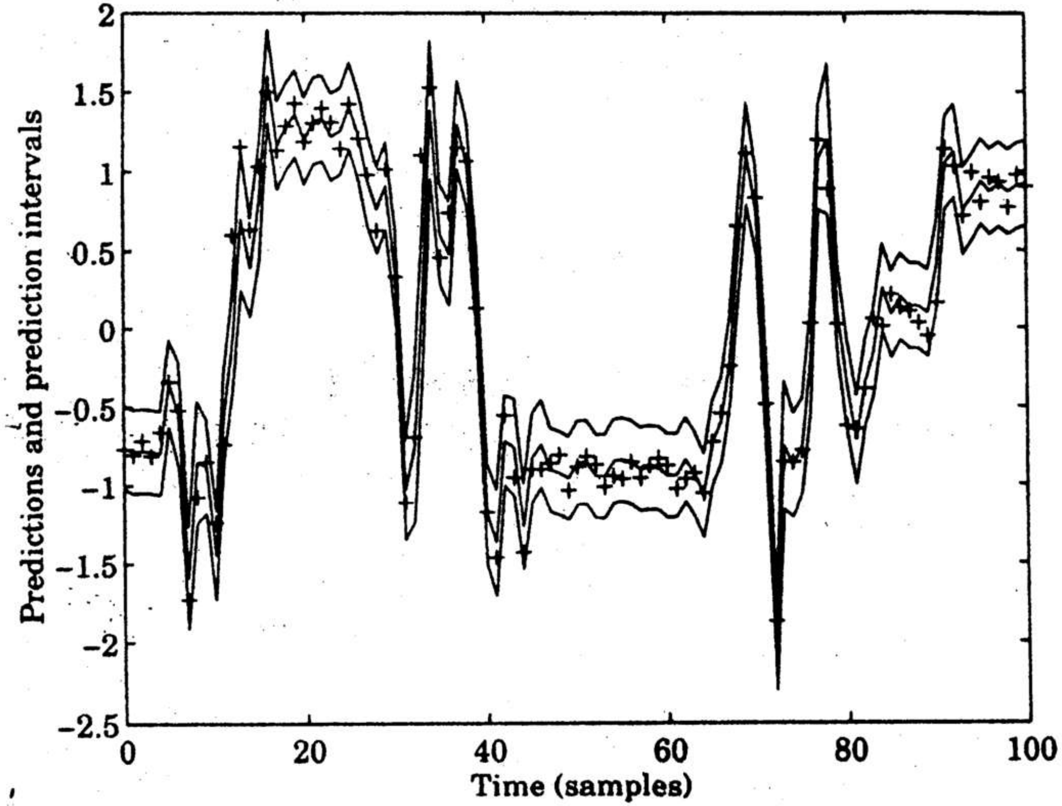
- Different regimes of the operating range?



Prediction intervals

$$y(t) \in \left[\hat{y}(t|\hat{\theta}) - c\,\sigma_p ; \hat{y}(t|\hat{\theta}) + c\,\sigma_p\right]$$

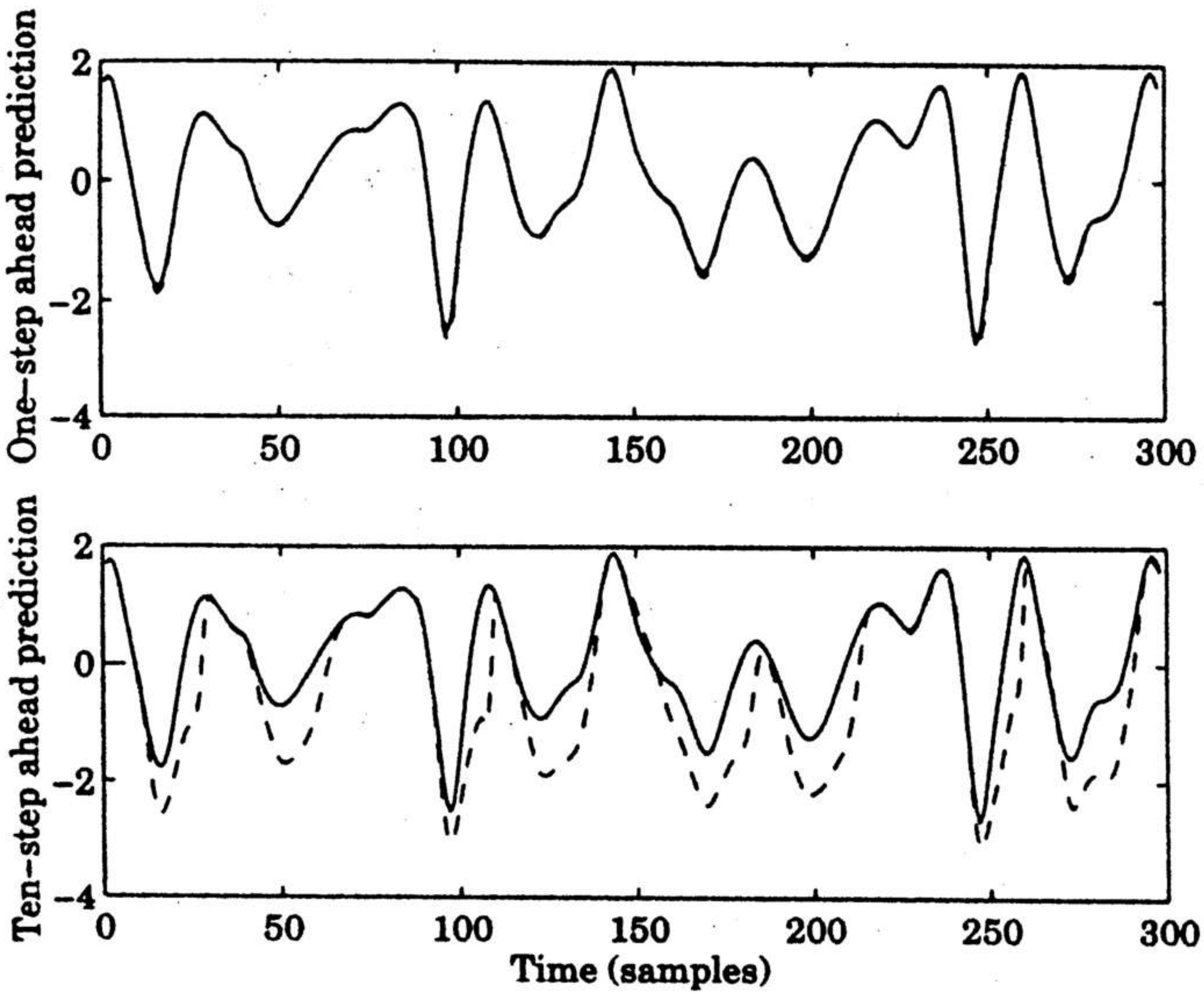
$$\sigma_p^2(t) = \boldsymbol{E}\left\{\varepsilon^2(t,\hat{\theta}) \middle| \varphi(t)\right\}$$



K-step ahead prediction 1

- High sampling frequency compared to the dynamics of the system?
  - ➔ One-step ahead prediction may not be any better than the “naive prediction”: $\hat{y}(t|\hat{\theta}) = y(t-1)$
- K-step ahead prediction: One-step ahead prediction with predictions substituting for outputs that have not yet been observed

K-step ahead prediction 2



Summary

- Correlations in the residuals
- Average generalization error
  - Akaike's FPE
  - LULOO
- Visualization of the predictions
  - Prediction intervals
  - K-step ahead prediction
- Questions?