Functional Models and Probability Density Functions.

Javier Nuñez-Garcia and Olaf Wolkenhauer*

Department of Biomolecular Sciences and
Department of Electrical Engineering and Electronics,
UMIST, Manchester, U.K.
* *Author for correspondence.*
Address: Control Systems Centre, P.O. Box 88, Manchester M60 1QD, U.K.
E-mail: o.wolkenhauer@umist.ac.uk, Tel./Fax: +44-(0)161-200-4672.

Abstract. There exist many approaches to discern a functional relationship between two variables. A functional model is useful for two reasons: Firstly, if the function is a relatively simple model in the plane, it provides us with qualitative information about the relationship. Secondly, given a fixed value for one variable, the other one can be calculated as a means for prediction. In this paper an approach for the extraction of functional models from probability density functions is proposed. The transformation of the conditional probability density function into a single value or a set of values is the basis for our discussion. Several transformations such as the mean value, the median and the modal intervals are well established. Regression models are compared to the functional models introduced here and as a consequence, two indicators to relate functional models to probability density functions are provided.

1 Introduction

Let f be a probability density function (*pdf*) of the variable $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ and $\mathbf{y}$ are a multivariate and a univariate variables, respectively. There exist many techniques to find the best function that explain the relation between those variables. The most frequently used objective function is the mean squared error which is the sum of the square of the Euclidean distances between the functional models and the data. In this paper we introduce a general functional model based on the probability density function of the joint variable $\mathbf{z}$. The idea is to summarize the conditional probability density function for each value of $\mathbf{x}$ into a single value or set of values. Firstly, we introduce the general formula and secondly we provide the application of some commonly used statistics to the conditional *pdf*s. Some cases where the conditional *pdf*s are summarized into intervals instead of single values are considered. It is interesting to know whether or not a functional model defined on the state space falls inside high density areas. In this case the functional model provides the most likely responses. The last section introduces two new indicators to measure how close the forecast generated by a functional model is from the most likely response.

2 Extracting functional models from probability density functions

The marginal *pdf* for $\mathbf{x}$ is defined

$$f_{\mathbf{x}}(x) = \int_{-\infty}^{\infty} f_{\mathbf{z}}(x, y)\ dy,\ \forall x \in \mathbb{R},$$

and the conditional probability density function, denoted by $f_{\mathbf{y}|\mathbf{x}}(\cdot|x)$, of $\mathbf{y}$ knowing that $\mathbf{x} = x$ is defined as

$$f_{\mathbf{y}|\mathbf{x}}(y|x) = \frac{f_{\mathbf{z}}(x, y)}{f_{\mathbf{x}}(x)},\ \forall y \in \mathbb{R}, \tag{1}$$

if for the marginal *pdf* $f_{\mathbf{x}}(x) > 0$.

Similarly than in regression analysis, supposing that $\mathbf{x}$ is the independent variable and $\mathbf{y}$ the dependent variable, a general formulation to extract functional models from a *pdf* by means of the conditional *pdf* is

$$\hat{y}_o = \mathrm{G}\left(f_{\mathbf{y}|\mathbf{x}}(\cdot|x_o)\right),\ \forall x_o \in \mathbb{R}, \tag{2}$$

where G is a function defined on the space of *pdf*s F, and $\hat{y}_o$ is a real number or a set of real numbers. In the last case it is denoted by $\hat{Y}_o$. In next section we introduce some functionals that summarize the conditional *pdf* (1), into some familiar statistics and intervals.

3 Forecasting with density functions

Forecasting with *pdf*s comprises two important steps: estimation of the *pdf* [Ros56,Par62,Sil86] and choice of functional G. In what follows several functional G are introduced. To start we classify G depending on the range space, $\mathbb{R}$ or $2^{\mathbb{R}}$, i.e., whether the forecast is a single point or a of set points.

3.1 Single-value forecast

In this case functional G maps elements from F to $\mathbb{R}$. The following are some examples of functions that provides single-value forecasts.

1. From function

$$\mathrm{G}(f) = \int_{\mathbb{R}} y f(y)\ dy\ ,$$

the *conditional mean forecast* is obtained:

$$\hat{y}_o = \int_{\mathbb{R}} y f_{\mathbf{y}|\mathbf{x}}(y|x_o)\ dy,\ \forall x_o \in \mathbb{R}. \tag{3}$$

In Figure 1 (red curve) an example of conditional mean functional is shown.

2. From function

$$\mathrm{G}(f) = \{y_o \colon \int_{-\infty}^{y_o} yf(y)\ dy = \alpha\},$$

where $\alpha \in [0, 1]$, the *conditional quantile* forecast is obtained:

$$\hat{y}_o = \{y_o \colon \int_{-\infty}^{y_o} yf_{\mathbf{y}|\mathbf{x}}(y|x_o)\ dy = \alpha\}. \tag{4}$$

When α is 0.5 we obtain the *conditional median* forecast.

3. From function

$$\mathrm{G}(f) = \arg\max\{f(y) \colon y \in \mathbb{R}\} \tag{5}$$

the *conditional mode* forecast is obtained:

$$\hat{y}_o = \arg\max\{f_{\mathbf{y}|\mathbf{x}}(y|x_o) \colon y \in \mathbb{R}\},\ \forall x_o \in \mathbb{R}, \tag{6}$$

In Figure 1 (blue curve) an example of conditional mean functional is shown. Note that the conditional mode forecast is not a single point when the maximum of the *pdf* is achieved in more than one point.

Depending on the situation, a different approach may be used. For example, when the mean square error[1] cost function is applied, the mean value of Equation 4 is the only value of the support of the variable that minimizes it. When the distance $d(x, \dot{x}) = 0$ iff $x = \dot{x}$, 1 otherwise, is applied, the single values that minimize it are the global modes[2] of the density function given by Equation 6.

3.2 Set forecast

In this case functional G maps elements of F onto elements $2^{\mathbb{R}}$, which is defined as the set of subsets of $\mathbb{R}$ or power set of $\mathbb{R}$. There are different techniques to summarize the support of a variable into a set according to the uncertainty reflected in its *pdf*. For unimodal and symmetric distributions such as a Gaussian, symmetric intervals about the mean are the most reasonable choice. For example in [KLRK97] the authors refer to the "three sigma" rule: values for which $|x - \mu| > 3\sigma$, where μ is the mean and σ the standard deviation, are classified as "impossible" to occur. We define the symmetrical interval about the mean forecast by applying the functional

$$\mathrm{G}(f) = [\mu - r\sigma, \mu + r\sigma],$$

[1] Mean of the square Euclidean distances between the actual response of the system and predicted values.

[2] A global mode is a value where the probability density function achieves its global maximum.

where r is a positive real number,

$$\mu = \int_{\mathbb{R}} y f(y)\ dy \quad and \quad \sigma = \int_{\mathbb{R}} (y-\mu)^2 f(y)\ dy.$$

Thus we obtain a *conditional interval about the mean* forecast

$$\hat{Y}_o = \mathrm{G}(f_{\mathbf{y}|\mathbf{x}}((\ \cdot\ )|x_o)),\ \forall x_o \in \mathbb{R}. \tag{7}$$

This type of intervals are often used to eliminate outliers in sample of data. The integral outside of this type of intervals, is the error that the random variable is erroneously classified as an outlier, or in other words, that an observation of the random variable takes a value outside the interval. For non-symmetric distributions, functional G could be defined as

$$\mathrm{G}(f) = [y_{\alpha/2}, y_{1-\alpha/2}],$$

where $\alpha \in [0,1]$,

$$\begin{aligned} y_{\alpha/2} &= \{y_o \colon \int_{-\infty}^{y_o} y f(y)\ dy = \alpha/2\} \quad \text{and} \\ y_{1-\alpha/2} &= \{y_o \colon \int_{-\infty}^{y_o} y f(y)\ dy = 1-\alpha/2\}. \end{aligned} \tag{8}$$

Thus we obtain a *conditional quartile interval* forecast

$$\hat{Y}_o = \mathrm{G}(f_{\mathbf{y}|\mathbf{x}}(\cdot|x_o)),\ \forall x_o \in \mathbb{R} \tag{9}$$

as a set of "possible" values of the random variable knowing that $\mathbf{x} = x_o$. There is a crucial difference with the previous defined above: opposed to the Gaussian case, it could occur that some values outside of interval have a higher density than values inside it, i.e., for a fixed value y inside the interval there exists one or more values y' outside of the interval such that $f(y) \leq f(y')$. This clashes with the following principle:

If a value y is "possible", a "more" probable value y' is "possible" too.

Although the probability for any single value is equal to zero since it is the integral of the density function on a null-Lebesgue set, it still makes sense when comparing single values probabilities by using the density function. For non-symmetric distributions, there is not such a "centre" of the distribution from which to construct symmetric intervals verifying the above principle. The only approach that avoids this inconvenience consists of choosing the set of "possible" values composed by the values with highest density, which corresponds to the level set of the density function, which probability is equal to α, being $1-\alpha$ a given fixed error. The level sets of density functions correspond to regions with the minimum volume or Lebesgue measure for a given error $1-\alpha$. This

means, that given a real value $t \in [0, \sup f(y)]$, level set $A_t = \{y\colon f(y) \geq t\}$ with $\mathrm{P}(A_t) = \alpha$ and for all $A \subseteq \mathbb{R}$ such that $\mathrm{P}(A) = \alpha$, we have $\lambda(A) \geq \lambda(A_t)$, i.e., A_t has minimum volume [NK99,NnGKCW03]. This regions are also known as modal sets [Pol95]. Thus we define the functional G

$$\mathrm{G}(f) = \{\{y\colon f(y) \geq t\}\colon \int_{\mathbb{R}} f(y)\ dy = \alpha\}.$$

for $\alpha \in [0, 1]$. Thus that we obtain a *conditional modal interval* forecast

$$\hat{Y}_o = \mathrm{G}(f_{\mathbf{y}|\mathbf{x}}(\cdot|x_o)),\ \forall x_o \in \mathbb{R}. \tag{10}$$

Note that in this case $\hat{Y}_o$ could be composed for more than one interval. This is discussed in greater detail in [NnGKCW03].

4 Relation with regression models

Models from Equations (3) and (6) are similar when for all $x \in \mathbb{R}$ function $f_{\mathbf{y}|\mathbf{x}}(\cdot|x)$ is symmetric and unimodal. This is the case for which $f_{\mathbf{y}|\mathbf{x}}(\cdot|x)$ is the *pdf* of a normally distributed random variable. Note that this conditions are commonly held by regression models. Suppose a regression model r such that

$$y_i = r(x_i) + e_i,\ \forall i = 1, \ldots, n\ , \tag{11}$$

where n is the sample size and e_i are n independent and identically normally distributed random variables with mean equal to zero an variance equal to σ^2. Consequently, the response random variable y_i is normally distributed with zero mean and variance equal to σ^2. In Figure 1, an illustration of the similarities between the regression model (black line) and functional models 3 (red curve) and 6 (blue curve) is shown. In the bottom of Figure 1 the conditional *pdf* for the regression model (plain curve) and the conditional mean *pdf* (dashed curve) for $x = 0$ is shown. The data set has been simulated from $\mathbf{y} = \mathbf{x} + e$, where $e \sim N(0, 0.3)$ and $\mathbf{x} \sim U(-2, 2)$. The probability density function was estimated using two-dimensional kernel Gaussians.

Although the mean value minimizes the mean squared Euclidean distance, it is not always the most adequate statistic for forecasting. This is more obvious if interval forecasts are considered. Since a *pdf* indicates the distribution of the uncertainty of the random variable for each possible value, the common sense tell us that the best representative interval for forecasting is such that bounding a given α probability, is the smallest possible respect to the Lebesgue measure. This interval is the level set A_t of the *pdf* such that $\mathrm{P}(A_t) = \alpha$ as indicated in Equation 10. If the *pdf* is unimodal and symmetric, this interval is centered at the mean since it is the same as the mode, i.e., interval forecasts given by Equations 7 and 10 are equal. For other shaped *pdf*s, for example, for a symmetric about the mean but bimodal distribution, the modal intervals are the smallest (in terms of

their Lebesgue measure) for a given α and so may be preferred over the intervals about the mean value. As an example of this, in Figure 2 a Gaussian kernel estimator of the *pdf* for the Old Faithfull geyser data from [Sil86] is shown. For $\alpha = 0.77$, the corresponding level set is the union of the intervals $[1.58, 2.14]$ and $[3.38181, 4.8204]$ which length is 2. The mode and the mean value are 4.07 and 3.46, respectively. For the same probability, any symmetric interval around the mean value will have a longer length than 2 or for any interval around the mean value of length 2, the probability of the geyser to have an eruption is less probable than 0.77.

As shown above methods to built functional models, such as regression, where the errors are assumed independent and identically normally distributed, produce similar forecasts as the conditional models from *pdf*s introduced in this paper. If this assumption on the errors is not verified the regression model could provide forecasts that according to the *pdf* are little probable to occur. Thus, the regression model and the model of Equation (6) will present important differences which could indicate that the assumption on the error may be wrong. Large differences between these models could also indicate a case of multimodality of the conditional *pdf*s: for example if there are two well separated clusters of points in a plane, one over the other. The regression model would be a curve crossing the middle of both clusters whereas the conditional *pdf*s is bimodal or achieve the global mode inside of one of the clusters.

5 Functional models and the conditional modal model

For a given functional model constructed from a sample of data, using for example regression techniques, we will define two indicators of the differences respect to the model of conditional mode of Equation (6), which corresponds to the most probable values to occur according to the *pdf* of the sample of data. Let $(x_i, y_i),\ i = 1, ..., n$ be a sample of data and let $\hat{y}_i,\ i = 1, ..., n$ be the forecasts generated by the regression model.

The first indicator M_1 is the mean value of the densities of the responses $\hat{y}_i,\ i = 1, ..., n$

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{y}|\mathbf{x}}(\hat{y}_i|x_i) = \frac{1}{n}\sum_{i=1}^{n} f(x_i, \hat{y}_i)$$

The upper limit of M_1, $\bar{M_1}$ is

$$\bar{M_1} = \frac{1}{n}\sum_{i=1}^{n} \sup f_{\mathbf{y}|\mathbf{x}}(\cdot|x_i). \tag{12}$$

The closer M_1 is to $\bar{M_1}$ the better the model is with respect to *pdf* f. The indicator M_1 achieves $\bar{M_1}$ when $f_{\mathbf{y}|\mathbf{x}}(\cdot|x_i)$ achieves its maximum at $\hat{y}_i$ for all $i = 1, \ldots, n$. This means that

$$\hat{y}_i \in \arg\max f_{\mathbf{y}|\mathbf{x}}(\cdot|x_i),\ \forall i = 1, ..., n.$$

Note that the real outputs $y_i,\ i = 1, ..., n$ are not directly involved in the calculation of the indicator M_1 but indirectly when *pdf* f is calculated.

$\underline{M}_1$ is always equal to zero since $\inf f_{\mathbf{y}|\mathbf{x}}(\cdot|x_i) = 0$, for all $i = 1, ..., n$. Summarizing, the bounds of M_1 are

$$0 \leq M_1 \leq \bar{M}_1 \leq \sup f(\cdot) \tag{13}$$

For a second indicator denoted M_2, we normalize the conditional possibility measure $f_{\mathbf{y}|\mathbf{x}}(\cdot|x)$ for each data input x_i,

$$f'_{\mathbf{y}|\mathbf{x}}(y|x) = \begin{cases} \dfrac{f_{\mathbf{y}|\mathbf{x}}(y|x_i)}{\sup f_{\mathbf{y}|\mathbf{x}}(\cdot|x_i)}, & \text{if } \sup f_{\mathbf{y}|\mathbf{x}}(\cdot|x_i) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$M_2 = \frac{1}{m} \sum_{i=1}^{n} f'_{\mathbf{y}|\mathbf{x}}(\hat{y}_i)$$

In this case, defining upper and lower bounds as for M_1, we have

$$0 = \underline{M}_2 \leq M_2 \leq \bar{M}_2 \leq 1 \tag{14}$$

$\bar{M}_2 = 1$, iff for all $i = 1, \ldots, n$ function $f_{\mathbf{y}|\mathbf{x}}(\cdot|x_i)$ is not null, i.e., exists y such that $f_{\mathbf{y}|\mathbf{x}}(\cdot|x_i) > 0$. Note thus that for the sample of data used to built the *pdf* $\bar{M}_2 = 1$. The main difference between M_1 and M_2 is that M_2 does not depend on the variations of the heights of function f throughout the region of the space where the data are placed.

As an example, we apply these indicators to a nonlinear plant studied in [YW98,WY99,YW99,SR00]. The process is explained by the equation

$$y(k) = g\big(y(k-1), y(k-2)\big) + u(k) \tag{15}$$

where the nonlinear component g is

$$g\big(y(k-1), y(k-2)\big) = \frac{y(k-1)y(k-2)(y(k-1)-0.5)}{1 + y^2(k-1)y^2(k-2)} \tag{16}$$

with initial condition $(0, 0)$ and error u, uniformly distributed in $[-1.5, 1.5]$. A sample of 200 training data is generated. The validation data are another 200 data where the input signal $e(k) = sin(2\pi k/25)$. The vector of the input variables $\mathbf{x}$ has the form $[y(k-1), y(k-2)]^T$, and points in the product space are $[y(k-1), y(k-2), g(y(k-1), y(k-2))]^T$. This means that the state space is $\mathbb{R}^3$. We have fitted a two-dimensional linear model by least squares regression and a nonlinear model built by combining 200 local linear models weighted according to 200 general Gaussian functions. Both models are shown in Figure 3. The dots represent the set of training data and the squares a set of validation data. To

estimate the three dimensional *pdf* from the training data, kernel estimation has been used. For each data point in the state space, a uniform kernel *pdf* was defined over the minimum sized hyperellipsoids containing its eight nearest neighbours. The mean of these 200 kernels is the estimator of the *pdf* for the training data (See [NnGW01,NnGW02,NnG02,Wol01]). In Table 1 different results are summarized for this example. For both sets of data the indicators M_1 and M_2 are calculated. Then $\bar{M}_1$ and $\bar{M}_2$ are the maximum values that the indicators could achieve according to the *pdf*. Note that these two values are independent of the studied model.

Table 1. Comparison of a linear and a nonlinear models, according to the conditional mode functional of the *pdf* .

Training Data

	M_1	$\bar{M}_1$	M_2	$\bar{M}_2$
Linear	$2.31e^{-2}$	$5.24e^{-2}$	0.428412	1
Nonlinear	$5.12e^{-2}$	$5.24e^{-2}$	0.982	1

Validation Data

	M_1	$\bar{M}_1$	M_2	$\bar{M}_2$
Linear	$1.13e^{-2}$	$3.7e^{-2}$	0.276	0.96
Nonlinear	$3.5e^{-2}$	$3.7e^{-2}$	0.865	0.96

Note in Table 1 that $\bar{M}_2 = 1$ for the training data and $\bar{M}_2 = 0.96$ for the validation data. This difference indicates that the functional models are trying to generalize into new areas where no experience (training data) is available. The difference between both maxima gives us the amount of validation data for which the model is extrapolating. In the example, the difference $1 - 0.96 = 0.04$ means that is 4% of the forecasts, i.e., eight data points, the models are extrapolating. The conditional *pdf* for these input data are equal to zero. This means that there is no evidence, based on the experience from the training data and according to the uncertainty function, to ensure that the forecasts given by the functional model for those points, behave as for the rest of the data.

6 Conclusions

The idea of using probability density functions for forecasting multiple-input single-output systems is introduced. Estimation of probability density functions in more than two dimensions could be a complex process. Nevertheless the approach explained here could also be applied to other uncertainty functions extracted from the experimental data, such as possibility measures [Zad68,DP86,DP93].

In [DS85,Wol01,NnG02] one can find some examples of how to built possibility measures from experimental data. Some techniques to summarize a *pdf* into a single value or into a set of values are reviewed. As a further work, the properties of functional G depending on the properties of *pdf* f could be determined. For example it is trivial that if function f is continuous, then the conditional mean functional G (3) is continuous too. Other functionals need additional properties for function f in order to be continuous. We have pointed out that given an input vector, the most likely response of the system is an alternative forecast to the commonly used mean value. This is especially indicated when the distribution on the error is unknown or the distribution of the data difficultly can be explained with a regression model. To finish two indicators that relate functional models to the *pdf* are introduced. The more M_1 and M_2 approximate to $\bar{M}_1$ and $\bar{M}_2$ respectively for a functional model, the more confidence we have that the assumption on the error used to calculate such a model are true.

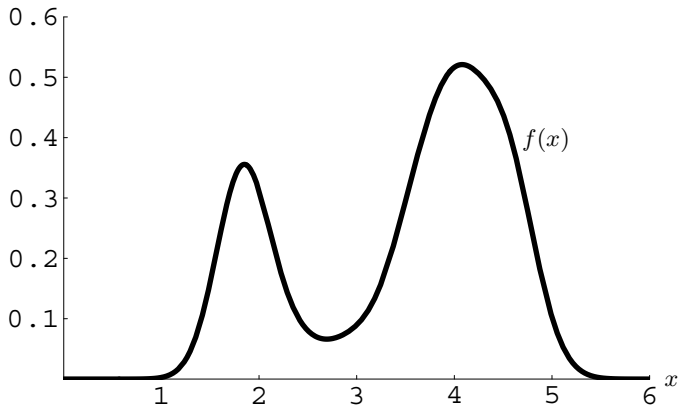

Fig. 2. Gaussian kernel estimator of the *pdf* for the Old Faithfull geyser data.

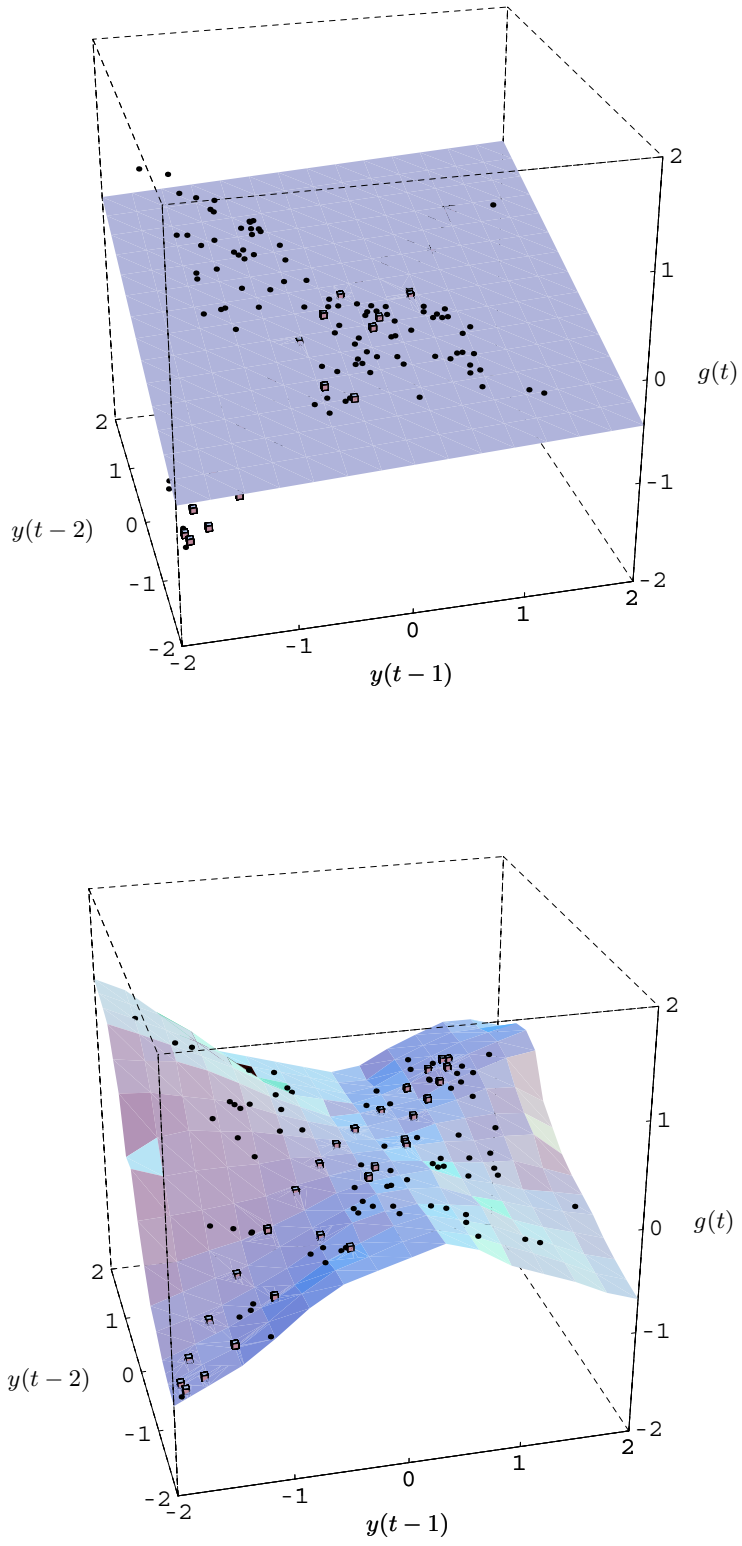

Fig. 3. A linear and a nonlinear models fitted to the training data of the nonlinear plant.

References

[DP86] D. Dubois and H. Prade. *Possibility Theory. An Approach to Computerized Processing of Uncertainty.* Plenum Press, 1986.

[DP93] D. Dubois and H. Prade. Fuzzy sets and probability: Misunderstandings, bridges and gaps. In *Proceedings of the Second IEEE Conference on Fuzzy Systems*, pages 1059–1068, 1993.

[DS85] B.B. Devi and V.V.S. Sarma. Estimation of fuzzy memberships from histograms. *Information Sciences*, 35:43–59, 1985.

[KLRK97] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl. *Computational complexity and feasibility of data processing interval computations.* Kluwer, Dordrecht, 1997.

[NK99] H.T. Nguyen and V. Kreinovich. How to divide a territory? a new simple differential formalism for optimization of set functions. *International Journal of Intelligent Systems*, 14, 3:223–251, 1999.

[NnG02] J. Nuñez Garcia. Random set theory applied to the forecasting of stochastic point processes. *Ph.D. Thesis submitted to UMIST. Department of Electrical Engineering and Electronics. Manchester. U.K.*, 2002.

[NnGKCW03] J. Nuñez Garcia, Z. Kutalik, K.H. Cho, and O. Wolkenhauer. Level sets and minimum volume sets of probability density functions. *To appear in International Journal of Approximate Reasoning*, 2003.

[NnGW01] J. Nuñez Garcia and O. Wolkenhauer. Random sets and histograms. *Proceedings of The 10th IEEE International Conference on Fuzzy Systems, Melbourne*, 2:1183 –1186, 2001.

[NnGW02] J. Nuñez Garcia and O. Wolkenhauer. Random set system identification. *IEEE Transactions on Fuzzy Systems*, 10:287–296, 2002.

[Par62] E. Parzen. On estimation of a probability density function and mode. *Annals Mathematical Statistics*, 33:1065–1076, 1962.

[Pol95] W. Polonik. Measuring mass concentrations and estimating density contour clusters - an excess mass approach. *The Annals of Statistics*, 23:855–881, 1995.

[Ros56] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals Mathematical Statistics*, 27:832–837, 1956.

[Sil86] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, 1986.

[SR00] M. Setnes and J.A. Roubos. GA-fuzzy modelling and clasification: complexity and performance. *IEEE Transactions in Fuzzy Systems*, 8:509–522, 2000.

[Wol01] O. Wolkenhauer. *Data Engineering: Fuzzy Mathematics in Systems Theory and Data Analysis.* John Wiley & Sons, New York, 2001.

[WY99] L. Wang and J. Yen. Extracting fuzzy rules for system modelling using a hybrid of genetic algorithms and kalman filter. *Fuzzy Sets and Systems*, 101:353–362, 1999.

[YW98] J. Yen and L. Wang. Application of statistical information criteria for optimal fuzzy model construction. *IEEE Transactions on Fuzzy Systems*, 6:362–371, 1998.

[YW99] J. Yen and L. Wang. Simplifying fuzzy rule-based models using orthogonal transformation methods. *IEEE Transactions on Systems, Man and Cybernetics*, 29:13–24, 1999.

[Zad68] L.A. Zadeh. Probability measures of fuzzy events. *Journal of Mathematical analysis and Applications*, 23:421–427, 1968.