The intersection between Bayesian Modeling, Kalman Filters and Adaptive Filters

Jesús Fernández Bes (jfbes@unizar.es)
June 3, 2016

http:\jesusfbes.es
1. Who and why?
Ph. D. Thesis at Signal Theory and Communications Dept. of Universidad Carlos III de Madrid.

- **Algorithms for Energy-Efficient Adaptive Wireless Sensor Networks.**
  - Stochastic Approximation + Dynamic Programming
    \(\sim\) Reinforcement Learning
  - Adaptive Networks \(\rightarrow\) Adaptive distributed estimation.

**What I will present today:** Probabilistic Signal Models and Adaptive Filters.

Other works (Small contribution):

- Feature Selection methods for **neuroimaging** classification.
  - SVMs + Bagging.
- Machine Learning applied to **Large-Scale Web mining**.
  - Topic Modeling, Clustering, SVMs, ...
Why? and Who else?

Why I talk here about this topic?

- Useful framework to improve signal estimation and modeling. (IMHO)
- Trending topic in Signal Processing.

Joint work with:

Victor Elvira
UC3M

Steven Van Vaerenbergh
Universidad Cantabria

Luis A. Azpicueta Ruiz
UC3M

Jesús Cid Sueiro
UC3M

Manel Martinez Ramón
University of New Mexico
2. Bayesian/Probabilistic Modeling
“If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts he shall end in certainties.”
Francis Bacon - The advancement of learning
Bayesian point of view

We do not care about the “real nature” of variables and parameters. Everything is a random variable.

- Some observed (Data) and some hidden/latent.
- All about using probability to model grades of belief.

To infer unknown quantities, adapt our models, make predictions and learn from data all we need is inverse probability rule (aka Bayes rule):

\[ P(\text{Explaination}|\text{Data}) = \frac{P(\text{Data}|\text{Explaination})P(\text{Explaination})}{P(\text{Data})} \]
Metaphysical Reason. It is the correct way of reasoning:
Metaphysical Reason. It is the correct way of reasoning:
Bayesian Modeling: Why?

♦ **Metaphysical Reason.** It is the correct way of reasoning:

![Image](image.jpg)

♦ **Engineering Reasons:**
  ♦ Well-defined framework: Conceptually easy.
  ♦ *Easy* to fix hyper-parameters.
  ♦ *Easy* to include domain knowledge from experts.
  ♦ We can generate pseudo-data.
  ♦ For **Estimation - Classification** Problems: Not just a numeric output but a distribution ⇒ **Uncertainty.**
Going from classical algorithms to the use of probabilistic Models

<table>
<thead>
<tr>
<th>Classical Algorithm</th>
<th>Probability Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>Gaussian Regression</td>
</tr>
<tr>
<td>K-means</td>
<td>Gaussian Mixture Models</td>
</tr>
<tr>
<td>Kernel Ridge Regression</td>
<td>Gaussian Processes</td>
</tr>
<tr>
<td>PCA</td>
<td>Probabilistic PCA</td>
</tr>
<tr>
<td>Softmax Classifier</td>
<td>Multinomial Regression</td>
</tr>
</tbody>
</table>
Standard PCA: Synthesis View

From a data set \( \{ \mathbf{x}_i \}_{i=1}^N \) where each \( \mathbf{x}_i \in \mathbb{R}^D \) we want to find an orthogonal set of \( L \) linear basis vectors \( \mathbf{w}_j \in \mathbb{R}^D \), and the corresponding \( \mathbf{z}_i \in \mathbb{R}^L \), such that we minimize the average reconstruction error

\[
J(W, Z) = \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{x}_i - W \mathbf{z}_i \|_2^2
\]

**Its solution** is SVD of \( \Sigma = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T \).

\( W = V_L \), i.e. \( L \) dominant eigenvectors of \( \Sigma \) and \( \mathbf{z}_i = W^T \mathbf{x}_i \).
**Probability Model**

- Assume a Linear-Gaussian model between data $x_i$ and latent variables $z_i$.
  \[ x_i = Wz_i + \epsilon_i \]
  where $\epsilon_i$ is zero-mean Gaussian noise of variance $\sigma^2$.
- Integrate out $z_i$.
  \[
p(x|W) = \prod_{i=1}^{n} \mathcal{N}(x_i|0, WW^T + \sigma^2 I)\]

**Maximum Likelihood solution**

- Projection Matrix
  \[
  \hat{W} = V_L(\Lambda - \sigma^2 I)^{\frac{1}{2}}
  \]
- Uncertainty of our estimation → Unexplained Variance
  \[
  \hat{\sigma}^2 = \frac{1}{D - L} \sum_{j=L+1}^{D} \lambda_j
  \]
Figure 1: Standard PCA (left) and Probabilistic PCA (right) [7].

Some advantages of probabilistic PCA

1. Use EM algorithm instead of SVD.
2. Integration within other models.
3. Model selection through Bayesian treatment of parameters.
4. Marginalization of missing data.
Bayesian Modeling: Why not?

- Computation and memory complexity.
  - Sometimes true.

- Some problems are not well modeled using probabilities.
  - Really??

- You have to make assumptions and write them down.

- Some of our favorite techniques does not seem to fit in this framework.
  - SVMs
  - Adaptive Filters, ...
3. Quick review of Adaptive Filters
Adaptive Filters

FILTERING
Extract information about a quantity of interest from noisy sources using only past and present data.

Adaptive Filter: Recursive algorithm to perform online filtering in a non stationary regression problem.

♦ Main Applications: Communication, acoustics, . . .
♦ Main Problems: Non stationary noise cancellation, acoustic echo cancellation, . . .
Classical Adaptive Filters

Figure 1.1: General adaptive-filter configuration.

The complete specification of an adaptive system, as shown in Fig. 1.1, consists of three items:

1) **Application**: The type of application is defined by the choice of the signals acquired from the environment to be the input and desired-output signals. The number of different applications in which adaptive techniques are being successfully used has increased enormously during the last two decades. Some examples are echo cancellation, equalization of dispersive channels, system identification, signal enhancement, adaptive beamforming, noise cancelling, and control [14]-[20]. The study of different applications is not the main scope of this book. However, some applications are considered in some detail.

2) **Adaptive-Filter Structure**: The adaptive filter can be implemented in a number of different structures or realizations. The choice of the structure can influence the computational complexity (amount of arithmetic operations per iteration) of the process and also the necessary number of iterations to achieve a desired performance level. Basically, there are two major classes of adaptive digital filter realizations, distinguished by the form of the impulse response, namely the finite-duration impulse response (FIR) filter and the infinite-duration impulse response (IIR) filters. FIR filters are usually implemented with nonrecursive structures, whereas IIR filters utilize recursive realizations.

Adaptive FIR filter realizations: The most widely used adaptive FIR filter structure is the transversal filter, also called tapped delay line, that implements an all-zero transfer function with a canonic direct form realization without feedback. For this realization, the output signal

\[ \hat{f}_w(k) = \arg \min_{f_w} \mathbb{E} \| d_k - f_w(x_k) \|^2 \]

Figure 2: Adaptive Filter [5].

Input-output pairs \( \{d_k, x_k\}_{k=0}^T \)

\[ d_k = f_w(x_k) + n_k \]
Classical Adaptive Filters

1.2 Adaptive Signal Processing

then used to form a performance (or objective) function that is required by the adaptation algorithm in order to determine the appropriate updating of the filter coefficients. The minimization of the objective function implies that the adaptive-filter output signal is matching the desired signal in some sense.

\[ \hat{w}_k = \arg \min_w \mathbb{E}\| d_k - w_k^T x_k \|^2 \]

**Figure 1.1** General adaptive-filter configuration.

The complete specification of an adaptive system, as shown in Fig. 1.1, consists of three items:

1. **Application**: The type of application is defined by the choice of the signals acquired from the environment to be the input and desired-output signals. The number of different applications in which adaptive techniques are being successfully used has increased enormously during the last two decades. Some examples are echo cancellation, equalization of dispersive channels, system identification, signal enhancement, adaptive beamforming, noise cancelling, and control [14]-[20]. The study of different applications is not the main scope of this book. However, some applications are considered in some detail.

2. **Adaptive-Filter Structure**: The adaptive filter can be implemented in a number of different structures or realizations. The choice of the structure can influence the computational complexity (amount of arithmetic operations per iteration) of the process and also the necessary number of iterations to achieve a desired performance level. Basically, there are two major classes of adaptive digital filter realizations, distinguished by the form of the impulse response, namely the finite-duration impulse response (FIR) filter and the infinite-duration impulse response (IIR) filters. FIR filters are usually implemented with nonrecursive structures, whereas IIR filters utilize recursive realizations.

**Figure 2**: Adaptive Filter [5].

Input-output pairs \( \{d_k, x_k\}^T_{k=0} \)

\[ d_k = w_k^T x_k + n_k \]

\( w_k \) is a vector of \( M \) parameters.
1.2 Adaptive Signal Processing

then used to form a performance (or objective) function that is required by the adaptation algorithm in order to determine the appropriate updating of the filter coefficients. The minimization of the objective function implies that the adaptive-filter output signal is matching the desired signal in some sense.

The complete specification of an adaptive system, as shown in Fig. 1.1, consists of three items:

1) **Application**: The type of application is defined by the choice of the signals acquired from the environment to be the input and desired-output signals. The number of different applications in which adaptive techniques are being successfully used has increased enormously during the last two decades. Some examples are echo cancellation, equalization of dispersive channels, system identification, signal enhancement, adaptive beamforming, noise cancelling, and control [14]-[20]. The study of different applications is not the main scope of this book. However, some applications are considered in some detail.

2) **Adaptive-Filter Structure**: The adaptive filter can be implemented in a number of different structures or realizations. The choice of the structure can influence the computational complexity (amount of arithmetic operations per iteration) of the process and also the necessary number of iterations to achieve a desired performance level. Basically, there are two major classes of adaptive digital filter realizations, distinguished by the form of the impulse response, namely the finite-duration impulse response (FIR) filter and the infinite-duration impulse response (IIR) filters. FIR filters are usually implemented with nonrecursive structures, whereas IIR filters utilize recursive realizations.

**Adaptive FIR filter realizations**: The most widely used adaptive FIR filter structure is the transversal filter, also called tapped delay line, that implements an all-zero transfer function with a canonic direct form realization without feedback. For this realization, the output signal

\[ d_k = w_k^T x_k + n_k \]

\( w_k \) is a vector of \( M \) parameters.

\[ \hat{w}_k = \arg \min_w \mathbb{E} \| d_k - w_k^T x_k \|^2 \]

**Two main families of algorithms:**

- **Least Mean Squares (LMS):**
  \[ \hat{w}_k = \hat{w}_{k-1} + \mu_k e_k x_k \]

- **Recursive Least-Squares (RLS):**
  \[ \hat{w}_{k-1} = \hat{w}_{k-1} + K_k e_k x_k \]

**Figure 2**: Adaptive Filter [5].

Input-output pairs \( \{d_k, x_k\}_{k=0}^T \)

\[ d_k = w_k^T x_k + n_k \]

\( w_k \) is a vector of \( M \) parameters.
Main Idea

Derive **Adaptive Filtering** algorithms from a probabilistic perspective.

- Useful theory to analyze the algorithms and compare them with other approaches, i.e. **Kalman Filters, Particle Filters**.
- Physical **interpretation** of parameters in terms of model.
- Get new variants that **work better**. Difficult to derive from classical perspective.
4. Probabilistic State-Space Models
• Transition / Process / Dynamics Equation:

\[ w_k = g(w_{k-1} | \theta_w) \]

• Measurement Equation:

\[ y_k = h(w_k | \theta_y) \]

Infer \( w_k \) from the observed sequence \( y_{1:k} = \{y_{k'}\}_{k'=0}^{k} \).
Bayesian Filtering

For each measurement $y_k$:

1. Compute Prediction:
   \[ p(w_k|y_{1:k-1}) = \int p(w_k|w_{k-1})p(w_{k-1}|y_{1:k-1})dw_{k-1}. \]

2. Update using new data point:
   \[ p(w_k|y_{1:k}) = \frac{p(y_k|w_k)p(w_k|y_{1:k-1})}{p(y_{1:k})}. \]

If the model is Linear and Gaussian

\[ w_k = A_{k-1}w_{k-1} + q_{k-1} \]
\[ y_k = H_kw_k + r_k \]

$r_k \sim \mathcal{N}(0, R)$ and $q_k \sim \mathcal{N}(0, Q)$.

The solution is known as Kalman Filter.
Some extra words about Kalman Filters

Applications: GPS, target tracking, biological processes modeling, patient monitoring, stochastic control, . . .

COOL EXAMPLE: Apollo 11’s Eagle IMU System

Extensions:

- Nonlinear: Extended KF, Unscented KF, Cubature KF, . . .
- Non Gaussian: Particle Filters
5. Probabilistic Adaptive Filters. Putting everything together
A probabilistic model for adaptive filters

Transition

\( w_k \) follows a diffusion process (random-walk model)

\[ p(w_k | w_{k-1}) \sim \mathcal{N}(w_k; w_{k-1}, \sigma_d^2 I) \]

Measurement

\[ p(y_k | w_k) \sim \mathcal{N}(y_k; x_k^T w_k, \sigma_n^2) \]

Prior

\[ p(w_0) \sim \mathcal{N}(w_0; 0, \sigma_d^2 I) \]
Objective: Estimate Posterior over $w_k$

$$p(w_k|y_{1:k}) \sim \mathcal{N}(w_k; \mu_k, \Sigma_k),$$

In which the mean vector $\mu_k$ is given by

$$\mu_k = \mu_{k-1} + K_k(y_k - x_k^T \mu_{k-1})x_k,$$

where we have introduced the auxiliary variable

$$K_k = \frac{(\Sigma_{k-1} + \sigma_d^2 I)}{x_k^T (\Sigma_{k-1} + \sigma_d^2 I) x_k + \sigma_n^2}.$$

The covariance matrix $\Sigma_k$ is obtained as

$$\Sigma_k = \left( I - K_k x_k x_k^T \right) (\Sigma_{k-1} + \sigma_d^2),$$
Assumed Density Filtering: The LMS filter

Reduce computational cost

- $O(M^2)$ can be too complex.
- Problem is computation of $\Sigma_k$ and matrix-vector products.
- Assume a scalar matrix $(\hat{\sigma}_k^2 I)$ in each step.

\[
\hat{p}(w_k|y_{1:k}) = \mathcal{N}(w_k; \hat{\mu}_k, \hat{\sigma}_k^2 I).
\]

How can we approximate a probability distribution?

Kullback-Leibler (KL) Divergence:

\[
\{\hat{\mu}_k, \hat{\sigma}_k\} = \underset{\hat{\mu}_k, \hat{\sigma}_k}{\arg \min} D_{KL}\{p(w_k|y_{1:k})||\hat{p}(w_k|y_{1:k})\}.
\]
If MAP estimation is performed, we obtain an adaptable step-size LMS estimation

\[ \hat{w}_k = \hat{w}_{k-1} + \eta_k (y_k - x_k^T \hat{w}_{k-1}) x_k, \]

with

\[ \eta_k = \frac{(\hat{\sigma}^2_{k-1} + \sigma_d^2)}{(\hat{\sigma}^2_{k-1} + \sigma_d^2)\|x_k\|^2 + \sigma_n^2}. \]

Posterior variance \( \Rightarrow \) measure of uncertainty

\[ \hat{\sigma}^2_k = \left(1 - \frac{\eta_k\|x_k\|^2}{M}\right)(\hat{\sigma}^2_{k-1} + \sigma_d^2). \]
1. The adaptive rule has $O(M)$.
2. Only two Hyper-parameters, $\sigma_d^2$ and $\sigma_n^2$, clear physical meaning.
3. Hyper-parameters can be estimated: EM, Variational Bayes,…
4. Easy to extend.
5. Change point detection.
Direct Extension: Nonlinear function but linear in the parameters:

\[ y_k = \sum_{i=1}^{M} w_k^{(i)} \phi_i(x_k) + n_k \]

\( \phi_i \) are some kind of basis functions:

♦ Polynomials: **Volterra Filter**.
♦ Radial Basis Functions.
♦ **Fourier** or **Wavelet** basis.

Use Kernel methods to represent a nonlinear mapping: **Kernel Adaptive Filters**.
• **Respiratory Motion** traces at CyberKnife treatment at Georgetown University Hospital [12].
Potential of probabilistic KAFs

- **Respiratory Motion** traces at CyberKnife treatment at Georgetown University Hospital [12].

![Diagram of 10-step ahead prediction of PROBLSMS on respiratory motion]
Potential of probabilistic KAFs

- **Respiratory Motion** traces at CyberKnife treatment at Georgetown University Hospital [12].

```
probLMS

???

KRLS-T
```
6. Other related work
Hidden Markov Models and Factorial Models

- When the hidden variable is discrete the models are usually called **Hidden Markov Models** (HMMs).
- We can assume several underlying factors. Inference means getting the set of active factors.

Application: Energy Dissagregation

From your **power consumption** signal infer list of **active** appliances.
Factorial Switching Models

• With Esther Pueyo and David Sampedro.
• Modeling Human Ventricular Cells.
  In particular:
  ♦ Temporal & Spatial variability of action potentials and its relation with autonomic nervous system.
  ♦ Some relevant parameters cannot be observed: Use state-space models + UKF to infer them.
My current work here

• With Esther Pueyo and David Sampedro.
• Modeling Human Ventricular Cells.
My current work here

- With Esther Pueyo and David Sampedro.
- Modeling Human Ventricular Cells.
7. References
My references:


3. One letter and two journals almost finished.
Other useful Bibliography

- **To understand Adaptive Filters:**

- **To learn Bayesian modeling/ML:**

- **Kernel Adaptive Filters:**
• **Bayesian Filtering:**


• **Other Papers:**