PAOLA'S THESIS COMMENTS

LUIS SALINAS

1. Chapter 3: Machine Learning Models.

P 33, Abstract: It seems that some commas are missing: "...training period when ..." $\longrightarrow$ "...training period, when..."
"...forecasting where..." $\longrightarrow$ "...forecasting, where..."
But the whole sentence "Usually machines ... period of time" does not make too much sense to me. Please review.

P 33, L -4: Insert for this in the text: "The reason for this is that ML models are data driven and are able to examine large amounts of data."

P 34, section 3.2:
L 3: "In this thesis we will emphasised the supervised learning problem since it is the most common way of modelling financial problems."
COMMENT: (a) d in "emphasized" is wrong.
(b) That almost everybody uses this approach is not a good reason for you to use it as well! Think, for instance, of "almost everybody eats junk food; thus I will eat junk food as well"! I suggest instead the following phrasing:
"Supervised learning is most popular and most commonly used in modelling financial problems and the assessments of this method and their results in praxis are fairly good. Therefore, in this thesis we will adhere to this trend and we will use (an improved version of) supervised learning."

L 5: "...consists in to find..." $\longrightarrow$ "...consists in finding..." (?) CHECK!

L 6: "...set of examples called training set S which have..." $\longrightarrow$ "...set of examples, called training set S, which have..."

First formula: There are two "S": a slanted one and a calligraphic one; slanted-S belongs (in set theory sense) to calligraphic-S; slanted S is defined, but what is calligraphic-S?

First formula notation: Written "$\forall\ i = 1...n$"; but it seems to me that the standard form (in this format) is "$\forall\ i = 1..n$" (two points only) CHECK!

GENERAL: Put new names with `\em (\emphasized)`. Example: *hypothesis space*; What kind of functions are allowed in calligraphic H? Polynomials in several variables? Continuous functions? etc.

Last sentence: "...$V(f(x), y)$ which expected risk has to be minimised." $\longrightarrow$ "...$V(f(x), y)$, THE expected risk, WHICH has to be minimised." CHECK!

I sent you some of these last remarks in a mail from August 14. I will repeat them in this LaTeX file. These last remarks are discussed in a more expanded form below, because there are more serious mathematical problems here. There are repetitions among the remarks.

P 34, LL -16 to -1 (section 3.2): “A learning algorithm $\mathcal{A}$ takes as input a data set $S \in \mathcal{S}$ and output a function f_S:”

It should say “...and outputs a function...”

HOWEVER, there are more serious problems here and in the following paragraphs.

It is really a bad idea to work with a set S and something called $\mathcal{S}$, which seems to be a class of sets like S.

To clearly distinguish them, let us call S the set S and $\mathcal{S}$, the class $\mathcal{S}$ of sets S or simply the class $\mathcal{S}$ for short.

You see, the set S and the class $\mathcal{S}$, as fonts, are almost indistinguishable. It seems that the set S is a subset of $X \times Y$ and the class $\mathcal{S}$ would be the subset of the power set of $X \times Y$ containing all training sets S.

To avoid confusions I suggest to use `\mathscr` fonts and call $\mathscr{T}$ your set $\mathcal{S}$. The $\mathscr{T}$ stands for ”training”.

Achtung!: `\mathscr` needs `\usepackage{mathrsfs}`

However, is it really necessary to use $\mathcal{S}$ or $\mathscr{T}$?

Have all sets S in $\mathscr{T}$ the same cardinality?

If not, I think one can safely use the power set $\mathscr{P}(X \times Y)$ as the set $\mathscr{T}$ and forget $\mathscr{T}$.

But you have to judge whether this would be appropriate or not.

If your $\mathcal{S}$ ($=$ my $\mathscr{T}$) is really a subset of the power set of $X \times Y$, then I would suggest to define the set S simply as

$$S = \{(x_i, y_i) \in X \times Y : k = 1, ..., n\}$$

The more serious problem is related to the functions f_S.

What is a “learning function” f?

What are the domain and codomain of these functions? If your answer is “$f : X \to Y$”, where $X \subseteq \mathbb{R}^m$ and $Y = \mathbb{R}$ or $Y = \{1, -1\}$, then the problems is transferred to the “loss function” $V(f(x), y)$ because then both arguments of $V(\cdot, \cdot)$ lie in $Y = \mathbb{R}$. How do you define then the function V? Suppose that $Y = \mathbb{R}$ as allowed above. Then it seems that the loss function V is defined on $X \times Y = \mathbb{R}^m \times \mathbb{R}$, i.e. $V : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ assuming that the codomain of V is $\mathbb{R}$ (you must declare what the codomain of V is). But then, when you write $V(f(x), y)$, then $f(x)$ is in $Y = \mathbb{R}$ and hence $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Thus, what is true: $V : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ or $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$?

How is the functions f_S related to the sets S?

P 34, PARAGRAPH STARTING AT L -6: Something is wrong with this definition. It seems that a learning algorithm $\mathcal{A}$ assigns an object $\mathcal{A}(S)$ to each set $S \in \mathscr{T}$.

What is exactly this object $\mathcal{A}(S)$?

You say that the object $\mathcal{A}(S)$ is an element of a so called ”hypothesis space” $\mathcal{H}$. But what is this ”hypothesis space”? You do not really describe the members of this ”hypothesis space” $\mathcal{H}$. You just say that the members of $\mathcal{H}$ are functions. But what kind of functions? Domain? Range? What do they do these functions?

And furthermore you say that the algorithm $\mathcal{A}$ searches these functions in $\mathcal{H}$. How is this search exactly performed?

The whole PARAGRAPH is rather unintelligible. Before proceeding any further, a CONCRETE EXAMPLE is needed. We must come back to this definition after we discuss your example.

In the last 2 lines of this paragraph and in the first 2 lines of page 35 you write:

"The selection of f_S is based on a loss function $V(f(x), y)$ WHICH expected risk has to be minimised." Instead of WHICH you should use WHOSE, I guess. Then comes this formula:

$$E[V(f(x), y)] = \int V(f(x), y)\, dp(x, y)$$

"$V(f(x), y)$ denote the price paid for mistakes. Therefore, $V(f(x), y) = 0$ if $f(x) = y$."

There are many comments in order here:

1. By integrals always put a small space "\," or "\;" between integrand and differential ($dp(x, y)$ here).

2. What is the domain of integration for the integral? Judging by the differential, it seems that this domain is $X \times Y$ and hence the function f is a function from X into (onto?) Y.

3. Then another function appears: $V(f(x), y)$, so that V should be defined on $Y \times Y$. Is this true?

4. Whenever you define functions you must first clearly define its domain and its range (codomain). The same applies to algorithms. If there is not yet a concrete pseudo-code description of an algorithm, it is necessary at least to say something about the class of algorithms that you are going to consider, and what should your algorithms do. An algoritmh is just a kind of function, which operate on a set or space of some objects and produce an output, which a subset (consisting may be of only one member or none at all!) of objects taken fron some other set or space.

5. It is important to define the units that will be used to measure the different variables appearing in your proble. But this might be optative. For instance it seems that the value $V(f(x), y)$ is measured in USD or EUR.

6. Your formula has a major problem in the following sense: both sides depend on x and y, but on the right hand side they are integration variables (i.e., running presumably on the whole of $X \times Y$) and on the left hand side they seem to denote that the variable representing the first argument of V is x running on X, and the variable representing the second argument of V is y running on Y. This is cumbersome. I suggest:

$$E[V(f(\cdot), \cdot)] = \int_{X \times Y} V(f(x), y)\, dp(x, y)\,,$$

but this must be critically reviewed.

Recall that some lines above you wrote $X = \mathbb{R}^m$, i.e. dimension m, and $Y = \mathbb{R}$, i.e. dimension 1. Note that in this case $f : X \to Y$ means $f : \mathbb{R}^m \to \mathbb{R}$. If you write $V(f(x), y)$, then we are in troubles because the first argument of V is now $f(x)$, which is in $\mathbb{R}$. Thus, in this case, V should be defined as $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, but you wrote before $V : X \times Y \to \mathbb{R}$, i.e., $V : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$.

Must then $V(\xi, \eta)$ be defined –for a given f– only on the graph

$$\{\big(f(\xi), \eta\big) \ : \ \xi \in X, \eta \in Y\} \subset \mathbb{R} \times \mathbb{R}$$

and not on the whole of $X \times Y = \mathbb{R}^m \times \mathbb{R}$?

This feeling is re-inforce after looking to equations (3.1) and (3.2). But then all would depend on f and I think this is not very rational. I would dare to say that that the function V –as well as the differential $dp(x, y)$– is defined (somehow; examples?) on the whole of $X \times Y$:

$$\begin{array}{rccc} V: & X \times Y & \longrightarrow & \mathbb{R} \\ & (\xi, \eta) & \longmapsto & V(\xi, \eta) \end{array} \qquad \text{and} \qquad \begin{array}{rccc} dp: & X \times Y & \longrightarrow & \mathbb{R} \\ & (\xi, \eta) & \longmapsto & dp(\xi, \eta) \end{array}$$

Under this notation, when $X = \mathbb{R}^m$ and $Y = \mathbb{R}$ there are big problems with your equations (3.1) and (3.2):

$$V(\mathbf{x}, y) = (\mathbf{x} - y)^2 \qquad \text{or} \qquad V(\mathbf{x}, y) = |\mathbf{x} - y|\,, \qquad \text{where} \quad \mathbf{x} \in \mathbb{R}^m\,, \ y \in \mathbb{R}\,.$$

There is no way to assign a sound meaning to these equations.

However, if you now consider a particular function $f : X = \mathbb{R}^m \to Y$, $y = f(x) \in \mathbb{R}$, then these equations can have a sound meaning, namely:

$$V(f(x), y) = (f(x) - y)^2 \qquad \text{or} \qquad V(f(x), y) = |f(x) - y)|$$

because $f(\mathbf{x})$, for $\mathbf{x} \in \mathbb{R}^m$, and y are now both in $\mathbb{R}$. But then $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$!

All this is very different from your equations (3.1), (3.2). In these equations (3.1), (3.2) the domain of the function f would be X, but its codomain cannot be Y, because if this were the case, then V should be defined on $Y \times Y$ which seems to be nonsense at this point.

Note that for $X = \mathbb{R}^m$, $Y = \mathbb{R}$ and $f : X = \mathbb{R}^m \longrightarrow Y = \mathbb{R}$ we have:

$$\begin{array}{rccc} V(f(\cdot), \cdot): & \mathbb{R}^m \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ & (x, y) & \longmapsto & V(f(x), y) \end{array} \qquad \text{but} \qquad \begin{array}{rccc} V(\cdot, \cdot): & \mathbb{R} \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ & (\xi, \eta) & \longmapsto & V(\xi, \eta) \end{array}$$

which is not consistent with your definitions.

Summarizing: you must carefully review this section. In particular, you must clarify what is the domain of V: $X \times Y$? $Y \times Y$?

P.35, LL 9-12, Formula (3.3): The same objections as before must be applied here. The main problem is with the definition of $V(x, y)$ and $V(f(x), y)$, as before. We already had f and f_S and now, without any warning, a $\widehat{f}$ appears. How are all these functions (I suppose!) related?

You say "The objective is to estimate a function $\widehat{f}$ through empirical risk (training error) minimisation (ERM)."

To estimate from what set or from where? From the class $\mathscr{T}$ of sets S? This means that in your formula (3.3) you have (at least) to declare from where you select the f's and what they are. Discrete functions? Are they related to the sets S in $\mathscr{T}$?

Formulas (3.1) and (3.2) would be OK assuming $x \in X = \mathbb{R}^m$ and $f : X = \mathbb{R}^m \to \mathbb{R}$. But again: what is the relation to the sets S? In formula (3.3) I suppose that x should be x_i and y should be y_i. Is this so?

With this interpretation in mind, the right hand side of formula (3.3) would be a Riemann sum approximation of the integral in the first line of P 35, where $p(x,y)$ is the uniform distribution and hence $dp(x,y) \approx 1/n$. Thus, you have:

$$E[V(f(x),y)] \approx R_{\text{emp}}[f]\,.$$

But what is this $R_{\text{emp}}[f]$ conceptually? So, this is a well known joint (?) distribution. But in the previous page you said that the joint distribution $p(x,y)$ was unknown. Note that you can arbitrarily change the value of $R_{\text{emp}}[f]$ just by arbitrarily changing the unknown joint distribution $p(x,y)$. So what is going on here? What do you want to estimate and how exactly do you want to do it?

P 35, several lines: L 1:quad First formula: What is the domain of integration for the integral? Also, separate "$V(f(x),y)$" from "$dp(x,y)$" a little!

L 2: Separate "V(f(x),y)= 0 if" from "f(x) = y" a little!

L 3: Are you writing "$L2$" for "L^2" or "L_2"? This would be OK if you systematically use "$L2$". Idem for $L1$.

L 15: Complexity of the hypothesis class. What notion of complexity are you using here? There are two complexities: one of the learner model, and one of the data. An explanation is needed. You give a kind of explanation in the case of polynomial regression; is this the one you will be using? Anyway, an explanation is needed.

It says: "complexity of the learner model $\widehat{f}$ to the complexity of the data f"

Do you call "learner model $\widehat{f}$" the one f you get by maximizing (3.3) among nobody knows what?

What is it and how do you compute the "complexity of the learner model $\widehat{f}$"?

What is the data f? Has it something to do with the sets S?

What is it (definition) and how do you compute the "complexity of the data f"?

You say: "In polynomial regression, the complexity parameter is the order of the fitted polynomial".

Why this is so? The word "complexity" always suggest how difficult is to perform some computation. Is this "difficulty" related to the something called "order" of a polynomial?

What is the order of a polynomial? Do you mean the degree?

L -13: "...data: generally generalisation accuracy..."
Too many "general's" together!

L -11: "Very complex models will..." ⟶ "a very complex model will..."
(so that the grammar number of the subject coincides with "it could...".

P 36, some remarks: Figure 3.1. How do you define "Generalization Accuracy"? Idem, "classifier complexity" and hence "low complexity classifier"? How do you measure them? Without these definitions the diagram makes little sense.

Ad (3.4): **NEW (as for 10.10.2016):** The presentation here has some problems. Formula (3.4) and its proof appears very suddenly. How is it connected to the last parapraph 3.2.2. in page 35? Why are you interested in dealing with formula (3.4) here?

More exactly, something like a roadmap is missing here.

I think somewhere before or in an Appendix you should define (very well) a minimal set of concepts from Statistics in order to understand the various formulas appearing in your work. Good modern references should be given for those concepts.

You use variables like x and y which are usually assumed to be continuous variables and hence the base theory would be integration theory.

But it seems to me that in the whole problem you are dealing with, samples or discrete values of these variables is what you are using.

If this is so, I would suggest that you use discrete variables, like x_k, y_k, samples, etc. Then you would have simple formulas for defining bias, expected values, sigma, etc. You should clearly state, for instance:

- what are you assuming to be known;
- what is not known;
- what quantitities you want to determine;
- how you intend to determine those quantities;
- etc.

Is this (3.4) a theorem? a remark? It seems a standard formula to me; I think there are textbooks discussing this formula. But what is σ here? If it is indeed a formula from the mathematical folklore, you could omit the proof and simply put a reference where this proof can be found.

By the way, instead of ”Demo” you should write ”Proof”. In the ”Demo”-Proof, the ellipsis ”$\ldots$” is misleading, since in formulas ”$\ldots$” is used to show that some terms have been left out, but this not the case here since the only term missing in this line is the term written in the line immediately below

Why is it $E[y - f(x)] = 0$ here?

In the second line of the ”Demo” appears a ”B”. What is this ”B”? It seems to me, that it should be ”y”.

P 37, LL 1-5 (formula).: L 1: ”$E[f(x) - \widehat{f}(x))]^2$” has two mistakes:

(1) A parenthesis ”(” is missing between ”[” and ”f”.

(2) The ”^{2}” should go between ”)” and ”]”, i.e. the ”^{2}” should go inside the brackets ”$[\ldots]$” and not outside.

A general recommendation: use `\widehat` instead of `\hat`; `\hat` produces a too tiny wedge.

L 2: The parenthesis ”(” is still missing between ”[” and ”f”, and the ”^{2}” is still outside the brackets ”[...]”.

LL 3-4: Again the problem of the misleading ellipsis ”$+\ldots$”; it must be omitted

GENERAL REMARK: In a scientific paper or in a thesis, you should NOT present ALL algebraic details, but just the minimum necessary for the reader to understand what is going on. Details must always be left to the reader. NEVER, but really NEVER, use weird school signs to denote cancelation of terms!

P 37, L 6.: “Bias is introduced by the model selection.”
How does model selection introduces bias? This should be explained.

P 37, LL 6-17: ”Therefore the model building process is repeated (through resampling) and substantially different averages of prediction values are obtained, bias will be high.”

Is this sentence correct? It seems to me that a better versiob of this sentence would be:

"Therefore, if the model building process is repeated (through resampling) and substantially different averages of prediction values are obtained, then bias will be high."

Is this your idea? Please discuss!

But I still do not see why this would be so, i.e., why is it so that resampling could produce substantially different averages of prediction values. When or under what circumstances this occurs? You have to prove this implication, or cite some reference where this is proved.

You have not yet described the model building process. You now should describe it.

In the next sentences in this paragraph the notion of "model complexity" appears again. Of course everybody has an intuitive meaning of this concept, but here we need a mathematically consistent definition and not only intuitions.

Grammar is again to be reviewed: in the sentence

P 37, L 10: "Variance measures how inconsistent are the predictions from one another, over different training sets, not whether they are accurate or not."

Inconsistence? Accuracy? Definition and/or references are again needed here.

QUESTION: If this is as claimed, how this would affect the results of your predictions in the sequel? What would be the effect on the confidence of your predictions?

The effects on your results could be devastating. Thus, how are you planning to avoid these bad effects?

P 37, L 17: "The best model will be the one has a balance between bias and variance".

Why this is so? Something like "which" or "that" between "the one" and "has a balance".

P37, Figure 3.2: Somewhere near Figure 3.2 (before, after, or in the caption) clear definitions of Prediction error and Model complexity should be given, together with an explanation about how they are measured in order to produce the diagram.

Explain the reasons of the behaviour of the red (test sample) and the blue (training samples) curves.

Why the blue one is decreasing?

Intuitively one can think that the more complex the model (whatever this could mean), the more over-fit results and hence less prediction error, but precisely this over-fit makes that the generalisation capacity of the model will be bad, and hence the error when the model is applied to the testing samples will increase (red curve). Is this correct? Please discuss, providing the right definitions.

P 38, L 2: "...in order to obtain as accurate model as is feasible ..."
→ "...in order to obtain as an accurate model as is feasible ..." You must discuss before (somewhere) the concepts of accuracy and feasibility you are using.

P 38, L 4: "truth target" Is it correct? Or is it "true target"? Two times in this line.

GENERAL REMARK: Pay attention to USA and UK English dialects!

Generalization ⟷ Generalisation
Optimization ⟷ Optimisation
Are "penalisation", "regularisation", etc., English words? Etc.

P 38, L 5: "…is the reason why in order to obtain the best model, data…"
→ "…is the reason why, in order to obtain the best model, data…"
Note the comma between "why" and "in".

P 38, L 6: "Training set is used to determine the model, validation set is used to estimate the generalisation error and finally testing set is used to estimate the accuracy of the model."
→ "The training set is used to construct the model, the validation set to estimate the generalisation error and finally the testing set to estimate the accuracy of the model."
Note that the suggested sentence has only one "is used to". We are using here the ellipsis feauture of indoeuropean languages.

Where did you rigorously define generalisation error and model accuracy?

P 38, L 10: "…procedure [Gei75], usually used when…" Too many "use" of "uses"!

P 38, L 12: "In K-fold cross-validation" References? This is most important because there should be an statistical theory as a basis for this procedure.

Must the K subsets of data be disjoint?

P 38, LL -4, -3; and P 39, LL 1,2: This must be expanded. You now talk about an optimisation function. Which is the optimisation function? How is it related to the "complexity of the model" (whatever it might be)? How is the generalisation error included in this optimisation function? Linearly? What are the "high complex models"?

PP 38/39, paragraph stating on L -2, and finishing on L 2: Did you discussed already what "ill/well-posed problems" are? Did you include Jean Jacques Hadamard as the source of these concepts? References for this circle of ideas? How do you prove that ERM is an ill-posed problem? Where is this proof?

What is a regularisation here? This concept is usually represented by a functional over a class of functions. The functional have many different forms and some additional terms are introduced as penalisations. These additional terms are usually additive, like Lagrange multipliers. But not always.

How does it looks like the Tikhonov regularisation? References?

Why and how does exactly the Tikhonov regularisation prevents overfitting?

You should present clearly your function class (which seems to be your $\mathcal{H}$ space), your optimisation functional, your penalisation terms, your regularisation parameters, etc. Of course, references are necessary.

Achtung! In L -2 you write "Regularization" but in L -1 you you write "Regularisation".

P 39, LL 3-6: It says: "Tikhonov regularisation minimised over the hypothesis space $\mathcal{H}$ for a fixed positive parameter λ the following:"

The sentence is too cumbersome and has mistakes ("minimise" should be "minimises"). I suggest:

"Tikhonov regularisation considers a functional of the form

$$(FORMULA(3.5)) \qquad R_{\text{emp}}[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x), y) + \lambda \mathcal{R}(f)$$

where λ is a real parameter and $\mathcal{R}(f)$ is the regulariser, which is a penalisation on f. The goal is to minimise the functional $R_{\text{emp}}[f]$ over the hypothesis space $\mathcal{H}$ for a fixed positive parameter λ."

There are many things unclear here:

(1) In formula (3.5) the variables x, y on the right hand side have no subindex. I assume they are x_i, y_i. Is this correct?
(2) Where do they run x, y, resp x_i, y_i?
(3) The same objection discussed in a previous item applies here: What is known? What is unknown? etc.
(4) Is V known? Is it one of those L^p-norms? If yes, which one in your specific case?
(5) What form has your penalisation $\mathcal{R}$?
(6) See the problems related to $V(f(x), y)$ discussed above.
(7) What exactly is the space of the f functions? $\mathcal{H}$? How will you represent your functions f?
(8) How exactly do you plan to perfom the minimisation? At least a pseudocode of your minimisation algorithm would be necesary at this point.

P 39, L 9: "...to solve a regression problem."
→ "...to solve a linear regression problem. In this case $f(x)$ has the form $f(x_t) = X^T . x_t$, where X^T is an appropriate matrix and $x_t \in X = \mathbb{R}^m$." Is it correct that $X = \mathbb{R}^m$ for some $m \in \mathbb{N}$?

P 39, equation in lines 11 - 13: What are $\mathbf{A}$, $\mathbf{X}$, $\mathbf{x}_t$, $\mathbf{Y}$?

What is the relationship between $\mathbf{A}$ and $\mathbf{X}^T$?

What is the relationship between $\mathbf{x}_t$ and $\mathbf{X}^T$?

A random reader of your thesis could think that:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_t \\ \vdots \\ x_N \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_t \\ \vdots \\ y_N \end{bmatrix}.$$

But these are vectors of dimension N. In equation (3.6) you say that $\mathbf{X}$ and $\mathbf{Y}$ are matrices of dimensiosn $n \times l$ and $m \times l$, respectively.

But then, what could $\mathbf{X}^T$ be? Is it a matrix built from sequentially shifted copies of the *row vector* $\mathbf{X}^T$? If yes, you must explain this.

In seems that the equation, as it is, makes no sense! Clear definitions are needed.

I suggest you review and present this equation using elementary matrix notation. This would make the incoherences evident.

From P 39, L -7, until P 40, L 3: You put this paragraph as someting called "Demo". What is it? A proof? A proof of what? What is your claim before starting with "Demo"?

Then you say "to solve equation (3.8)". But (3.8) is just the definition of More-Penrose!

Are you trying to solve (3.6)? If this is true, what is known and what is not known?

You should prove the equivalence you claim. At least write $\mathbf{C}_\lambda$ to make explicit the dependence of $\mathbf{C}$ from the parameter λ.

I think you shoud better write something like:

$$\min_{\mathbf{X}\in ???} \|\mathbf{C}_\lambda \mathbf{X} - \mathbf{F}\| \tag{3.9}$$

You should make clear in which space $\mathbf{X}$ is running for the optimisation.

The outcome of this minimisation problem depends on λ, but equation (3.6) does not depend on λ.

Thus, why is (3.6) equivalent to the minimization problem (3.9)? A proof is needed.

Which value of λ will you consider for a solution of (3.6) and why?

What is $\mathbf{A}$ and what is the rationale for introducing $\mathbf{C}$ and $\mathbf{F}$.

The equation for $\mathbf{X}(\lambda)$ in LL 1, 2 in P 40, seems to be a More-Penrose solution of $\mathbf{C}_\lambda \mathbf{X} = \mathbf{F}$. But why this solution coincides with the solution –if it exists– of the minimisation problem?

By the way, is it sure that the minimisation Problem has a solution? Is it unique? etc.

Reviewing work in progress/DATE: 19.10.2016

References

[1] Paola Arce, *Online Learning Method for Financial Time Series Forecasting*. Dissertation, UTFSM, June 2016.

CCTVal and Departamento de Informática, UTFSM, Valparaíso;
Mathematisches Institut, Universität Würzburg;
October 19, 2016
E-mail address: luis.salinas@usm.cl