Phd thesis "On smooth models for complex domains and distances" by David Miller — list of minor corrections from external examiner

1. P26, line 4: "this data" ⇒ "these data". Similar instances of "data" being treated as singular can be found throughout the thesis, for example: p52, lines 17 and 18 (two instances); p159 line -9; p162 line 5; p163 line 10; p169 lines -6 to -5; p202 line 9; p213 line 10.

2. P26, half-way down "In both of these examples ...": you talk about an "underlying process" when you mean "an underlying mean", "regression function" or "trend surface". This is an example of the sloppy use of vocabulary, discussed in the viva.

3. P28, last sentence of §1.1.1: "complimentary" ⇒ "complementary".

4. P29, line -10: "effect" ⇒ "affect". Ditto caption to Figure 1.2; p66 line 13; p227 line 6.

5. P30: the caption to Figure 1.2 claims that the fit in the middle panel interpolates the data, but as plotted it clearly doesn't.

6. P34, paragraph 2: explain the difference between thin plate splines and thin plate regression splines? (or if you're using the terms interchangeably, don't!).

7. P38, line 6 of §1.1.4: "datum" ⇒ "data". "Datum" is correct in the next line though.

8. P39, bottom line: should read "$\ldots V(\cdot)$, such that $\mathrm{Var}(Z_i) = \phi V(\mu_i) \ldots$".

9. P40: half-way down the page, I suspect the Hardin and Hilbe reference does *not* cover PIRLS — IRLS maybe. Line -9: how is $\mathcal{V}_g$ defined in the generalised case? What about other methods of smoothing parameter selection — doesn't `mgcv` use a different approach for distributions without a dispersion parameter?

10. §1.3, title: "practise" ⇒ "practice". Ditto p183, line -3.

11. P42, line before (1.9): typo "esimated". A few lines later, typo "to analysing the results ...".

12. §1.3.2: the definition of the Brier score at (1.10) is incorrect I think: it should have the observations in it. The quoted expression is simply the MSE for the estimated probability surface. From this perspective, there is nothing special about the binary case.

13. §1.3.4: explain the rationale for thinking of $\mathrm{tr}(\mathbf{A})$ as effective degrees of freedom (it is merely a useful analogy with the linear model case, where the trace of the hat matrix does indeed evaluate to the length of the coefficient vector — as you started to explain in the oral examination)

14. P44, line -11: typo "pats". Line -3: typo "a flawed" — and this assertion needs qualification (also in the Abstract).

15. P48, line -16: typo "lie to instead lie".

16. P49, lines 7–8: how does leakage "manifest itself as a breakdown of stationarity in a kriging context"? What does this mean?

17. P49, line 12: typo "to as".

18. P52, line -5: typo "soap silm".

19. p53: comment needed on missing pixels in Figure 2.1 (and, if possible, change the colour scale so that you don't use the same colour for missing data as for high incidences).

20. §2.2.1: before jumping straight into the Tweedie distributions here, it would be useful to give some insight into what you tried *before* this. It would also be worth clarifying that the Tweedie family is available within `mgcv` so that it isn't (in principle, at least) any more difficult to do this than to use any other standard family; and to explain (and defend) the need to fix the Tweedie parameter p.

21. P54, line -2: the point about anistropy arising from the use of latitude and longitude is correct; however, a sentence should be inserted to explain why this is the case.

22. Pp58–59: following the explanation of the soap film smoother, it would be helpful to indicate what software you're using to implement it.

23. P59, lines -4 to -3: what do you mean by "The smoothing parameter λ_{space} ... only exists as a way of making ... the explanation ... clearer"? I suspect it means that for soap film smoothers the term involving λ_{space} in (2.3) should actually be replaced by two terms.

24. §2.2.4: what are the assumptions required to claim that (2.6) is a posterior distribution?

25. P62, line -10: Chandler (2005) is not in the bibliography. If you're going to try and buy approval from your external examiner, you should do it properly ☺.

26. §2.3: How are the spatial groups defined in Figure 2.3(b)? Comment on the line that looks like $|r| = 1 + \sqrt{\hat{\mu}}$ in Figure 2.3(d)? (as discussed in the oral examination, my guess is that this corresponds to exact zeroes in the data).

27. P65, para 2: add a comment on the width of confidence intervals for temporal trends in Figure 2.5.

28. P66, bottom: there is a comment here about the computational cost of model-fitting. This has not been mentioned previously. How long *does* it take?

29. In Chapter 3, the background material on Schwarz-Christoffel transforms would benefit from careful checking; as discussed in the oral examination, it isn't always clear which direction you're going in (e.g. when describing the transform for the rectangle case, you say thatyou map from Γ to the upper half-plane and then map a rectangle to the upper half-plane — but surely this second stage should go in the opposite direction i.e. upper half-plane $\rightarrow$ rectangle?). Some other specific points here:

    - P72: two lines after (3.1), the domains are described as complex but the meaning of this (i.e. that the co-ordinates are represented as complex numbers) is not explained until a few lines later.
    - p74: Presumably t in the definition of J_e is complex-valued — in this case, how is $\sin t$ defined?
    - P74, line -2: typo "can computed".
    - Are there any restrictions on γ to ensure that the integral is independent of the path chosen? (Actually, Figure 3.2 suggests that the prevertices $\{w_k^*\}$ are all on the real line, so perhaps the integrals are along $\mathbb{R}$ as well, but this hasn't been stated anywhere in the main text).
    - What is meant by "mapping to the strip"? (five lines after definition of J_e).
    - P75, line -5: typo "it is know".

30. P97, line -6: "principle" $\Rightarrow$ "principal". Ditto line -4; p160, line -10; .

31. The description of MDS in §4.2.1 would benefit from tightening up a bit. E.g. the reader is left to infer that the MDS projects the points into n dimensions (which seems unnatural given that we started in just two). In equation (4.2), as presented the x^*s should be xs — although I think what is *actually* going on here is that you're pretending you don't have the original xs and are therefore simply trying to find a spatial configuration that reproduces the observed distance matrix. Perhaps this could be stated explicitly? Some other points in this section:

    - P101, line 11: "The first two terms in on ..." $\Rightarrow$ "The contributions from the first two terms on ...".
    - P101, line -9: why does the eigendecomposition of $\mathbb{S}$ have this particular form? It suggests that $\mathbb{S}$ must be positive semidefinite (or is it just symmetric?). The Choleski option certainly requires that $\mathbb{S}$ is positive definite.
    - P102, lines 19–21 "Hence ... dimensions": this sentence is unclear, although I know what you mean. I don't think it adds anything though: the next sentence says the same thing but more clearly. Suggest deleting this one.

- P102: what is the last paragraph of §4.2.1 about? Please clarify (or delete).
- P103, line 11: $\mathbf{\Lambda}$ and $\mathbf{X}^*$ are claimed to be 2×2 and $n \times 2$ here, but previously (pp. 100–101) they were both $n \times n$. A few lines later: $\text{diag}(\mathbb{S})_{ii}$ implies that $\text{diag}(\mathbb{S})$ is a matrix, whereas previously it was a vector.
- P103, first sentence of §4.2.3: this assertion about MDS can't be true in general — if it was, the results would depend on the choice of initial points.
- P103, line -2: typo "in a similar to check".

32. P104, line 17: typo "is show".

33. P107, line 8: "novel" ⇒ "a novel".

34. P108, line -4: typo "the the".

35. P110: it would be useful to spell out "EDF" the first time it appears in the chapter (line -13) — we haven't encountered this acronym for a while. In the first sentence of 'Setup', what is meant by "at three levels"? Presumably the three *models* in the next sentence are actually three *fitting methods* (the model is the same in each case). Similarly on p156, line -6.

36. P115, line 2: explain why you use a "slightly easier domain" here than in the previous chapter.

37. P118, line -3: typo "sever".

38. P120: the axis labels in Figure 4.13 (b) and (c) seem incorrect.

39. P125, after (4.11): what do you mean by "contracted by factors of ..."? Similarly, in the bottom half of p131, what is meant by "squashing" the mixture surface?

40. P127, after (4.12): when discussing the finite differences for the second derivatives, it might be helpful to indicate that this issue didn't arise in earlier chapters because the derivatives there were available analytically.

41. P128, line -2: why use † here — it wasn't used before?

42. P131, lines -10 to -9: what are the "known stretch factors", and how are they calculated? As discussed in the oral examination, you didn't estimate $\mathcal{L}^*$ "from the data" but rather calculated it empirically from the MDS transformation results.

43. P134, bottom: for clarity, it would be helpful to cross-reference back explicitly to §1.1 here for a description of the data.

44. P135 line 2: typo regarding *log* chlorophyll with gamma errors.

45. P141, lines 2–4 "Given that ...": correct grammar in this sentence.

46. P144, line 6: according to Figure 4.28, *three* dimensions seems optimal for the comb rather than four as claimed. Line -3: clarify what is meant by "a dimensional selection method that also takes into account the response"?

47. P146, line 12: typo "an as such".

48. P147, line -4: "massive" $\Rightarrow$ "a massive".

49. P148, lines 15–18: the claim that smoothing parameters must be chosen for every stage of the backfitting algorithm is incorrect. There are some arguments against backfitting in Simon Wood's book, but this isn't one of them. The curse of dimensionality *is* a problem though.

    In the next paragraph: kriging hasn't been discussed previously, except very briefly on pp48–9. The discussion here seems to demonstrate lack of awareness of modern developments in Gaussian process modelling (which is used increasingly for high-dimensional smoothing, for example in the analysis of computer experiments). It might be helpful to tone down some of the criticisms of kriging in general.

50. P152, line -6: typo "fequencies".

51. P154: as discussed in the oral examination, it might be helpful to say something about the need to optimise GCM simultaneously over the choice of projection dimension and smoothing parameter, and to explain how this is done.

52. P157: caption to Figure 5.3 erroneously states that the boxplots are in groups of four.

53. P161, lines -6 to -5 "Here the divisions of taken ...": something wrong with this sentence.

54. P168, Figure 5.10: as a general comment about the preparation of graphics, be aware that one in twenty males suffers from red-green colourblindness. This figure would be completely incomprehensible to such an individual. You might be interested in submitting your graphics for inspection to `http://www.vischeck.com/vischeck/vischeckImage.php` to verify this ...

55. P177, para 3: how does ML_P suggest that the response isn't smooth? In the previous paragraph you said that it was fitting planes to the data! A plane is just about the smoothest function you can think of ...

56. P181, line -11: typo "it's".

57. P183, line 3: typo "disirable".

58. P186 para 2: you need to clarify what you mean when you say that there is a problem in terms of interpretability (or remove the paragraph)? As discussed in the oral examination: there's an underlying model here, and all you're trying to do is to estimate the regression surface so I don't really see that there *is* a problem.

59. P196, line 10: it would be helpful to give a slightly more formal justification for this density estimator. I think it is that n individuals were observed in an effective area of $2Lw$ with detection probability $p = \int_0^w g(x)dx/w$; hence $E(N) = 2\lambda Lwp$ where λ is the intensity, and λ can be estimated as $\hat{\lambda} = n/2Lwp$.

    Three lines later: the probability density function of what? For the expected number of objects to be independent of x, you need some assumption of stationarity that has not been stated.

60. P199, line 3: typo "one can of".

61. P200 line 3: what is meant by "probability of the ith observation being within the sampled area"? *All* of the observations are within the sampled area! From the subsequent discussion it looks as though you mean the probability of an individual being within the sampled area at the time of the survey. Continuing in a similar vein, there is something odd about (7.9). First, I think the LHS should be $\hat{N}$ rather than N. Second, there must be some assumptions about the individuals *not* in the sample in order for this to be a valid estimator e.g. if there was a sub-population of individuals with a very low probability of being in the sampled area then it wouldn't work — considerations of this sort also suggest that some kind of randomisation of the sampling transects is necessary (this is mentioned subsequently in §7.6 in fact — but in general for a coherent discussion it is necessary to introduce the assumptions at the outset).

62. P201, bottom line: acronyms CDS and MCDS not defined until the end of §7.7.

63. P202, line 2: log-normality of what?

64. P203, caption to Figure 7.7, line 2: typo "first first". Lines 6–7: the third panel shows conditional rather than marginal detection functions I think. Similarly in the main text further down the page.

65. P220, caption to Figure 8.5: the second and third panels do *not* show "the quantiles ... of nest size" — my guess is that they show the conditional detection functions for these selected quantiles. Also, there is no legend in the figure to indicate which line corresponds to which quantile.

Richard Chandler
27th January 2012