List of Corrections

Joe Wandy

December 27, 2016

The following is the list of corrections that I have done for the dissertation. The page numbers listed below come from where the examiners placed their marginal comments on the initial (uncorrected) dissertation, used during the examination.

- Changed the title of the dissertation to be less boring. It is now *Unsupervised Bayesian Explorations of Mass Spectrometry Data*.
- Corrected various grammatical and typographical mistakes that were spotted by the examiners throughout the entire dissertation.
- **Pg. 1**: The second paragraph of the Introduction can be structured more clearly. It now reads as follows:

  > The processing of the raw LC-MS data in a pipeline is often necessary before biological conclusions can be drawn from the studies. The pipeline begins with peak detection, where observed peaks having mass-to-charge (m/z), retention time (RT) and intensity values are extracted from the raw data. In most studies, multiple samples are obtained and measured (producing *biological replicates*) or alternatively a sample is run through the LC-MS instruments multiple times (producing *technical replicates*). Correspondent peaks that are the same across multiple LC-MS runs are matched in the peak alignment step of the pipeline. During identification, the identities of compounds that generate the observed peaks are deduced. At this stage, fragmentation data, produced when parent peaks are further fragmented in a tandem mass spectrometry process, can also be used to aid in the annotations of compound identity through matching against spectral databases.

- **Pg. 2-3**:
  - Formatted the lists of contributing papers at the bottom of pg. 2 and the top of pg. 3 to be consistent with the bibliography.
  - Since the examination, the *PNAS* paper [10] has been accepted for publication, so it is now moved to the list of contributing papers in pg. 2.
  - The manuscript under review (listed at the top of pg. 3) was not accepted at *Bioinformatics*, so it has now been submitted to *BMC Bioinformatics* instead and is still under review.
- **Pg. 4**: Some words were previously missing from the first paragraph of pg. 4. It now reads as follows:

In particular, the chapter provides a brief context of how mass spectrometry based -omics fit in the broader picture of computational biology. In addition, the chapter also explains the nature of liquid chromatography mass spectrometry data and what necessary pre-processing steps have to be performed before the data can be used for further analysis.

- **Pg. 9**:
  - In the first paragraph of Section 2.2.1, the examiners noted that the use of *radio frequency* is incorrect. This has been replaced with *electromagnetic radiation*. The word 'small' is removed from *atoms are the small building blocks of matter* to improve clarity. Finally, *a nuclei* is grammatically incorrect and has been corrected to *a nucleus*.
  - In the second paragraph of Section 2.2.1, the phrase *Before data analysis is possible* is potentially confusing. This has been replaced with *Before statistical analysis can be performed*
- **Pg. 10**: An additional explanation on the different types of high resolution mass spectrometry instruments is requested by the examiners. The second paragraph of Section 2.2.2 has been rewritten to include this:

Modern high-precision MS instruments have very high resolving power, with accuracy up to several parts-per-million. MS instruments can be ranked in ascending order by the resolving powers of their mass analyser: **(1)** time-of-flight MS, **(2)** orbitrap MS, and lastly **(3)** Fourier transform ion-cyclotron MS. In time-of-flight MS, ions are accelerated in an electric field. The m/z ratio of an ion is measured from how long the ion reaches the detector at a known distance. In orbitrap MS, ions are trapped in an orbital motion around an electrode. Currents are generated from the trapped ions and, through Fourier transform, converted to a mass spectrum. In ion-cyclotron MS, ions are trapped and excited in a magnetic field, inducing a charge that is transformed into mass spectrum.

A higher resolving power corresponds to a better ability of the MS instrument to detect small differences in m/z ratios. Having a higher resolving power is generally very useful when trying to identify which metabolites are present in the sample as spectral peaks close in m/z values can be resolved, allowing for e.g. a broad peak at low resolution to be measured as multiple sharp peaks in high resolution. The difference between the observed m/z value to the exact m/z value of a compound is known as the mass accuracy of a mass spectrometry instrument, measured in parts-per-million, i.e. $\text{mass accuracy} = 1,000,000 * \frac{(\text{observed m/z} - \text{exact m/z})}{\text{exact m/z}}$. In this manner, compounds with identical nominal (integer) masses but different exact masses can be distinguished, allowing for greater confidence that a measured peak represent an actual distinct molecular species.

- **Pg. 15**:
  - The first three sentences in Section 2.3.3 is a bit unclear, in particular *to spike a known amount of internal standards*. This has been clarified and now reads as follows:

Standards are compounds of known concentration that produce peaks at well-defined m/z and RT values. Since they are known, peaks generated from standards can be used as the 'landmark' peaks to determine how retention time shift occur across samples. This simplifies the problem of alignment for these limited number of standard compounds that can be spiked into the samples before running them through LC-MS.

  - Replaced the incorrect usages of the open quotation mark (using ' instead of ‘) throughout the whole dissertation.

- **Pg. 17, 20**: Corrected the broken references to [94] that were shown as [?] before.

- **Pg. 30**: What are the disadvantages of the generative modelling approach? The last paragraph of Section 3.1 has been rewritten to consider this:

> Other applications of generative modelling on mass spectrometry data include modelling the assignment of formulae to peaks [98, 107], modelling the fragmentation events of tandem mass spectrometry data, where the separation is performed using liquid chromatography (CMF-ESI, [2]) or gas chromatography (CFM-EI, [3]). However, generative modelling also have some drawbacks. During generative modelling, we seek to model the joint distribution of the random variables of interest. This requires making certain assumptions on the process that generate the observed data, and it may be that our assumptions are poor approximations of the true underlying process. In this case, discriminative models (which seek to model the decision boundary between classes for prediction tasks) may perform better while requiring fewer training data [57] and having fewer parameters to tune. Discriminative models, such as logistic regressions and Support Vector Machines, have have also been applied to mass spectrometry data for the predictions of retention time [22, 13, 116] and the characteristic fingerprints of compounds from fragmentation data [48, 31].

- **Pg. 32**: Some sentences were added to explain why having an infinite mixture model is a good idea.

> An infinite mixture model introduces the flexibility to add and remove components as needed, depending on the data. This is useful when we assume that a component corresponds to a chemical compound, and the number of compounds present in the sample is not known in advance and can be inferred from the data.

- **Pg. 54**:

  - The examiners asked to make clear whether anything in Chapter 3 is original work. This is added to the fourth paragraph in Section 3.2.

    > We now introduce mixture modelling for this example peak data. Note that this chapter does not introduce any original work. Instead, the materials introduced in this chapter serve as the building blocks for the subsequent chapters in this thesis, where novel models to group related peaks will be introduced.

  - What the alternatives to LDA are, and why LDA is better. Additionally, what are its weaknesses? The second paragraph of Section 3.5 is rewritten to answer this:

The classical application of LDA is for topic discovery in the text domain, although LDA-like models have been applied to continuous data [2, 8, 12]. In the text application, data points are individual words, which can be grouped, forming a document. In a document, certain words tend to co-occur — for instance, 'Bayesian' and 'probability' are two such words – and ignoring word orders, we can represent this pattern of co-occurrences by a multinomial distribution on the counts of words in a document. In the standard mixture model construction, a document is assigned to a cluster, and all words from the same document are generated by sampling from the same multinomial distribution linked to the cluster. Following its generative assumption, the multinomial mixture model therefore provides a 'hard' clustering result where documents belongs to one cluster and all words in the document are generated by that cluster alone.

An alternative to the multinomial mixture is the probabilistic Latent Semantic Analysis (pLSA) [5], which extends the generative process by allowing for a document to contain words drawn by a mixture of different *topics*. A topic in this case corresponds to a multinomial distribution over the entire vocabulary space and serves the same purpose as a cluster in the multinomial mixture model. However, in pLSA, the document-to-topic proportion is a point estimate that is inferred from the document collection. The resulting pLSA model trained on a particular document collection may over-fit and have problems generalising to new and unseen documents from a different collection (having different topic proportions). LDA extends upon pLSA by placing a Dirichlet prior on the document-to-topic proportions, allowing for unseen documents to be account by the model in a principled way. Particularly for smaller data, the choice of document-to-topic prior distribution in LDA is also shown to be important and may lead to better modelling performance [11].

And also in the last paragraph of Section 3.5:

Since its introduction in [1], numerous extensions have been proposed to the standard LDA model. Here we highlight some interesting extensions that address the shortcomings of the standard LDA model. In the LDA model, the number of topics (K) has to be chosen in advance or estimated via some model comparison procedure, such as cross-validation. To address this, the Hierarchical Dirichlet Process model, introduced in Section 3.4, can also be used to perform topic modelling, while allowing for topics to be created and deleted as necessary based on the data. Correlations among topics are also absent from the original LDA model. This is introduced in the Correlated Topic Model [6]. Topics are also flat in the original LDA model, which is addressed in the hierarchical Latent Dirichlet Allocation (hLDA) model [3]. In the hLDA model, a tree hierarchy of topics are introduced using a nested Chinese Restaurant Process. Higher level topics in the tree corresponds to more general concepts, while low-level topics are more specific. Lastly, topics in LDA are defined over a set of finite words that cannot change over time,. An infinite vocabulary model [13] extends upon this by drawing topics from a Dirichlet Process with a base distribution over all possible words. This allows topics to contain words that change and evolve over time as the model is continuously trained on new data in a streaming context.

– The examiners also asked how many data points are there, and whether this is enough for LDA. This is added to the first paragraph of Section 7.5.2:

> We performed model selection via a 4-folds cross validation approach on one of the data file (Beer3 positive ionization mode). This data file contains 1422 fragmentation spectra (i.e. documents) over 4496 fragment or loss features (i.e. words). During MCMC inference via Gibbs sampling, the log likelihood of the model is monitored to ensure that convergence has been achieved. For each test fold being held out in the Beer3 data file, an estimate of the model evidence is also computed after training the model on the remaining training folds in the file. The number of Mass2Motifs was also selected in this manner from cross-validation.

– **Pg. 89**: More descriptions can be added about the Beer dataset used in the experiment. This goes to the first paragraph of Section 5.4.1.

> A Beer dataset of three runs from one batch that is representative of the typical biochemical diversity in a complex metabolomics study is introduced. The beer extract was acquired from a bottle of 'Seven Giraffes Extraordinary Ale brewed by William Bros. Brewery Company, with approximately 10 ml of the beer sampled from the bottle directly after opening. As a routine procedure in the facility, the standard compounds were also mixed into the beer sample. LC-MS measurement was then performed using a Thermo Scientific Ultimate 3000 RSLCnano liquid chromatography system, coupled to a Thermo Scientific Q-Exactive Orbitrap mass spectrometer. After acquisition, all RAW files containing mass spectrometry measurements in a proprietary vendor-dependant format were converted to the open mzXML/mzML formats. Mass spectra were centroided and separated into positive and negative ionization modes using MSconvert (ProteoWizard).

– **Pg. 106**:

  – The examiner asked how much data is needed for the HDP-Align model. The entire Section 6.6.1 already discusses this.
  – How much expert tuning is required on the prior hyper-parameters? The first paragraph of Section 6.5.3 is expanded to discuss this:

HDP-Align is a complex model, with many model hyperparameters that can be tuned for optimal performance. For HDP-Align (Table 6.2), we perform the experiments based on our initial choices on the appropriate parameter values. These are almost certainly less than optimal and can be optimised further. Nevertheless we found that the most important parameters can be set based on empirical knowledge of mass spectrometry, while others can be left at their default. The most important hyper-parameters are $\sqrt{\rho^{-1}}$ set to the equivalent value in parts-per-million (ppm), which is the standard deviation when separating ionisation product peaks in the same global cluster into mass clusters, and $\sqrt{\delta^{-1}}$, which is the standard deviation when assigning local to global clusters. These two values can be seen as equivalent to the tolerances when matching peaks across runs based on their m/z and RT values, and can generally be estimated from the settings of the MS instruments. For the Proteomic dataset, $\sqrt{\rho^{-1}}$ is set to 500 ppm while for the Glycomic and Metabolomic datasets, $\sqrt{\rho^{-1}}$ is set to 3 ppm. The local (within-run) cluster RT standard deviation $\sqrt{\gamma^{-1}}$ is assumed to be fairly constant and set to 2 seconds for all datasets, while the global cluster standard deviation $\sqrt{\delta^{-1}}$ is set in the following dataset-specific manner: 50 seconds for the Proteomic dataset and 20 seconds for the remaining datasets. The larger tolerance values are required for the Proteomic dataset to accommodate for greater m/z error and RT drifts across runs.

Other hyperparameters in HDP-Align are fixed to the following values and we expect that they can be left mostly unchanged on different datasets: $\alpha' = 10$, $\alpha_t = 10$, $\alpha_m = 100$. The values of the precision hyperparameters for global cluster RT (σ_0) and mass cluster (ρ_0) are set to a broad value of 1/5E6. No significant changes were found to the results when these hyperparameters for the DP concentrations and cluster precisions were varied. The mean hyperparameters μ_0 and ψ_0 are set to the means of the RT and m/z values of the input data respectively. During inference, 10000 posterior samples were obtained with the first 5000 used as burn-in, and taking every 10-th sample after burn-in for the posterior probabilities of peaks to be matched.

– It would helpful to explain the random variables of the model in the caption of Figure 6.2. This has been added, and the caption now reads as follows:

Graphical model for HDP-Align. j indexes over the input files (LC-MS runs), while n indexes over the peaks in that file. k indexes over within-file local clusters, while i indexes over across-file global clusters. In a global clusters, peaks are separated by their m/z values into mass clusters, indexed by a. x_{jn} is the observed RT value of peak n in file j, while y_{jn} is the observed m/z value. The indicator variables z_{jnk}, v_{jni} and v_{jnia} are used to denote the assignment of peak n in file j to local cluster k, to global cluster i and to mass cluster a in global cluster i respectively. Next, t_{jk} and t_i are the RT values of the local cluster k in file j and the global cluster i respectively, while μ_{ia} is the mass cluster mean. In addition, γ, δ and ρ as the precision parameters on the Gaussian components for the local RT clusters, global RT clusters and mass clusters respectively. Next, μ_0 and σ_0 are the mean and precision of the base Gaussian distribution for the global mixture component RT values, while ψ_0 and ρ_0 are the mean and precision of the base Gaussian distribution for the mass cluster means. Finally, α', α_t and α_m are the Dirichlet Process concentration parameters.

– **Pg. 116**: Made Figures 6.4 - 6.8 larger.

– **Pg. 117**: The examiners asked how reliability of the MCMC procedure can be assessed for such a complex model. A discussion is added to the fifth paragraph of Section 6.6.1:

We see from the results here that by reducing recall, it is possible to extract from HDP-Align alignment results that have a higher precision than what the baseline methods can achieve. However, care must be taken when using the HDP-Align model in actual analytical situations as the reliability of MCMC inference on such a complex model can be difficult to assess. As discussed in Section 6.5.3, the choices of user-defined model parameters play an important part in influencing the results. The most important are the Gaussian precisions of the local RT clusters (γ), global RT clusters (δ) and mass clusters (ρ). It is important for these parameters to be set sensibly based on expert knowledge on the settings of mass spectrometry instruments that generate the data. A portion of the full dataset can be used at this point to test the initial choices of parameters. Figure 6.4 also shows that some poor mixing behaviours can be observed during inference, so during actual usages, further convergence diagnostics of the model can be performed by inspecting the trace plots from running a very long MCMC chain or from multiple parallel chains. All important variables of the model are already exported to text format for further diagnostics, and an option is also provided to initialise the sampler with the same user-defined random seed each time to facilitate debugging.

– **Pg. 131**: Discuss the benefits of extending the visualisation module MS2LDAVis from the Python port of the topic modelling visualisation interface LDAVis [9].

Given its hypothesis-generating nature, the analysis of Mass2Motifs to characterise and examine their correspondence to actual biochemical substructures is an iterative and exploratory process. This is made possible through the MS2LDAVis module, an interactive web-based visualisation build upon the combination of Javascript and the D3.js library (http://d3js.org). MS2LDAVis is extended from the topic modelling visualisation interface LDAVis [9] used in the text domain, but our adaptation introduces fragmentation-specific views that are not available in the original interface. Implementations of LDAVis are available in both the 'R' and Python programming languages. We opted to base MS2LDAVis on the Python implementation of LDAVis as it allows for a cleaner integration with the rest of the pipeline, which is also coded in Python. Additionally, MS2LDAVis has also been integrated into a Web application available at http://www.ms2lda.org. This application is developed in Django (a Python-based Web framework) and is a work in progress that aims to provide an accessible web interface for topic decompositions of fragmentation spectra and annotations of the resulting Mass2Motifs. All these integrations, while still possible, will be more difficult if we had based MS2LDAVis on the 'R' implementation of LDAVis. As such, the Python port of LDAVis provides a suitable starting point to develop our extensions.

– **Pg. 160**:

  – What are you doing with your code and data for open source and library storage? The following is added to the second paragraph of Section 8.3 to point the reader to implementations of all the models and methods introduced in the thesis:

> Implementations for all the methods introduced in this thesis, alongside the evaluation datasets, are publicly available online. From Chapter 4, a Python implementation of the method that combines clustering information to improve alignment can be found at http://github.com/joewandy/peak-grouping-alignment. The PrecursorCluster model alongside the methods for matching of IP clusters introduced in Chapter 5 have also been implemented in Python and can be found at http://github.com/joewandy/precursor-alignment. The HDP-Align model introduced in Chapter 6 has a Java implementation that can be found at http://github.com/joewandy/HDP-Align. The MS2LDA workflow from Chapter 7 is developed in Python and available at http://github.com/sdrogers/MS2LDA.

– How could your work generalise to another area of spectral analysis? A new subsection 8.2.5 is added to answer this.

> Throughout this thesis, much of the attention has been given to the analysis of liquid chromatography (LC) coupled to mass spectrometry data. However, other separation technologies, such as gas chromatography (GC), can also be used. GC is generally faster than liquid chromatography but requires the compound under analysis to be vaporised, so only volatile compounds can be separated by GC [4]. The models introduced in this thesis to group ionisation product peaks can be readily adapted to GC data, particularly since chromatographic peaks produced from GC separation tend to have higher resolutions and are more reproducible (so retention time drift is less of a problem). Another separation technique that is gaining popularity is ion mobility spectrometry, where ions are first separated by their electrical potential in a carrier buffer gas. The main advantage of ion mobility is high speed of separation, in the order of tens of milliseconds, allowing for very high throughout separation and mass spectrometry measurements [7]. The models introduced in this thesis can potentially be applied to ion mobility MS data as well, although further studies must be taken as the characteristics of the resulting ionisation product peaks might differ.

– **Bibliography**: Corrected references 33, 35, 47, 57, 71, 98, 106, 133, 141 in the Bibliography section of the thesis by adding various missing details, such as the volume number, page number etc.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[2] R. Das, M. Zaheer, and C. Dyer. Gaussian LDA for Topic Models with Word Embeddings. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, 2015.

[3] D. Griffiths and M. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems*, 16:17, 2004.

[4] J. H. Gross. *Mass Spectrometry: A Textbook*. Springer Science & Business Media, 2006.

[5] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM, ACM.

[6] J. D. Lafferty and D. M. Blei. Correlated topic models. In Y. Weiss, P. B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 147–154. MIT Press, 2006.

[7] J. A. McLean, B. T. Ruotolo, K. J. Gillig, and D. H. Russell. Ion mobility–mass spectrometry: a new paradigm for proteomics. *International Journal of Mass Spectrometry*, 240(3): 301–315, 2005.

[8] S. Rogers, M. Girolami, C. Campbell, and R. Breitling. The latent process decomposition of cdna microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(2):143–156, 2005.

[9] C. Sievert and K. Shirley. LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, 2014.

[10] J. J. J. van der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess, and S. Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113(48):13738–13743, 2016.

[11] H. M. Wallach, D. M. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. Curran Associates, Inc., 2009.

[12] D. Weinshall, G. Levi, and D. Hanukaev. LDA Topic Model with Soft Assignment of Descriptors to Words. In *Proceedings of the 30th Annual International Conference on Machine Learning*, volume 28.

[13] K. Zhai and J. L. Boyd-Graber. Online latent dirichlet allocation with infinite vocabulary. *ICML (1)*, 28:561–569, 2013.