

Parallel Coordinates for Exploratory Modelling Analysis

Antony Unwin, University of Augsburg, Germany

Chris Volinsky, AT&T Research, Shannon Laboratory

Sylvia Winkler, University of Augsburg, Germany

Contact: antony.unwin@math.uni-augsburg.de

Abstract

Parallel Coordinate plots have been shown to be useful in Exploratory Data Analysis, especially when they are implemented in interactive software. They are also very useful in Exploratory Modelling Analysis, that is in the evaluation and comparison of many models simultaneously, where they can be applied in at least three quite different ways. The advantages of this innovative graphical tool will be illustrated using the software CASSATT and a data set previously analysed using Bayesian Model Averaging.

Introduction

Fitting models to complex data sets is difficult. Mathematical procedures are available on how to find an optimal model within a specified class for a single criterion, but deciding which model terms to include, whether to transform variables, whether to exclude some points as outliers, and how other, possibly qualitative, criteria affect the analysis all complicate the issue. When calculating even one model took much effort and a lot of time, it was not possible to consider a range of models, a single model had to suffice. Nowadays it is easy to calculate many, many models of different kinds, so that it has become feasible to

select models based on a broader range of considerations. Another development has been the idea of combining the results from large numbers of models, as for instance in Bayesian model averaging (Hoeting et al [1999]).

Exploratory Data Analysis refers to methods and procedures for exploring the data space to learn about a data set and so, by analogy, Exploratory Modelling Analysis (EMA) refers to methods and procedures for exploring the space of models, which may be fit, to a data set. There is little published work on the process of model building. Lindsey [1997] describes a “standard” approach and works primarily with the AIC criterion. Cook and Weisberg ([1982], [1994], [1999]) have discussed various methods for determining and checking regression models, generally concentrating on model diagnostics. Cox and Wermuth [1996] covers a subgroup of graphical models. Bayesians concentrate on comparisons of a previously specified group of models using Bayes factors (Kass et al [1995], De Santis et al [1999]). In this vein, Bayesian Model Averaging (BMA) accounts for the uncertainty in the model selection process by finding a collection of the best models in a model class, and averaging over them in accordance with posterior model probabilities. Averaging over many models in this manner provides superior out-of-sample performance compared to the typical approach of selecting a single model (Madigan and Raftery 1994). BMA has also proven successful for prediction in graphical models, generalized linear models, and survival models (see Hoeting, et al 1999 for a review).

Interactive graphics are excellent for EDA and will be used here for EMA. They are designed for exploring rather than presenting information. In particular, there is no need for labelling a display because any such information (and more) can be obtained by directly querying the graphic. The figures in this paper have been left as they appear on screen. While it would be helpful to add labelling, this would give a false impression of what is shown when an analysis is carried out and the importance of empowering the graphic with interactive features might not be apparent.

For exploratory modelling analysis two different kinds of tools are necessary: views which summarise the model space and views which concentrate on detailed assessments of models and which facilitate model comparisons.

The data set and the model space

The data set analysed in this paper comes from a trial of a treatment for primary biliary cirrhosis of the liver. There were 418 patients and 17 potentially explanatory variables. Some of the cases were used for later model validation and some of the variables had missing values, so the actual model fitting was done on a data set of 304 cases and 15 variables. 5 of the explanatory variables were binary, while the remaining 10 were treated as continuous (although two were ordinal with 3 and 4 values respectively). The data have been analysed in depth by Fleming and Harrington [1991], who used a stepwise procedure to identify a single best Cox regression model with 5 covariates, and by Raftery et al [1996] who identified 36 models and used Bayesian model averaging to combine the results. Therneau and Grambsch [2000] also include an analysis in their recent book. The model found by Fleming and Harrington was:

$$h(t) = \exp\{0.03313 \cdot \text{Age} + 0.82712 \cdot \text{Edema} + 0.89481 \cdot \ln(\text{Bilirub}) - 3.03048 \cdot \ln(\text{Alb}) + 0.23814 \cdot \ln(\text{ProthTime})\}$$

Notice that three of the variables have been log-transformed. A model with the same variables, but untransformed, would count as a different model, as would stratified models (considered in Therneau and Grambsch [2000] for this data set). It would, of course, be possible to consider many different models with exactly these variables but with different coefficients. Such models could have very similar criteria values to those of the selected model and might be more acceptable on other criteria, for example consider the model with coefficients rounded to one significant figure. Allowing this would enlarge the model space considerably and so we shall always assume there is only one model for each set of (possibly transformed) variables. For expository reasons we shall also restrict the model space in another way. It will be assumed that all models are fit to the same data points. In practice it is likely that a group of points might be classified as outliers and some models fitted without them. Again, this could enlarge the model space to be explored substantially and it would complicate some of the comparisons. Hoeting et al [1996] suggest pre-screening the data set to identify possible outliers and then defining a model by the set of variables included and by the set of observations identified as outliers. This is reasonable if there are few outliers (though even with the 21 case stackloss data set they identify 9 points as potential outliers, inflating the model space by a factor 2^9 — what might happen with a larger data set?).

For simplicity, we will define a "model" to be a subset of independent variables (already transformed if necessary), to be fit to a data set of fixed size. This implies that if there are m explanatory variables and they are all continuous or binary, the set of Cox models we are considering are the ones fit to the 2^m possible subsets of the m independent variables. Taking other types of model into consideration (perhaps accelerated failure time models) would enlarge the model space still further.

Summarising views

An obvious first requirement should be a table comprising all models of interest, giving various global model statistics such as AIC, G^2 or the posterior probability (where relevant) and acting as an interactive index to more detailed views of the models. Few statistical software packages offer even a static table of models calculated, let alone an interactive index. When hundreds of models are calculated a table listing each one individually would become unwieldy. A display of the models by criterion values may be used additionally. For posterior model probabilities a weighted histogram is a good approach. Figure 1 shows the histogram of the posterior probabilities of the 36 models fitted to the survival data set already described. Since this is a display by area, there is no limit to the number of models, which may be incorporated. Raftery et al [1996] used an approximate approach, starting with a uniform prior distribution across all models, to identify a subgroup of good models, and then normalising the posterior probabilities over this subgroup. The histogram has been weighted by the posterior probabilities themselves so that the single most probable model is represented by the bar on the right of the display corresponding to a value of 0.17. The bars on the left represent collections of models of low probability and the height of each bar reflects the sum of the probabilities of the models in that bin. A model close to the Fleming and Harrington model has been selected. Notice that this is far from having the highest posterior probability. (Log-transforming the variable Prothrombin time makes only a small difference to the fit and is mainly beneficial because of a single outlier. It is more parsimonious not to transform and that is the model selected here. Interestingly Therneau and Grambsch [2000] claim that the outlier was in fact an error.)

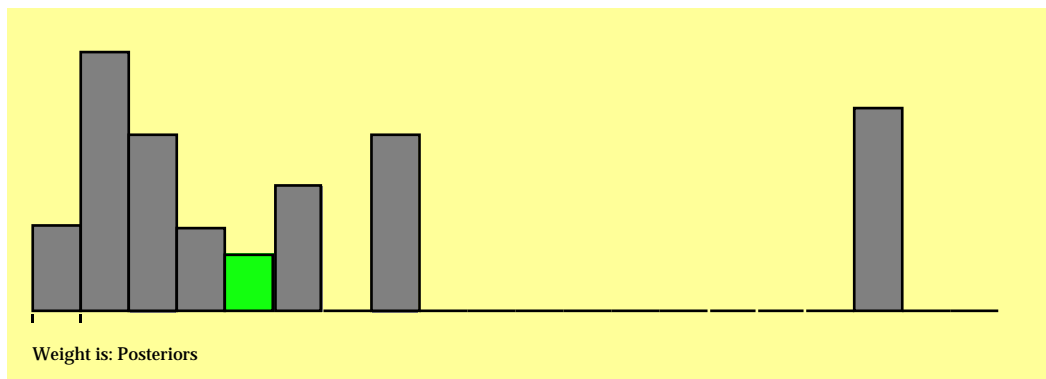


Figure 1 A weighted histogram of the model posterior probabilities with the model equivalent to Fleming Harrington highlighted (The horizontal scale is from 0 to 0.2 with equal class widths of 0.01)

A key goal of modelling is the determination of which explanatory variables should be included and how much effect they have. Hoeting et al (1999) suggest that for each variable the posterior probabilities of the models they are in should be summed giving their posterior effect probabilities (see Table 1) to measure their importance.

Variable	Posterior Effect Prob	No of Models	Min Coef	Max Coef	Wtd Coef
age	1	36	0.025	0.037	0.031
bili	1	36	0.670	0.937	0.790
albu	1	36	-3.523	-2.274	-2.783
edema	0.876	31	0.744	1.070	0.887
copp	0.788	25	0.309	0.431	0.358
thromb	0.750	25	0.215	0.326	0.259
hist	0.369	16	0.220	0.371	0.279
SGOT	0.113	6	0.0021	0.0033	0.0026
alphi	0.084	5	-0.00005	-0.00003	-0.00005
sex	0.063	5	-0.437	-0.183	-0.318
plate	0.061	4	-0.0015	-0.0007	-0.0010
hepa	0.032	2	0.186	0.243	0.203
asc	0.019	1	0.176	0.176	0.176
treat	0.018	1	0.091	0.091	0.091
spid	0.016	1	0.032	0.032	0.032

Table 1 Posterior effect probabilities for the variables included in at least 1 of the 36 models. (The number of models in which a variable appears is also given, as are the minimum and maximum estimated values of the coefficients and the weighted average.)

This is a useful first step, but has the disadvantage that each variable in a given model receives the same probability regardless of its contribution to the model. It also gives no insight into which combinations of variables appear in the same models. For subsets of up to 10 variables a weighted fluctuation diagram (Hofmann [2000]) can be helpful. This will be discussed in another paper. Table 1 also shows that the

estimated coefficients do not vary substantially across this set of models, valuable additional information regarding the interpretation of effects.

Parallel coordinate plots

Parallel coordinate plots were introduced by Inselberg (see Inselberg [1998] for some recent work) and also discussed by Wegman [1990], who proposed applying them in data analysis. They enable the display of multi-dimensional data in two-dimensional space. Each dimension is represented by a vertical axis (hence the name — all axes are parallel to each other) and values for a particular case are linked by lines. There must be some loss of information but this can be partly counteracted by varying the order of the axes. Different orders give different views of the multi-dimensional space. Interactivity is valuable for reordering the axes flexibly and fast. Interaction is also valuable for dealing with the dense mass of lines produced by large data sets. Being able to select subgroups of cases, highlight the selected lines, switch between different subgroups, all assist in interpreting the otherwise intricate displays which arise.

R-plot

Parallel coordinate plots can contribute in several different ways to EMA. Their most immediate value lies in exploring the raw data. Data analysis must always precede modelling and parallel coordinate plots are one of several interactive graphical tools that are useful. Figure 2 shows a plot from the Cirrhosis data set. The 10 cases with the longest survival times have been selected and we can see that all are censored observations. “We can see that...” is always a dubious expression in any paper, especially one concerning graphical displays, where authors tend to make extravagant claims that a first-time reader can see just as much as someone who has substantial experience of working with the plots. With interactive graphics there is the additional factor that interactive tools are not available to the reader. This is very apparent in Figure 2. For an individual to interpret what is shown, they need to be able to query the graphic, reorder it, rescale it, and generally view it in a variety of different ways, to convince themselves of the information contained in it. To distinguish the different applications of parallel coordinate displays in EMA, we refer to this one as the R-plot, R standing for Raw data.

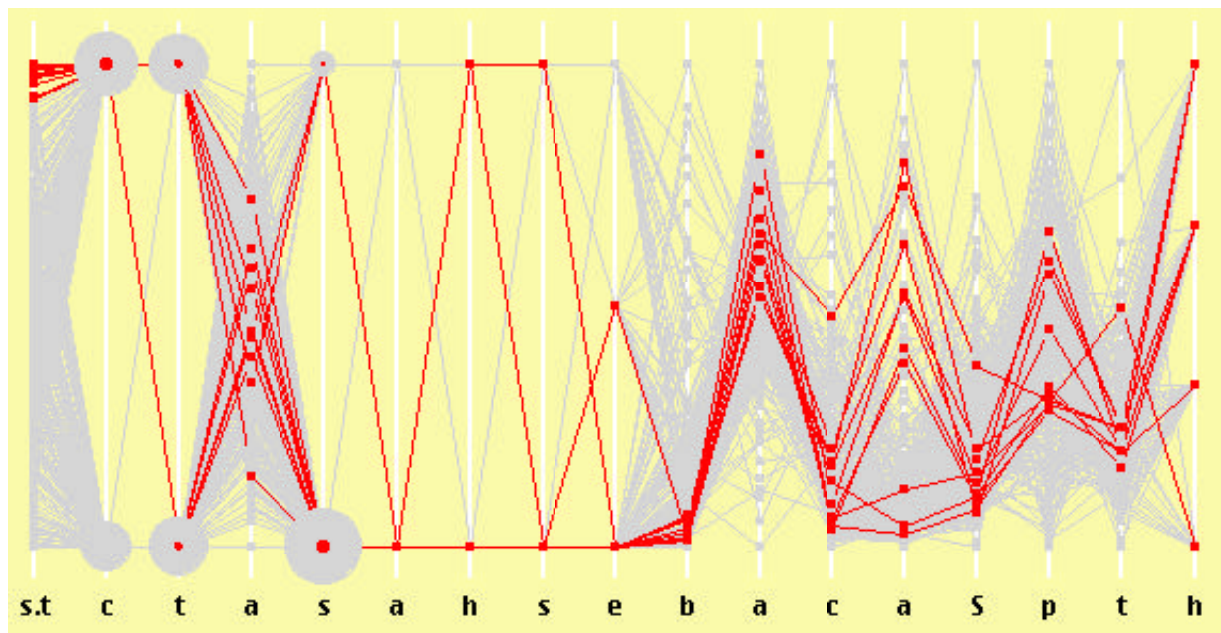


Figure 2 A parallel coordinate plot (R-plot) of the data. The first two variables on the left are survival time and censoring. The 10 longest survival times have been selected. They are all censored. Categorical variables with text codes are represented by circles proportional to category size.

t-plot

The second application is the t-plot, which displays the t statistics from the different models in a parallel coordinate plot. Evaluating the contribution of an explanatory variable to a number of models is difficult, as has already been discussed. In the t-plot each variable has its own vertical axis and the t statistics are plotted as data. This assumes that it is reasonable to compare t-statistics both within and across models. It is well known that t-statistics are not ideal, and we look forward to better suggestions. As a first approximation they are certainly very helpful.

Models are cases in the t-plot, so they are shown by lines. Figure 3 shows the 36 models fitted to the Cirrhosis data set by Raftery et al [1996]. Note that the axes have all been scaled to the same scale, that t is set to 0 when a variable does not appear in a model, that absolute t values are shown, since no variable was fitted with both negative and positive coefficients in different models, and that the axes have been ordered (in this case by the maximum of the t-statistics of the variables, but any statistic of the t values could have been taken). One can see that the rightmost three variables appear in all the models.

The use of absolute t values saves space and makes the ordering more informative. Variables with more big t statistics are likely to be more important in the modelling. All of these adjustments to the plot can be

made interactively with intuitive interface controls. Querying, selection and ordering need to be flexible and interactive to extract the information from the diagram.

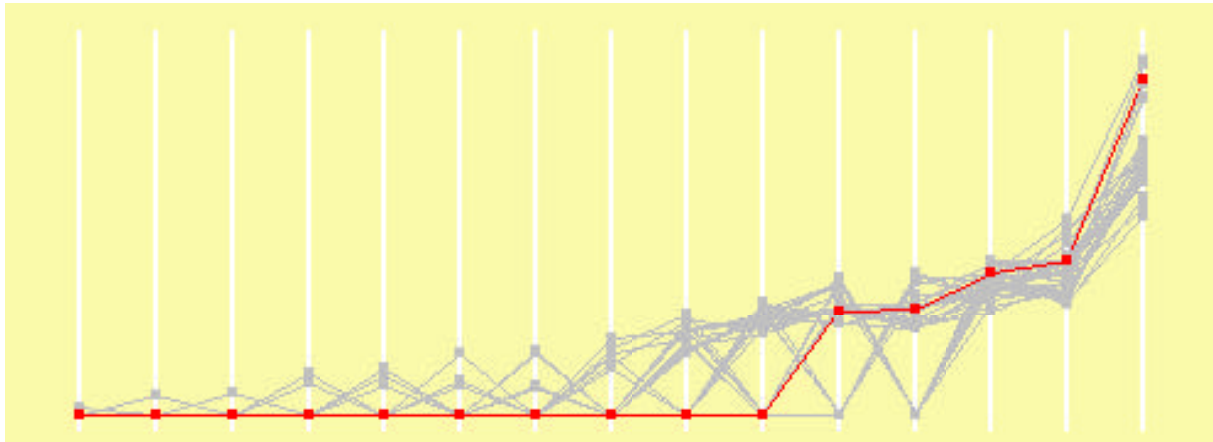


Figure 3 The absolute t-values of the coefficients of the 15 variables in the 36 models, one axis for each variable, sorted by the maximum. The FH equivalent model has been selected. t-plot

The Fleming and Harrington equivalent model (i.e. without their log transformation) has been selected in Figure 3. Interestingly, it incorporates the first 5 explanatory variables on the right of the t-plot. The t-plot shows very clearly which variables appear in all models and which variables are most/least important. It summarises the information about the influence of the explanatory variables in the modelling very effectively. (Chris Volinsky would like to order the axes by the posterior effect probabilities. This would be possible by hand, but not interactively, as a more sophisticated data structure than a data matrix would be needed in CASSATT.)

The basic t-plot could be improved and refined in various ways:

- a) Data sets where both positive and negative coefficients arise for a variable will have strong interrelationships between the explanatory variables and will need to be analysed taking account of this.
- b) Variables, which may appear in a model in transformed form (e.g. $\ln(x)$ instead of x), should also be handled with care. In principle we may take the t statistics to be comparable, but in practice we are dealing with rather different objects. A further problem arises when models include more than one term for a variable (e.g. both x and x^2). Including additional separate axes would be one solution.
- c) All explanatory variables in the Cirrhosis example were modelled as continuous or binary. Categorical variables with multiple categories will need multiple coefficients. How should the corresponding axes in the t-plot be displayed to call attention to the connections?

The M-plot and C-plot of residuals

Residual plots have been used for a long time as diagnostic tools to inspect the fit of individual models, but what kind of residual plots can be used for comparing the fit of many models? The M-plot is a parallel coordinates plot, in which the deviance residuals of each model are plotted along an axis assigned to that model (M stands for residuals by Model). Figure 4 shows the M-plot for 22 loglinear models fitted to the Copenhagen housing data set (Venables and Ripley [1999]). All these models are non-dominated in the sense that for none of them is there another loglinear model with both a lower G^2 and more degrees of freedom. The axes have been ordered by hand by the G^2 statistics for the models. On the far left we have the independence model and on the far right the saturated model (with all residuals zero).

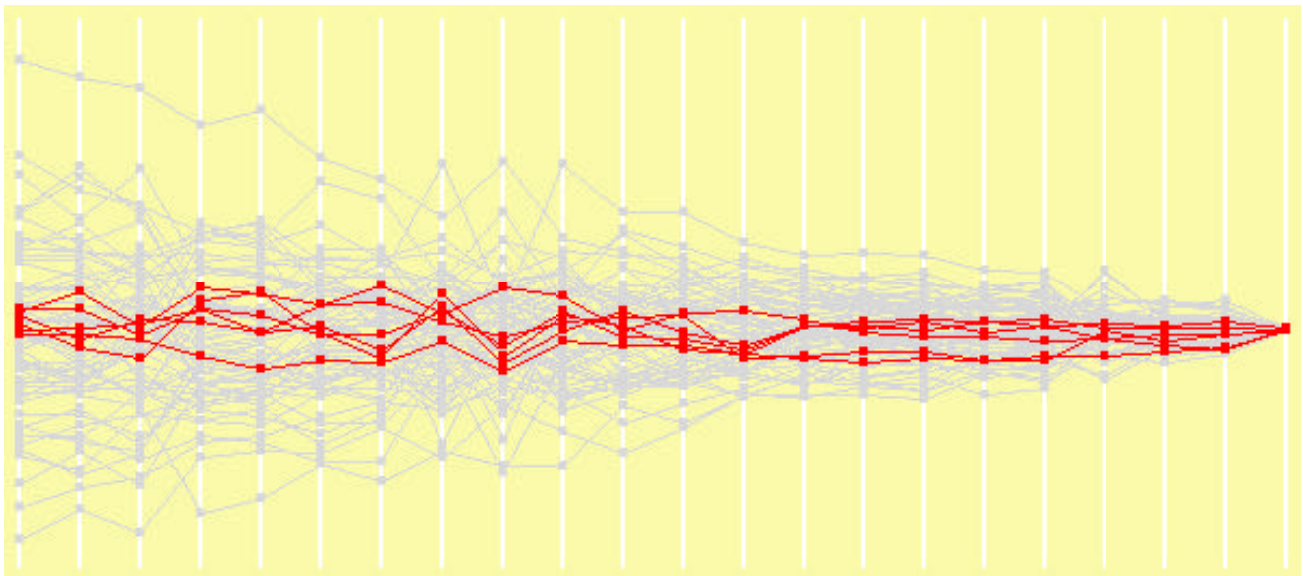


Figure 4 Residuals by model for Copenhagen Housing (ordered by declining G^2). Six cells, which are fitted well by all 22 models, have been selected.

The data has 72 cells and it is relatively easy in these circumstances to follow the gradual concentration of residual values, because all axes have been given a common scale and because the axes have been ordered. You can see which cells are always underfitted or overfitted and which seem to be fitted well by all models (as are the 6 cells highlighted in the figure). What happens in the case of the Cirrhosis data? In the case of censored survival data such as these, using the standard residual, observed time - expected time, is not appropriate. Since the observed time for a censored case is simply the time of last contact, the residual is not interpretable as an "error". Also, we are not interested in predicting event times, rather the goal is to

estimate the probability that an event has happened in a given time period (e.g. in the first five years). Because of this, residuals have been developed for survival models which have a different interpretation. The standard approach is to use a counting process formulation, where the residuals are defined as "observed number of events at time t" minus "expected number of events at time t". In our data set, onset of cirrhosis is the only event, so the first number is 0 or 1. The expected number of events can range theoretically from 0 to infinity since the model is not limited to single events, and so the residuals could range from -infinity to 1. In practice there will be few high negative residuals. Censored cases (with no event) must have non-positive residuals. We use deviance residuals, a scaled version of the residual, which retains the sign of the unscaled version and has better distributional properties.

Figure 5 shows the deviance residuals for the 36 models in an M-plot. This looks quite different, almost like a set of parallel lines. The most extreme positive residual is the same case for each model, not one of the 36 models fits it well. While we might have expected to find lots of line crossings suggesting that one model fitted one group of points well, while another fitted a different group well, it seems that all models fit the points pretty much the same way. It almost looks as if there is little difference between the models. Thinking back to Figure 4 the reason becomes clear: all of the 36 models are good models, which fit the data at least fairly well. There is no saturated model and neither is there a model with no explanatory variables. The M-plot will be most helpful for a full array of models, including not only ones which fit well, but ones which fit badly as well.

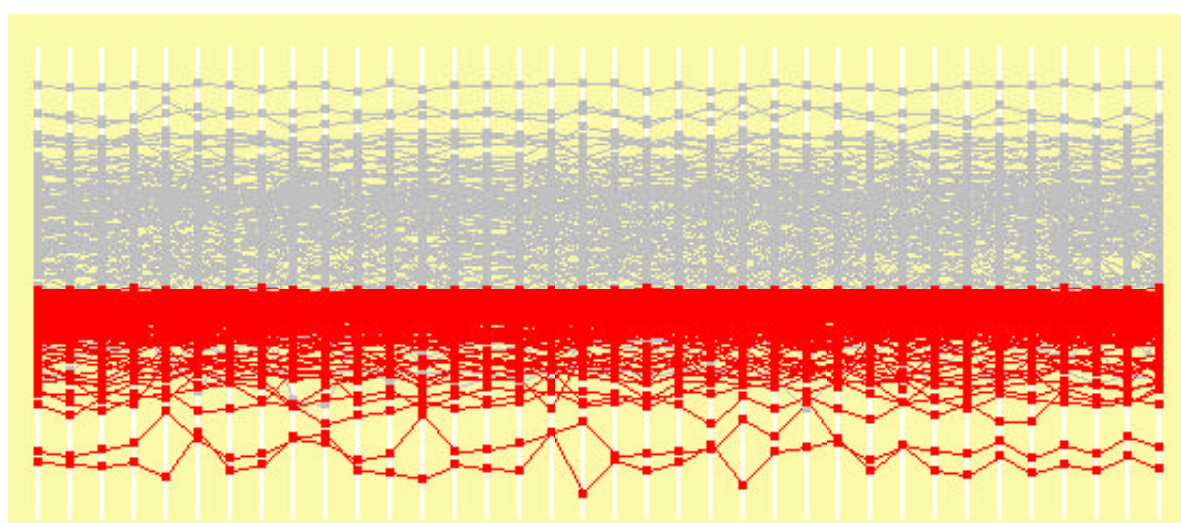


Figure 5 Deviance residuals of 36 Cirrhosis models with censored observations selected. M-plot (Querying the plot confirms that all censored case residuals are negative.)

The censored observations have been selected in Figure 5. It is a feature of survival models that all censored cases will have negative residuals, while uncensored cases will have positive residuals. It is unusual (and appears to happen here) that a few uncensored cases can have negative residuals. This arises because the model predicts more than one event for those cases (and in survival analysis only a maximum of one event is possible).

An alternative way of looking at the residuals from a set of models is offered by the C-plot (residuals by Case). There is now one axis for each case and the residuals from all models are plotted on the same axis for the same case. Lines join residuals from the same model. The data matrices behind the M and C plots are just transpositions of one another. Figure 6 shows the C-plot for the Cirrhosis data. There are 304 axes, one for each point, and they have been set to a common scale and ordered by median residual. It is very striking how little variation there is in some of the point residual distributions, suggesting that all 36 models make pretty much the same prediction for those points. Variability increases somewhat with the positive residuals, a particular feature of the survival data set.

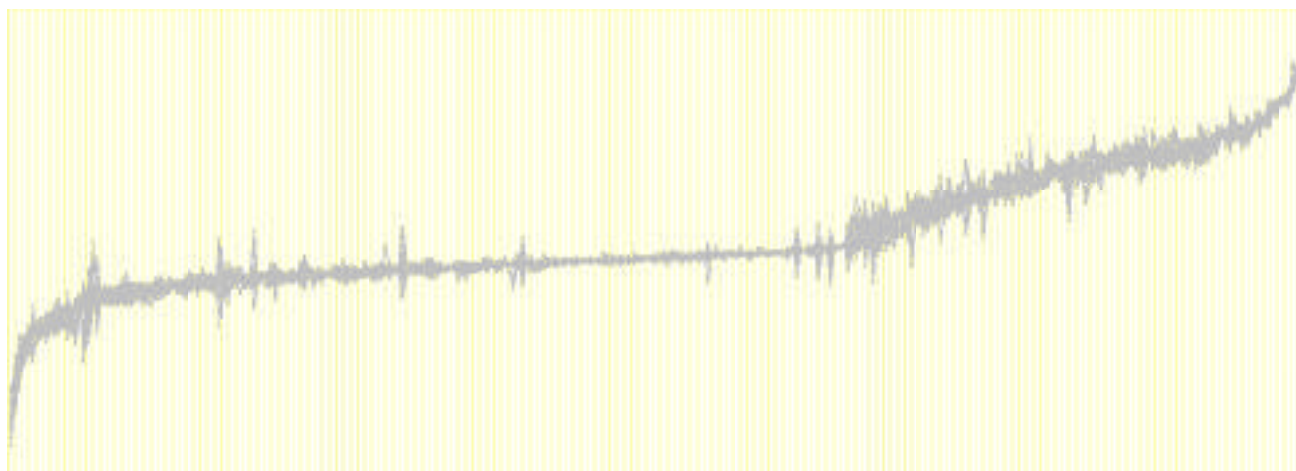


Figure 6 Parallel coordinate plot of the deviance residuals by case, ordered by median residual. C-plot

A point with a median residual far from zero and low variability (such as the point with highest positive residual referred to in the discussion of Figure 5) is fitted equally badly by all models. It can therefore not be explained by any combination of the variables that has been tried, given the other data. This suggests it is some kind of outlier. Points with high variability are predicted very differently by many of the models. Both the M- and C-plots treat the residuals equally, no matter which model they come from. Some way of taking account of how well the different models fit when examining the residuals could be helpful,

weighting them by the posterior probabilities of the models or by some other global criterion, for instance. When there are very many cases, the C-plot becomes unwieldy for the whole data set (although subsets of the data can still be profitably investigated). For large data sets we suggest a plot of the standard deviation of each case's residuals against its mean as in Figure 7.

For the Cirrhosis data set it reveals an interesting structural feature. The censored cases with high variability are those with a high negative residual, i.e. an event was expected but did not happen. The uncensored cases with high variability are those with a small residual, i.e. an event was expected and did happen. Variability of the residuals tends to increase as the model's predicted probability of an event goes up.

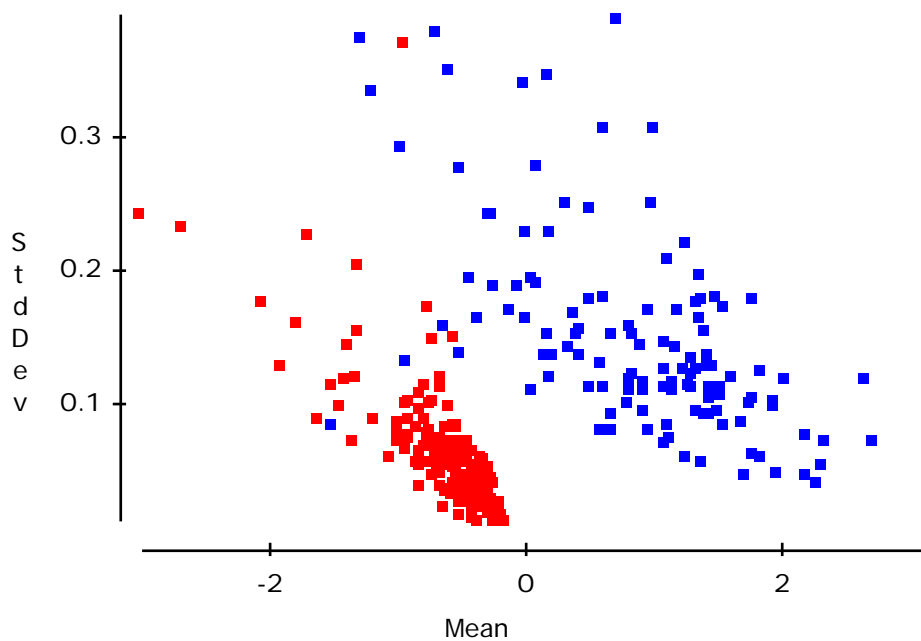


Figure 7 Scatterplot of the standard deviations and means of the deviance residuals of individual cases (censored cases are coloured red)

Conclusions

The interactive RtMC Parallel Coordinate plot family gives excellent overviews and generates thought-provoking insights:

R-plot for investigating the raw data

t-plot for exploring the importance of explanatory variables

M-plot for comparing residuals across models

C-plot for examining the distribution of residuals by individual case

Together with other interactive graphics displays, these tools provide a way of working with larger numbers of models fit to the same data set. Much progress is still to be made in Exploratory Modelling Analysis, but it is an exciting area and one where both theoretical developments and visualisation methods have important roles to play.

Software

The software used for parallel coordinates was CASSATT, which is research software for interactive graphics developed in the department of Computer-Oriented Statistics and Data Analysis at the University of Augsburg, written by Sylvia Winkler (see <http://www1.math.uni-augsburg.de> for more information and details of availability). Software for fitting the Bayesian models may be obtained from Chris Volinsky (<http://www.research.att.com/~volinsky/bma.html>).

References

- Cook, R. D., Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall
- Cook, R. D., Weisberg S. (1994). *An Introduction to Regression Graphics*. New York: Wiley
- Cook, R. D., Weisberg, S. (1999). Graphs in Statistical Analysis: Is the Medium the Message? *American Statistician*, 53(1), 29-37.
- Cox, D. R., Wermuth, N. (1996). *Multivariate Dependencies — Models, analysis and interpretation*. London: Chapman & Hall.
- De Santis, F., Spezzaferi, F. (1999). Methods for Default and Robust Bayesian Model Comparison. *International Statistical Review*, 67(3), 267-286.
- Fleming, T. R., Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
- Hoeting, J., Madigan, D., Raftery, A.E., Volinsky, C. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4) 382--417. Corrected version available at <http://www.stat.washington.edu/www/research/online/hoeting1999.pdf>
- Hoeting, J., Raftery, A.E., Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *CSDA*, 22(3), 251-270.
- Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots. *Metrika*, 51(1), 11-26.
- Inselberg, A. (1998). Visual Data Mining with Parallel Coordinates. *Computational Statistics*, 13(1), 47-63.

Kass, R. E., Raftery, A.E. (1995). Bayes Factors. *JASA*, 90, 773-795.

Lindsey, J. K. (1997). *Applying Generalized Linear Models*. New York: Springer.

Madigan, D., Raftery, A. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *JASA*, 89(428), 1535-1546.

Raftery, A. E., Madigan, D., Volinsky, C.T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. In J. Bernardo Berger, A., Dawid, A., Smith, A. (Ed.), *Bayesian Statistics 5*, (pp. 323-349). Oxford University Press.

Therneau, T. M., Grambsch, P.M. (2000). *Modeling Survival Data*. New York: Springer.

Unwin, A. R., Hawkins, G., Hofmann, H., and Siegl, B., Interactive Graphics for Data Sets with Missing Values - MANET. *Journal of Computational and Graphical Statistics*, 5(2) (1996) 113-122.

Venables, W. N., Ripley, B.D. (1999). *Modern Applied Statistics with S+* (3rd ed.). New York: Springer.

Wegman, E. J. (1990). Hyperdimensional Data Analysis using Parallel Coordinates. *JASA*, 85, 664-675.