A joint newsletter of the Statistical Computing & Statistical Graphics Sections of the American Statistical Association

Statistical COMPUTING & GRAPHICS

Word from Computing Section Chair

JOHN MONAHAN
COMPUTING SECTION CHAIR

Statistics as a Science: A Brief Rant

Simulation and Publishing Standards

In my last contribution to this Newsletter, I discussed the challenge of teaching simulation in our graduate computing course, and lamented at how poorly we statisticians practice what we preach. Since my students are still working on their assignments I cannot give a comprehensive report. Nonetheless, the students' questions appear to arise from two sources: 1) grasping the concept of statistical analysis of a statistical analysis, and 2) paying attention to the relatively simple big picture experimental design and analysis when their interest lies on the methodology being compared.

In the meantime, I have had conversations with colleagues about including the teaching of simulation

Continued on page 2 ...

EDITORIAL NOTE

What's Inside this Issue

We are excited to have some terrific contributions in this issue of the newsletter. Our feature article is from the AT&T KorBell team (Robert Bell, Yehuda Koren, and Chris Volinsky) who were just announced as winners of the first part of the $1,000,000 Netflix competition to improve the Netflix Cinematch movie recommendation system. The winning KorBell entry was a data mining ensemble 'on steroids', and the authors have written an edge-of-your- seat article including some high drama on the final day of the competition.

Continued on page 3 ...

Word from Computing Chair, cont. from page 1...

as a methodological tool in our undergraduate courses. Often we find ourselves in situations where assumptions undergirding our statistical tools are not met. We may be limited by the audience to standard statistical methodology, or a situation where discovering a new statistical method may not be warranted by time or resources. To illustrate, my wife (who works for the State of NC) is trying to forecast growth in the yearly number of participants in a program at the county level where the funding is erratic -- funding may be unavailable for a county for some months of a year and the mechanism might be modeled as the observed number being a random fraction of the potential participants. For her audience, the statistical tools must be easily explained and justified: means, simple linear regression with time, or some simple weighted average. Here, simulation experiments can be performed to find which forecasting method works best under different noise mechanisms.

In another direction, some colleagues of mine have been discussing establishing departmental standards for simulation experiments in doctoral theses. Four major points suggested by my colleagues include the following:

• Explain how the data were generated so that the experiment could be replicated.

• Explain the experimental design, including any pairing/blocking arising from the comparison of different methods on the same set of data.

• Present statistical justification of the interpretation of the results: tests, standard errors, etc.

• Follow proper scientific presentation guidelines

There are some computational aspects to some of these: 1) If a method of analysis uses random numbers, care must be taken to either generate all of the data beforehand, or to use a different generator or seed for data generation and analysis, 2) Some statistical software struggle to handle thousands of blocks. Not included in this list are some purely computational issues:

• Starting values for searches or iterative methods.

• Describing the algorithm or providing open-source code.

Many statistical methods involve search methods for finding nonlinear least squares estimates or MLE's or M-estimates. These search algorithms require starting values, and, more importantly the starting values may have important effects on the performance of a statistical method for a number of reasons. For something as basic as Newton's method, if there are multiple optima, the domain of attraction for the starting value dictates the optimum that the search algorithm finds. A pet peeve of mine is the use of the true values from which the data were generated as starting values for the search. If starting values from a user who does not know the true would entail different domains of attraction, then the choice of starting values become part of the statistical method. Moreover, from my students' recent questions, I have realized that starting values are not discussed in the literature in any general fashion, only in terms of a specific statistical method or nonlinear function.

A more problematic issue is the detailed description of the methodology. Good science dictates that another scientist must be able to replicate the experiment, and that principle leads to clearly specifying the tools that were used. For packaged routines, or commercial software, this may mean specifying as 'nls' from Version 2.5.1 or PROC GLM from Version 9.1.3 under XX operating system. But for user-written code, this must mean including the code that was used. But good science may lead to conflict when that code may have commercial value. As we consider the inclusion of code as a requirement for a departmental thesis protocol, we must pay attention to university technology transfer guidelines. I suggest here that we should also consider whether our profession ought to promulgate guidelines for publication of simulation experiments. And that raises questions for which I have no answers: How should we handle the situation where two scientists argue that each has the better method, where neither will share their code with the other?

Should our journals publish as science something that cannot be replicated? Should we sacrifice progress at the altar of science because someone can make some money? Should the public (read taxpayer) fund research that will not become part of the public domain?

Yes, I would like to begin a conversation on these issues, even if that means making trouble.

John Monahan, November 2007

Editorial Notes, cont. from page 1....

Andreas Krause Michael O'Connell

We have two other highly substantive articles in the newsletter. Tingling Song, Cameron Coffran and Knut Wittkowski from University of Rockefeller describe an innovative grid computing approach to non-parametric analysis of multivariate ordinal data with application to gene expression and epistasis screening. Joachim Moecks and Walter Koehler from the Bioscience Club of Heidelberg and Baseline Statistical Solutions in Mannheim, examine time-to-event data in context of population evolution charts.

Our conference round up includes reviews of JSM 2007 in Salt Lake City and a preview of the JSM 2008 Denver program by Program Chairs Wolfgang Jack (Computing) and David Hunter (Graphics). We also have reviews of the 2007 UseR! Conference in Ames, IA, the S-PLUS 2007 Impact Conference in Atlantic City, NJ and the BASS 2007 conference in Savannah, GA.

Our website has been newly revamped, thanks to Hadley Wickham, and it includes some presentation papers linked from the JSM 2007 conference review in the newsletter. It also includes photos from JSM 2007 and a link to a Flickr account that we set up to host our section's photos going forward. We encourage folks to send Hadley <h.wickham@gmail.com> presentations, photos, links and related materials re. their 2007 or 2008 JSM work for linking in to our new website.

This newsletter introduces Michael O'Connell as the Computing Section editor, joining Graphics Section editor Andreas Krause. Michael is Director, Life Sciences at Insightful Corp. He has a wide range of statistical computing interests including mixed models, calibration and non-parametric regression, and is currently active in statistical graphics and predictive modeling of drug safety in clinical trials. Michael and Andreas have worked together in the past including presenting a joint workshop at the 2006 Deming Conference on statistical graphics. Please feel free to contact Michael or Andreas on any aspect of the newsletter. They are always looking for content...

We want to thank Juana Sanchez for her tremendous work on the newsletter over the past 2 years and for her help in transitioning to us for this issue.

Merry Christmas and Happy New Year to everyone.

Andreas Krause
akrause@pharsight.com

Michael O'Connell
moconnell@insightful.com

Visit our newly revamped web site for updates on news and events

http://www.stat-computing.org

Juana Sanchez with Jeff Solka at JSM 2007. Thanks for the work on the newsletter Juana !

Feature Article

CHASING $1,000,000: HOW WE WON THE NETFLIX PROGRESS PRIZE

Robert Bell, Yehuda Koren, and Chris Volinsky
AT&T Labs – Research
{rbell,yehuda,volinsky}@research.att.com

1. Summary

In October 2006, Netflix Inc. initiated a contest designed to improve movie recommendations for customers of its popular online movie rental business. It released more than 100 million customer-generated movie ratings as part of the Netflix Prize competition. The goal of the competition is to produce a 10 percent reduction in the root mean squared error (RMSE) on a test data set of customer ratings, relative to the RMSE achieved by Cinematch, Netflix' proprietary movie recommendation system. Netflix promises a prize of $1,000,000 to the first team to reach the 10 percent goal (http://www.netflixprize.com/).

As might be expected, the allure of $1,000,000 has drawn much interest. Over the first year, more that 2500 teams from dozens of countries submitted entries. By most indications, participation has been greatest among people who identify themselves as computer scientists, with relatively less participation by the statistics community.

Although no team had reached the 10 percent improvement level after one year, Netflix awarded the first annual progress prize of $50,000 to the KorBell team of AT&T Labs-Research, which achieved the greatest improvement at that time, 8.43%. The KorBell team includes a computer scientist and two statisticians.

Figure 1 shows progress of the leaders over the first 12 months relative to the performance of Cinematch. To avoid clutter, the figure only shows leads that persisted for at least 48 hours. Within one month, the leader was halfway to the goal of a ten percent improvement, and two-thirds of the way after three months. Several different teams have had the lead, including academics, employees of consulting firms, industrial statisticians, computer scientists, and even undergraduates. Since summer 2007, progress has been slower and more erratic, but after a flurry of activity leading up to the progress prize, our "KorBell" team (originally known as "BellKor") squeaked by with the best score.

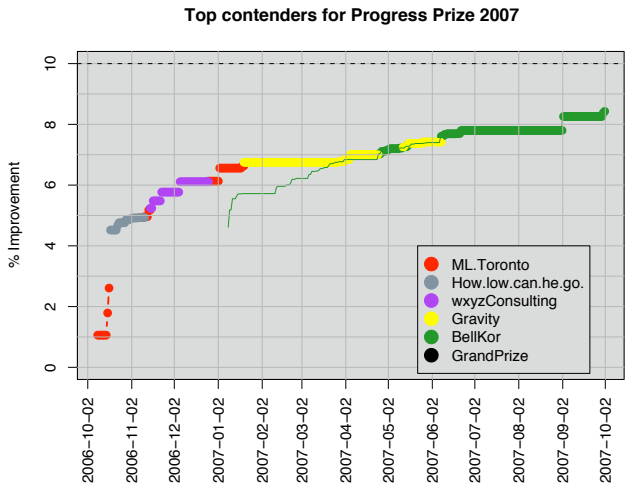


Figure 1. Scores of the leading team for the first 12 months of the Netflix Prize. Colors indicate when a given team had the lead. The % improvement is over Netflix' Cinematch algorithm. The million dollar Grand Prize level is shown as a dotted line at 10% improvement. The thin line represents the progress of our team (BellKor/KorBell).

Our winning entry was a linear combination of 107 separate sets of predictions [5]. These prediction sets used a variety of methods from the field of collaborative filtering, a class of methods that analyze past user behavior to infer relationships among items and to inform item recommendations for users. Many of the techniques involved either nearest neighbor methods or latent factor models. We found it was important to utilize a variety of models that complement the shortcomings of each other.

In addition, we developed several innovations that improved existing collaborative filtering methods, notably:

- A neighborhood-aware factorization method that improves standard factorization models by optimizing criteria more specific to the targets of specific predictions [4].

- Integration of information about which movies a user rated into latent factor models for the ratings themselves, adapting techniques from [9,10].
- A new method for computing nearest neighbor interpolation weights that better accounts for interactions among neighbors [2,3].
- New regularization methods across a variety of models, including both neighborhood and latent factor models.

Section 2 describes the Netflix Prize competition, the data, some of the resulting challenges, and the exciting race to the finish. Section 3 presents some basics of collaborative filtering. Section 4 discusses the need to utilize multiple, complementary models. Section 5 summarizes a few of the innovations developed in the process. Section 6 draws some conclusions about lessons learned from the competition.

2. The Netflix Prize Competition Structure.

Netflix published a comprehensive data set including more than 100 million movie ratings that were performed by about 480,000 users on 17,770 movies. Each rating is associated with a date between mid 1999 and the end of 2005. Ratings are integers from one (worst) to 5 (best) [6].

The contest was designed in a training-test set format that is illustrated in Figure 2. A Hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user (or fewer if a user had not rated at least 18 movies over the entire period). The remaining data made up the Training set. The Hold-out set was randomly split three ways, into subsets called Probe, Quiz, and Test. The Probe set was attached to the Training set, and labels (the rating that the user gave the movie) were attached. The Quiz and Test sets made up an evaluation set that competitors were required to predict ratings for. Once a competitor submits predictions, the prizemaster returns the root mean squared error (RMSE) achieved on the Quiz set, which is posted on a public leaderboard (http://www.netflixprize.com/leaderboard).

This unusual design helped to address some interesting qualities of the data. By design, the Hold-out set has a different distribution than the training set. Since it is constructed of the last nine ratings for each user, the distribution of time is quite different than in the Training set. In addition, the number of ratings per user is capped at nine, and so it does not display the skewness of movies-per-user that we see in the data at large.

Designating a Probe set facilitates unbiased estimation of RMSE for the Quiz/Test sets, even though the Training and Quiz/Test sets come from different distributions. Since submissions to the competition can only be done once per day, this Probe set allows for a tighter feedback loop for evaluation of promising models.

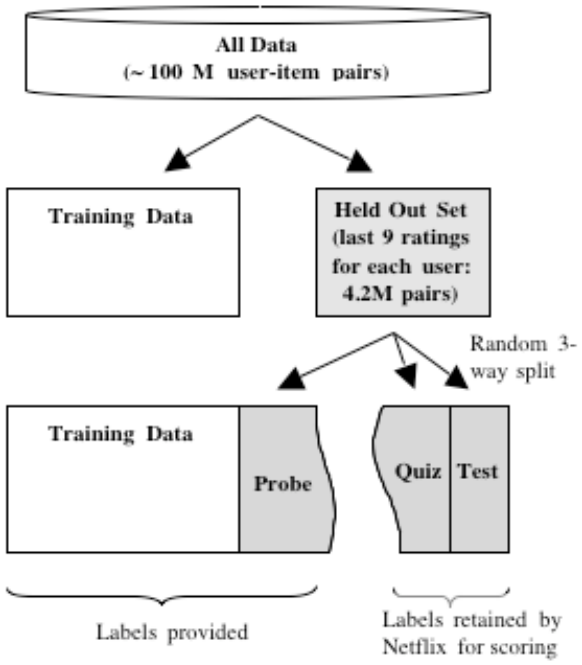


Figure 2: Structure of the Netflix Prize competition.

Also, splitting the remaining Hold-out data into Quiz and Test means that the competitor never knows which observations resulted in the achieved RMSE. Ultimately, the winner of the prize is the one that scorers best on the Test set, and those scores are never disclosed by Netflix. This precludes clever systems which might "game" the competition by learning about the Quiz set through repeated submissions.

Data Characteristics. The labeled Training/Probe data consist of more than 100 million 4-tuples: userid, movieid, rating, and date. Characteristics of the data

combine to pose a large challenge for prediction. Some of these characteristics include:

- There was a rapid rise in mean ratings starting in the first quarter of 2004 (Figure 3). We do not know the reason for this shift.
- Table 1 shows the 5 most rated and highest rated movies in the dataset. Movies with widespread consensus, such as the Lord of the Rings trilogy (average scores above 4.7), may not be very discriminative in models. Other movies, such as those shown in Table 1 with the highest variance, are polarizing, and might be more informative about an individual.
- Even with over 100 million ratings, almost 99% of the potential user-item pairs have no rating. Consequently, machine learning methods designed for complete data situations, or nearly so, must be modified or abandoned.
- The distribution of movies-per-user is quite skewed. Figure 4 shows this distribution (on a log scale). Ten percent of users rated 16 or fewer movies and one quarter rated 36 or fewer. The median is 93. But there are some very busy customers, two of which rated over 17,000 of the 17,700 movies!
- Similarly, ratings-per-movie is also skewed (Figure 5). The most-rated movie, "Miss Congeniality" was rated by over 220,000 users, almost half of the whole user base! One quarter of titles were rated fewer than 190 times and a handful were rated fewer than 10 times.

Peoples' taste for movies is a very complex process. A particular user's rating of a movie might be affected by any of countless factors ranging from the general—e.g., genre—to quite specific attributes such as the actors, the setting, and the style of background music. The wide variation in the numbers of ratings associated with individual users and movies only complicates the challenge of simultaneously modeling such a wide spectrum of factors.

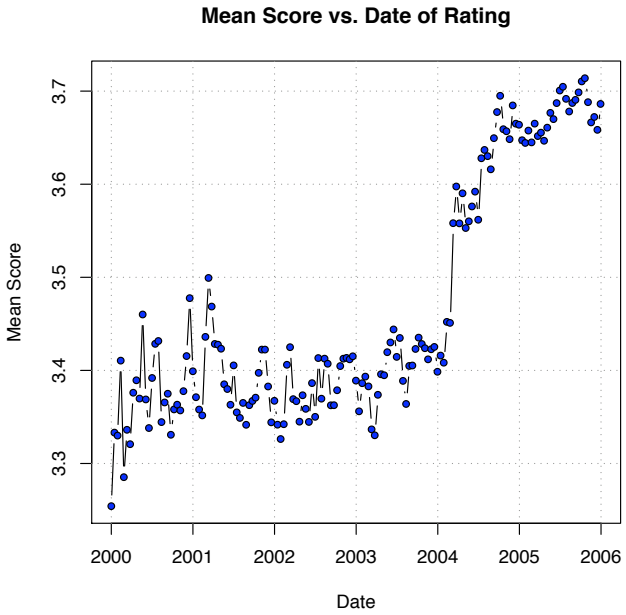


Figure 3. Shift in mean rating over time.

Most Rated	Highest Rated	Highest Variance
Miss Congeniality	The Return of the King	The Royal Tenenbaums
Independence Day	The Fellowship of the Ring	Lost In Translation
The Patriot	The Two Towers	Pearl Harbor
The Day After Tomorrow	The Shawshank Redemption	Miss Congeniality
Pirates of the Caribbean	The Empire Strikes Back	Napolean Dynamite

Table 1: Movies with the most ratings, highest ratings, and the highest variance. Highest rated and highest variance movies were out of those with at least 50,000 ratings.

About the Competition. The competition began in October, 2006 and generated wide interest from the data analysis community. The posting of top scores publicly on the website leaderboard generated a lot of enthusiasm early on, as the "horse-race" mentality set in.

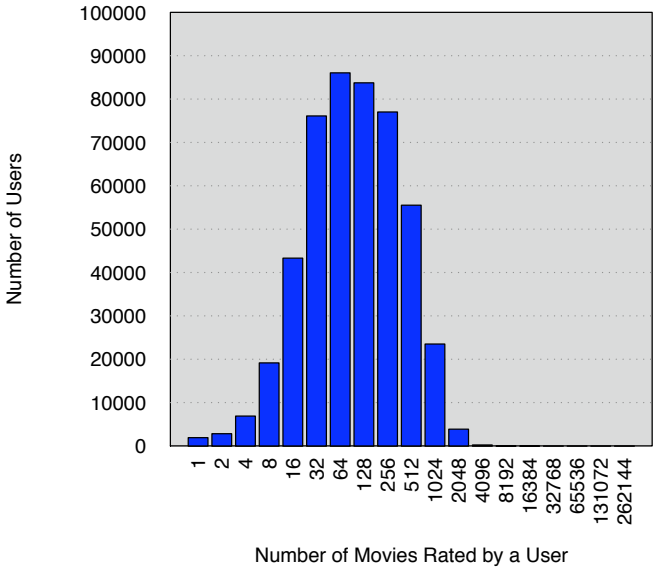


Figure 4. Histogram of number of ratings per user. Bins are on a logarithmic scale (base 2).

The competition has a very active forum where competitors join to share questions and ideas. In some cases people posted computer code that allowed fitting of particular models. The organizers are also active on this forum, answering questions and responding to concerns.

Also, there was a workshop organized at the KDD data mining conference in August 2007, where most of the top teams presented papers on their work. The overall spirit of collaboration and openness has played a key role in the enthusiasm for the competition and in its success.

Our "KorBell" team (originally known as "BellKor") held the lead for several months leading up to the one year anniversary of the competition. Nonetheless, the last hours of the competition held much drama. The progress prize was scheduled to go to the leading team as of midnight October 1, 2007. As of 9:00 P.M. on September 30, BellKor held a seemingly comfortable

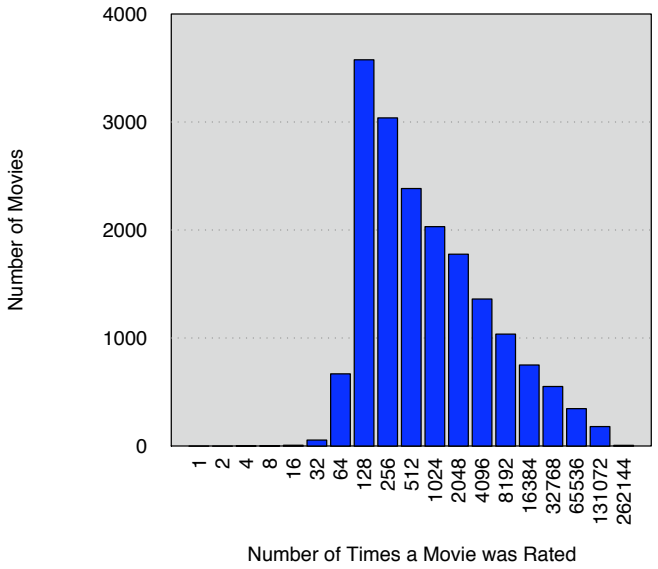


Figure 5. Histogram of number of ratings per movie. Bins are on a logarithmic scale (base 2).

lead at 8.28%, with two other teams at 8.02% and 7.99% and no one else above 7.64%. However, the contest rules allow (encourage) cooperation among teams. That evening, the fifth and sixth place teams united (averaging their previous best submissions, perhaps) to take second place at 8.07%. Later that evening, both BellKor and a new team -created by the union of the second and third place teams - both submitted entries (within 71 seconds of each other), resulting a virtual tie at 8.38%! Because each team is only permitted one submission per 24 hour period, the leaderboard remained quiet until late on October 1. In the last hour, both teams submitted their final predictions, and BellKor (now reconstituted as KorBell) posted an 8.43% improvement to win the progress prize.

3. Collaborative Filtering

Recommender systems analyze patterns of user interest in items or products to provide personalized recommendations of items that will suit a user's taste. Because good personalized recommendations can add another dimension to the user experience, e-commerce leaders like Amazon.com and Netflix have made recommender systems a salient part of their web sites.

Broadly speaking, recommender systems use either of two strategies. The *content based* approach profiles each user or product, creating descriptive attributes to associate users with matching products. For example, a movie profile might contain variables for its genre, the participating actors, its box office popularity, etc. User profiles could include demographic information or answers to a suitable questionnaire. Of course, content based strategies require gathering external information that might not be available or easy to collect. Although the contest rules did not prohibit the use of content data, it did not allow competitors to use data gathered through illegal scraping of commercial web sites.

For the Netflix data, we focused on an alternative strategy, known as *Collaborative Filtering* (CF), which relies only on past user behavior—e.g., their previous transactions or product ratings. CF analyzes relationships between users and interdependencies among products in order to identify new user-item associations.

Nearest Neighbor Methods. The most common CF tools are nearest neighbor methods, which predict ratings for unobserved item-user pairs based on ratings of similar items by the same users [11]. For example, suppose that r_{ui} is the unobserved rating of item i by user u. Then if $N(i;u)$ is a set of items similar to i that were rated by u (typically no more than 20 to 50 items), then the nearest neighbor prediction for r_{ui} might be

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in N(i,u)} s_{ij}(r_{uj} - b_{uj})}{\sum_{j \in N(i,u)} s_{ij}},$$

where b_{ui} is a baseline prediction for r_{ui}, and s_{ij} is a measure of similarity between items i and j [8].

The item similarities, denoted by s_{ij}, play a central role here, usually serving both to select the neighbors and to weight the mean. Common choices are the Pearson correlation coefficient and the closely related cosine similarity. These similarity scores depend on having adequate "overlap" (i.e., users that have rated the same movies). Having a large dataset like the Netflix data makes a huge difference in the quality of the similarities.

It is important to use the baseline values to remove item- and user-specific biases that may prevent the model from revealing the more fundamental relationships. Prior methods often take b_{ui} as the mean rating of user u or item i. In [2,3], we offered a more comprehensive approach to these baseline estimates.

Alternatively, one might calculate similarities between users instead of movies and use ratings of the target item by similar users [4,8].

Merits of the neighborhood-based approach that have made it popular include its intuitiveness, absence of the need to train and tune many parameters, and the ability to easily explain to a user the reasoning behind a recommendation.

Latent Factor Models. An alternative CF approach has the more holistic goal to uncover latent features that explain the observed ratings. The *factorization* method is based on singular value decomposition (SVD). Using this method we associate each user u with a user-factors vector p_u and each movie with a movie-factors vector q_i. Predictions are inner products of the form $\hat{r}_{ui} = p_u^T q_i$. In theory, these factor vectors capture the f most prominent features of the data, leaving out less significant patterns in the observed data that might be mere noise.

Estimation requires some form of regularization to avoid over fitting. One approach minimizes the regularized cost function:

$$\sum_{u,i} (r_{ui} - p_u^T q_i)^2 + \lambda(\|p_u\|^2 + \|q_i\|^2),$$

where the summation is over all pairs (u,i) for which r_{ui} is observed. The regularization parameter λ prevents over fitting; a typical value is $\lambda = 0.05$. We minimized this function by an alternating least squares scheme (Sec. 5.1 of [3]). Others have used gradient based techniques [7].

Restricted Boltzmann machines (RBM) provide another model based on estimating latent factors. An RBM is a stochastic artificial neural network that generates the joint distribution of users, movies, and ratings. It contains two layers of units: (1) a visible layer corresponding to all observed ratings, and (2) a hidden layer that models the various users' tastes. Weighted links associate each hidden and visible unit [10].

4. Utilizing a Complementary Set of Models

None of these models by themselves could get us to the top of the leaderboard. Instead, our best results came from combining predictions of models that com-

plemented each other. While our winning entry, a linear combination of many prediction sets, achieved an improvement over Cinematch of 8.43%, the best single set of predictions reached only 6.57%. Even that method was a hybrid based on applying neighborhood methods to results of a restricted Boltzmann machine. The best improvement achieved purely by a latent factor model was 5.10%, and the best pure nearest neighbor method was further behind [5].

We found that it was critically important to utilize a variety of methods because the two main tools for collaborative filtering—neighborhood models and latent factor models—address quite different levels of structure in the data.

Our best neighborhood models typically used 20 to 50 neighboring movies for any single prediction, often ignoring the vast majority of ratings by the user of interest [2]. Consequently, these methods are unable to capture the totality of weak signals encompassed in all of a user's ratings.

Latent factor models are generally effective at estimating overall structure that relates simultaneously to most or all movies. However, these models are poor at detecting strong associations among a small set of closely related movies such as *The Lord of the Rings* trilogy, precisely where neighborhood models do best.

While, in theory, either methodology might be extended to encompass the full spectrum of relationships, such an effort seems doomed to failure in practice. Finding niche clusters of items such as Hitchcock films or boxing movies might require thousands of latent factors. And even so, those factors might be swamped by the noise associated with the thousands of other movies. Similarly, for neighborhood methods to capture the degree that a user prefers action to dialogue may require a very large number of neighbors, to the exclusion of estimating the user's interest in other relevant features.

5. Improving Existing Methods

In this section, we summarize a few innovations on standard collaborative filtering methods.

Neighborhood-Aware Factorization. As noted in the previous section, factorization takes a very high level view of the relationship among items. But, for predicting the rating of a specific movie, certainly some movies are likely to be essentially uninformative. For example, for predicting how a man rates *Lethal Weapon*, it is probably irrelevant whether his kids liked *The Muppet Movie*.

Trying to make factorization more relevant to the target movie-user pair, we introduced *adaptive user factors* that model the behavior of a user within the neighborhood of the target movie [2]. We allow user factors to change according to the item that is being predicted. Specifically, we replace p_u with an item-specific user factor vector $p_u(i)$ that minimizes

$$\sum_{j \in N(u)} s_{ij}[r_{uj} - p_u(i)^T q_i]^2 + \lambda \left\| p_u(i) \right\|^2 ,$$

where a nonnegative similarity measure s_{ij} is used to overweighting movies similar to i in the estimation of $p_u(i)$. Here, $N(u)$ is the set of all movies rated by user u.

Accounting for Which Movies were Rated. Further improvement was achieved by explicitly integrating information about *which* movies a user rated. The idea is that someone who never watches sci-fi movies might be less likely to rate *Lord of the Rings* highly than someone who has watched several other sci-fi movies, even if that person has rated those movies somewhat low. By changing perspectives to whether or not the movie was rated, we have changed the nature of the data itself. Now, instead of having a user-movie rating matrix which is mostly missing values, we have a binary matrix (rated or not rated) that is complete.

We adopted a principle from Paterek's NSVD method [9], which refrains from explicitly parameterizing each user's ratings, but rather models each user based on the movies that he/she rated. Each movie i is associated with two movie-factors vectors q_i and x_i. In place of the user factors vector p_u, user u is represented by the weighted sum: $\left(\sum_{j \in N(u)} x_j\right) / \sqrt{|N(u)|}$, so r_{ui} is predicted as: $q_i^T \left(\sum_{j \in N(u)} x_j\right) / \sqrt{|N(u)|}$. This suggests a straightforward way to integrate information about which movies a user rated into the basic factorization model. A rating is modeled by

$$\hat{r}_{ui} = q_i^T \left(p_u + \left(\sum_{j \in N(u)} x_j\right) / \sqrt{|N(u)|}\right).$$

The parameters are estimated by gradient descent minimization of the associated regularized squared error.

An alternative way to incorporate which movies a user rated is utilization of conditional restricted Boltzmann machines (RBM) [10], which achieve a similar accuracy to the basic factorization model described above.

Our positive experience of integrating which movies a user rated into latent factor models may have an even larger payoff for real life systems that can access wide ranging implicit feedback such as purchase/rental history and browsing patterns. The most challenging part of the Netflix data proved to be the many users who provided very few ratings. We expect that many of those users do have a rich rental history that could allow a combined recommender system to make accurate, personalized recommendations for them, or even for customers who provided no ratings at all.

Interpolation Weights for Neighborhood Models. Standard neighborhood-based methods raise some concerns:

- The similarity measure s_{ij}, which directly defines the interpolation weights, is arbitrary. We could not find any fundamental justification for the various choices in the literature.
- Previous neighborhood-based methods do not account for interactions among neighbors. Each similarity between an item i and a neighbor $j \in N(j;u)$ is computed independently of the content of $N(j;u)$ and the other similarities: s_{ik} for $k \in N(j;u)$. For example, suppose that a neighbors set contains three movies that are highly correlated with each other (e.g., sequels such as *Lord of the Rings 1-3*). An algorithm that ignores the similarity of the three movies when determining their interpolation weights, may end up essentially triple counting the information provided by the group.
- By definition, the interpolation weights sum to one, which may cause over fitting. If an item has no useful neighbors rated by a particular user, it would be best to ignore the neighborhood information, staying with the current baseline estimate. Nevertheless, the standard neighborhood formula uses a weighted average of ratings for these uninformative neighbors.
- Neighborhood methods may not work well if variability differs substantially among neighbors.

To overcome these problems, we replaced the weighted average by a more general weighted sum

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in N(i,u)} w_{ij}(r_{uj} - b_{uj}) ,$$

which allows downplaying neighborhood information when lacking informative neighbors.

We learn interpolation weights by modeling the relationships between item i and its neighbors through a least squares problem:

$$\min_w \sum_{v \neq u} \left[r_{vi} - b_{vi} - \sum_{j \in N(i,u)} w_{ij}(r_{vj} - b_{vj}) \right]^2 .$$

The major challenge is to cope with the missing values, which we do by estimating all inner products between movie ratings [3]. This scheme provides interpolation weights that are derived directly from the ratings residuals ($r_{vj} - b_{vj}$), not based on an arbitrary similarity measure. Moreover, derivation of the interpolation weights explicitly accounts for relationships among the neighbors.

Our experiments showed that this scheme significantly improved accuracy relative to that based on more standard interpolation weights without a meaningful increase of running time [3]. Additional accuracy gains are possible by combining the local view of nearest neighbor methods with the higher level view of latent factor model. Accordingly, we can use predictions from a factorization or other latent factors model as the baseline estimates $\{b_{ui}\}$ for our nearest neighbor method.

Regularization. Regularization plays a central role in all of our methods. All the models described above include massive numbers of parameters, many of which must be estimated based on a small number of ratings (or pairs of ratings). Consequently, it is critically important to avoid over fitting the training data. We use shrinkage to improve the accuracy of similarity scores utilized for nearest neighbor selection [4]. Shrinkage of estimated inner products is essential to successful estimation of interpolation weights for the nearest neighbor model described above [2].

Regularization plays an even more critical role for factorization models, which might use 60 or more factors. Without any regularization, predictive accuracy for test data would begin to decline after about two factors. With proper regularization, accuracy continues to improve (very slowly) even beyond 60 factors.

Our factorization models typically use ridge regression for regularization. Although we have not fit formal Bayesian models, many of the methods have features that can be justified on Bayesian grounds.

6. Lessons Learned

Regarding the challenge of predicting ratings, we offer three main lessons.

First, the best predictive performance came from combining complementary models. In part, this may simply reflect the practical difficulty of incorporating diverse types of structure into a single model. Of course, predictive accuracy contrasts sharply with the goal of many statistical analyses—to improve understanding of a complex system. However, when recommending a movie to a user, we don't really care why the user will like it, only that she will.

Although the winning Netflix entry combined many sets of predictions, we note that it is possible to achieve a 7.58% improvement with a linear combination of three prediction sets that combine all the key elements discussed in this article: nearest neighbor models, neighborhood aware factorization, and latent factor models that incorporate information about which movies a user rated [5].

Second, although combining multiple imperfect models appears to outperform fine tuning any single model, details do matter. Using a principled approach to optimize an otherwise ad hoc method can help to improve predictive accuracy.

Third, to detect as much signal as possible from these data requires very complex models with millions of estimated parameters. With so many parameters, accurate regularization is key. If the there is one sentence that characterizes the challenge best, it is: How can one estimate as much signal as possible where there are sufficient data without over fitting where there are insufficient data?

An interesting remaining question is how well content-based models would do compared to our collaborative filtering method? In some regards, it is amazing that the CF models perform as well as they do without such metadata!

In a less technical regard, the Netflix Prize competition demonstrates the value of making industrial-strength data sets widely available. These data and the prize have generated unprecedented interest and advancement in the field of collaborative filtering. Many of these advances have been shared, most notably in the Netflix Prize forum

(http://www.netflixprize.com/community) and in a 2007 KDD workshop [1].

Certainly, the money spurred a lot of people to take a look at the data, but we doubt that money was the main stimulus for progress. Most participants (ourselves included) probably realize that their chance of winning the grand prize is small, precisely because there is so much interest. However, that interest provides great opportunity to build on the advances of others and to provide a rigorous test of any new methodologies. For that, we thank Netflix and all the other competitors, especially those who have been active in the forum and other outlets.

A final question remains. Will someone win the $1,000,000 by reaching the 10 percent target in the next year, or ever? We expect that someone will, although it may take a couple years. In our experience, it has become harder over time to achieve much improvement. Often, the seemingly best ideas only duplicate information that is already in the mix. Nonetheless, new insights continue to arise and be shared, so hope springs eternal. Stay tuned.

References

1. ACM SIGKDD, "KDD Cup and Workshop 2007," http://www.cs.uic.edu/~liub/Netflix-KDD-Cup-2007.html
2. R. M. Bell and Y. Koren, "Improved Neighborhood-based Collaborative Filtering," *Proc. KDD-Cup and Workshop at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.
3. R. M. Bell and Y. Koren, "Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights," *IEEE International Conference on Data Mining* (ICDM`07), 2007.
4. R. M. Bell, Y. Koren and C. Volinsky, "Modeling Relationships at Multiple Scales to Improve Accuracy of Large Recommender Systems," *Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.
5. R. M. Bell, Y. Koren and C. Volinsky, "The BellKor solution to the Netflix Prize," http://www.research.att.com/~volinsky/netflix/ProgressPrize2007BellKorSolution.pdf, 2007.

6. J. Bennett and S. Lanning, "The Netflix Prize," *Proc. KDD-Cup and Workshop at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.
7. S. Funk, "Netflix Update: Try This At Home," http://sifter.org/~simon/journal/20061211.html, 2006.
8. J. L. Herlocker, J. A. Konstan, A. Borchers and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," *Proc. 22nd ACM SIGIR Conference on Information Retrieval*, pp. 230-237, 1999.
9. A. Paterek, "Improving Regularized Singular Value Decomposition for Collaborative Filtering," *Proc. KDD-Cup and Workshop at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.
10. R. Salakhutdinov, A. Mnih and G. Hinton, "Restricted Boltzmann Machines for Collaborative Filtering," *Proc. 24th Annual International Conference on Machine Learning* (*ICML`07*), 2007.
11. B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," *Proc. 10th International Conference on the World Wide Web*, pp. 285-295, 2001.

Scientific Article

POPULATION EVOLUTION CHARTS: A FRESH LOOK AT TIME-TO-EVENT DATA

Joachim Moecks [1] and Walter Koehler [2]
[1]Bioscience Club of Heidelberg, Heidelberg, Germany
[2]BaseLine- Statistical Solutions, Mannheim, Germany

Correspondence to: J. Moecks Joachim@Moecks.net

Abstract

Time-to-event data analysis methods have a wide range of applications. The classical approach focuses on the survivor function and Cox proportional hazard regression when there are covariates to assess. Population evolution charts address these methods from a different angle by focusing on the distribution of the covariate for the developing cohort over time. Covariates that are related to the target events will show a change in distribution over time. Population evolution charts need fewer assumptions than the standard approaches. The present paper presents definition and interpretation of this approach for binary covariates. All concepts and the connection to Cox regression are detailed and illustrated with real data from clinical trials..

Introduction:

Time-to-event data (TTE data) is generated in many fields of science and engineering. The methodology for analyzing these data originates in the analysis of life table records – hence the name 'survival analysis', whereas in more technically-minded environments the same methodology is known as 'analysis of failure-time-data'. We prefer the more neutral term 'event' although the methodology cannot deny its origin by considering quantities called 'survivor function' or 'hazard'. Independent of the naming, this represents an emerging field of applied statistics enjoying a wide range of applications from social science, economics, quality and reliability testing to medicine. Medicine represents probably the largest area of application, where statistical TTE analysis became a key technology of data analysis, whether in epidemiology or in clinical trials. The methods considered in this article were developed and motivated from clinical trials and will be illustrated by means of those data, but may be applied in any other field of TTE analysis.

In many clinical trials the success of therapeutic strategies is described by time variables, e.g. "time to some clinical event" or "time to response to therapy or cure". In the field of serious diseases, e.g. cancer, many studies consider clinical endpoints like "time to disease progression", "time to death (overall survival, OS)", or, in combination, "time to progression or death (progression free survival, PFS) ". The success of a therapy is evaluated by investigating the distribution of the random variable T (time-to-event) through the survivor function $S(t) = 1 - F(t) = P(T > t)$ giving the probability that the event occurs after t.

The survivor function as such is not of primary interest in most cases, but relating survival to covariates

to compare the risk, e.g. for two treatment arms, for the two sexes, for a pre-existing disease (yes or no), or for two groups defined by a threshold in a metric covariate, say. (There could also be metric covariates – but we will focus on binary cases in this article).

Population Evolution Charts (PECs) conceive the follow-up of patients as a selection process relative to the covariate distribution: suppose at study start (baseline) there are p% males in the cohort. Whenever an event occurs, a patient leaves the cohort, i.e. is selected. If the selection of males proceeds with a same base probability of p% over the whole study duration, then the selecting events are indifferent (unrelated) with respect to the covariate. However, if systematically more or fewer males than p% are selected, a relation of events and covariate can be assumed. This simple idea behind the PEC may be written a little more formally as follows.

Consider a binary baseline covariate X with observations $x_j \in \{0,1\}$ where $j \in G(0)$, the studied cohort at time $t=0$ (baseline, start of the study). The initial cohort $G(0)$ is reducted to some $G(t)$ at study time t. In order to investigate changes in the cohort with respect to the covariate X simply compare the fraction of those with $X=1$ at baseline $\bar{x}(0) = \frac{1}{|G(0)|} \sum_{j \in G(0)} x_j$ with $\bar{x}(t) = \frac{1}{|G(t)|} \sum_{j \in G(t)} x_j$, the fraction of those with $X = 1$ for the remainder cohort at time t. A plot of $\bar{x}(t)$ represents the basic Population Evolution Chart (PEC). If the events are not associated with the covariate, we expect a constant course of the PEC, i.e. the selection process selects indifferently from the two groups and thus will not change the composition. Systematic deviations from being constant support an association of X with the selection process, e.g. if $\bar{x}(t)$ tends downwards then those with $X = 1$ are at a higher risk respectively earlier risk to perceive an event. An attractive feature of PECs is that the association and its time pattern can be studied without any assumptions that might be difficult to validate.

Two examples of PECs from real clinical data are given in Figure 1. The PEC for male percentage indicates that sex is clearly related to the considered events: after about 120 days of treatment the percentage of males in the study population dropped down from 70% to 60%, the development shows a clear monotone trend (interpretations will be studied below). The green trace demonstrates how independence shows up in a PEC, here for a random covariate that by construction is not associated with the event times. In the PEC example, we have tacitly ignored the problem of censoring in time-to-event data. We will turn to the topic of censoring in the estimation section.

Setting and Notation

To avoid any technical difficulties, we will assume in the following that all quantities studied are mathematically well-behaved. In brief, these are the main 'coordinates' of the standard time-to-event terminology: The analysis is not based on the distribution function, $F(t) = P(T \le t)$, but rather makes use of the survivor function $S(t) = 1 - F(t) = P(T > t)$, the probability that the event occurs after t. The ubiquitous quantity in TTE-analysis is the hazard function $h(t)$, describing the risk for an immediate event at t conditional on survival up to t. It can be obtained from the survivor function together with the integrated hazard $H(t)$ as follows:

$$H(t) = -\log(S(t))$$

$$h(t) = {}^{d}\!/_{dt} H(t) = -{}^{d}\!/_{dt} \log(S(t)) = \frac{f(t)}{S(t)},$$

where $f(t)$ denotes the density of T (Kalbfleisch & Prentice, 1980, p. 6; Lee 1992, chapter 2).

The standard approach for studying the influence of covariates is to consider the conditional distribution of T, $P(T > t \mid X = 1)$, and fitting some (semi-) parametric model (e.g. Cox proportional hazard model, Kalbfleisch & Prentice, chapters 4, 5).

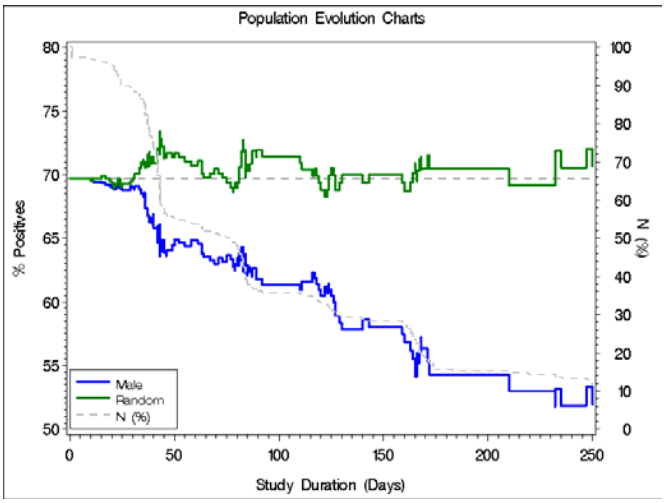


Figure 1: Population Evolution Chart Examples. Blue trace: The percentage of males in development over treatment time. Green trace: binary random covariate with no relation to events (same event times as for blue trace)

In contrast, population evolution charts emphasize the covariate and take the opposite view to the TTE-data: The PEC can be straightforwardly defined as a sequence of conditional probabilities:

(1) $\Psi_X(t) := P(X = 1 \mid T > t)$

The population evolution chart thus represents the probability for $X = 1$ for the survivors up to t.

The PEC was introduced as describing a selection process, underlined by the following derivation. Consider the distribution of the covariate X at time o (baseline): $P(X = 1)$. The simple identity

$$P(X = 1) = P(X = 1; T > t) + P(X = 1; T \le t) \ ,$$

leads, by elementary probability algebra, to

$$P(X = 1) = P(X = 1 \mid T > t) \times P(T > t) + P(X = 1 \mid T \le t) \times P(T \le t)$$

Using $S(t) = P(T > t)$, the overall survivor function, we thus get:

(2)

$$\Psi_X(t) = P(X = 1 \mid T > t) = \frac{1}{S(t)}\left[P(X = 1) - (1 - S(t)) \times P(X = 1 \mid T \le t)\right]$$

This underlines that the events can be perceived to select from the baseline distribution $P(X = 1)$, leading to a time-dependent composition of the population.

A further representation of the PEC can be seen directly from its definition as conditional probability:

(3)

$$\Psi_X(t) = P(X = 1 \mid T > t) = \frac{P(T > t \mid X = 1)}{P(T > t)} P(X = 1) = \frac{S_1(t)}{S(t)} P(X = 1)$$

Estimation

The well known complication of time-to-event data is the fact that not all events are actually observed, but the observation may have terminated at t before an event occurred (censoring), e.g. due to a fixed administrative date for the study end. For censored observations it is only known that the event did not occur up to t. The well known technique for this issue are special estimates of the survivor function, like the celebrated Kaplan-Meier graph.

In the case of no censoring, there is no estimation problem and we simply proceed as indicated in the Introduction. However, if there is some censoring, then the cohort $G(t)$ is not complete and could miss some of those censored before t, for it is unknown whether they would have survived t. Censoring usually comes with assumptions e.g. that it is random or independent of the time variable T. In most cases it is also custom to assume that censoring is not dependent on the covariates, e.g. that any demographic property of the patient or pre-existing co-morbidities would not influence the censoring. In clinical trials, we actually face two types of censoring; the pure administrative date of study end or time of interim analysis and censoring that occurs during the conduct of the study, patients dropping out voluntarily or for unknown reasons, e.g. due to insufficient effect of the control treatment (placebo) or side effects of the test treatment. A complete independence of these censored observations from the events or from covariates is usually assumed, but may be overoptimistic.

In order to address different assumptions regarding censoring, we introduce censoring formally:
Let C denote the random variate of the censoring process. The random variate pertaining to all actually

observed times (censoring and events) is denoted by A. Note that $A = \min(T, C)$. If censoring and events are independent, we get for the total process multiplicative survivor functions:

$$P(A > t) = P(T > t) \times P(C > t)$$

This property carries directly over to the Kaplan-Meier (KM) estimates of the involved survivor functions:

$$\hat{P}(A > t) = \hat{P}(T > t) \times \hat{P}(C > t)$$

Note that $\hat{S}(t) = \hat{P}(T > t)$ is the usual KM-estimate of the survivor function, while for calculating $\hat{P}(C > t)$, the role of events and censored observations is to be interchanged.

Let us now consider the PEC for the total process pertaining to A according to representation (3):

$$P(X = 1 \mid A > t) = \frac{P(A > t \mid X = 1)}{P(A > t)} P(X = 1)$$

By independence of T and C we get:

$$P(X = 1 \mid A > t) = \frac{P(T > t \mid X = 1)}{P(T > t)} \times \frac{P(C > t \mid X = 1)}{P(C > t)} \times P(X = 1)$$

The assumption of independence of covariate X and censoring C leads to $P(C > t \mid X = 1) = P(C > t)$ and hence we get:

$$P(X = 1 \mid A > t) = \frac{P(T > t \mid X = 1)}{P(T > t)} \times P(X = 1) = P(X = 1 \mid T > t)$$

In this case the PEC could be estimated based on the total process A. The simple estimate presented in the introduction also refers to the total process and thus could serve as first try in case of independent censoring:

$$(4) \quad \hat{\Psi}_I(t) = \frac{1}{|G(t)|} \sum_{j \in G(t)} x_j$$

A further way of estimation is offered by the representation from (3): Let $\bar{x}(0) = \frac{1}{|G(0)|} \sum_{j \in G(0)} x_j$ denote the baseline estimate of $P(X = 1) = \Psi_X(0)$ and let $\hat{S}(t)$, $\hat{S}_1(t)$ denote the KM- estimate of $P(T > t)$, $P(T > t \mid X = 1)$, respectively, which leads to:

$$(5) \quad \hat{\Psi}_{II}(t) = \frac{\hat{S}_1(t)}{\hat{S}(t)} \times \bar{x}(0)$$

In comparing the two estimators, we observe that (5) requires the least assumptions, since there is no need to assume that the covariate is independent of censoring. A further advantage of (5) is that we may investigate the censoring process by defining a PEC for censoring:

$$(6) \quad \hat{\Psi}_C(t) = \frac{\hat{P}(C > t \mid X = 1)}{\hat{P}(C > t)} \times \bar{x}(0) \quad .$$

The estimates derive from KM-estimates with censoring and events interchanged. Since (4) refers to the total process, censoring and events cannot be separately studied.

The two estimators of the PEC are illustrated in Figure 2: the example shows the results for a biomarker where a positive value indicates a worse medical prognosis. It can be seen that there is a strong decline in the PEC for events for the estimator II, which means that indeed positive values for this parameter leads to more events. At the same time a different development is seen for the censor CPEC. It shows a slight trend in the opposite direction, meaning that those censored are more likely to have negative values of the biomarker. The trace for the estimator I shows that we have to deal with variance for larger t, and that in particular for larger t, the estimates are affected by the censored events.

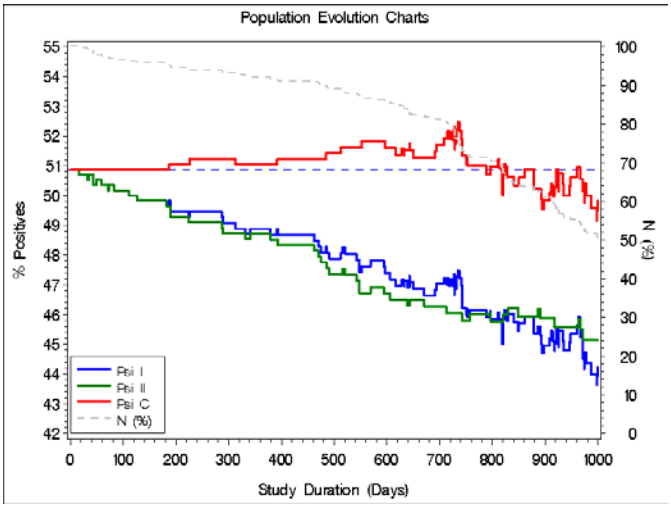


Figure 2: Estimates of the Population Evolution Chart. The real data example of a baseline biomarker. Patients with biomarker positivity are predominantly selected by the considered events. Censoring events show a weak trend in the opposite direction.

Interpretation of PECs

The basic goal of a PEC analysis is to investigate dependencies in a nonparametric way. The overall null hypothesis is the constancy of the PEC, i.e. that there is no relationship of events and covariate. Formally we have:

Ho: $\Psi_X(t) := P(X=1 \mid T>t) = \text{const.}$

The constant is the value of the PEC at t=0, i.e. $P(X=1) = \Psi_X(0)$. From representation (3) we get the equivalent hypothesis $S(t) = S_1(t)$ which is also equivalent to $S_1(t) = S_0(t)$, where $S_0(t) = P(T>t \mid X=0)$ denotes the survivor function for $X=0$.

Thus for testing the overall constancy of a PEC, we are facing a standard problem in TTE methodology, namely to compare two survivor functions. The comparison can be carried out non-parametrically e.g. by the logrank or Wilcoxon test.

In order to look a bit deeper into the guts of the PEC approach, we note that $S(t) = S_1(t) \times P(X=1) + S_0(t) \times P(X=0)$.

By taking the derivatives of the log of this relation,

$$h(t) = -\frac{d}{dt}\log\{S_1(t)\times P(X=1) + S_0(t)\times P(X=0)\} = S(t)^{-1}\{f_1(t)\times P(X=1) + f_0(t)\times P(X=0)\}$$

some algebraic rearrangement leads to:

$$(7)\quad h(t) = h_1(t)\times\Psi_X(t) + h_0(t)\times(1-\Psi_X(t))$$

This provides a useful relation for the overall hazard linking it to the PEC and pertaining subgroup hazards.

Further, we can write the PEC as a function of underlying hazards for $h_1(t) - h_0(t) \neq 0$:

$$(8)\quad \Psi_X(t) = \frac{h(t) - h_0(t)}{h_1(t) - h_0(t)}$$

The interesting case for a PEC is the deviation from constancy – with the focus on the interpretation of the degree of deviation from constancy. Taking logarithms in the representation (3) and taking subsequent derivatives leads to:

$$\log(\Psi_X(t)) = \log(P(X=1 \mid T>t)) = \log S_1(t) - \log S(t) + \log P(X=1)$$

$$(9)\quad \frac{\Psi'_X(t)}{\Psi_X(t)} = h(t) - h_1(t)$$

The dynamics in the change of the PEC is determined by a hazard difference. This relation can be transformed to involve the subgroup hazards $h_1(t)$ and $h_0(t)$:

$$(10)\quad \Psi'_X(t) = (h_0(t) - h_1(t))\times\Psi_X(t)\times(1-\Psi_X(t))$$

Since the PECs are positive functions, only the hazards determine whether the PEC increases or decreases: whenever $h_1(t)$ exceeds $h_0(t)$, the PEC decreases, and vice versa, as local instantaneous property. PECs will exhibit a monotone course if one of the hazards dominates the other for all times considered. The PECs shown in Figures 1 and 2 follow this pattern very clearly.

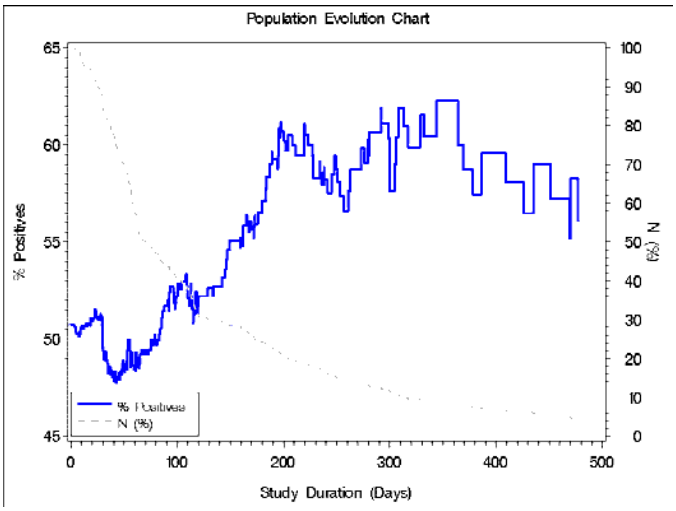


Figure 3: A non-monotonic Population Evolution Chart (real data). A turn point is visible approximately after 40 days.

The PEC in Figure 3 shows a different pattern: after an initial phase with very little difference in hazards, there was a clear drop, when the positives had more events. After about 40 days the picture changed and the negatives suffered from more events. Such a pattern fits to the idea that a positive status is associated with a bad prognosis – however, after a lag time of about 40 days, the therapy is able to change the picture creating specifically for positives a benefit from this therapy, in a clearly higher degree than for the negatives. Later, after approx. 200 days, the PEC becomes more or less constant, which could mean that those of the negatives, who have survived up to this point, do no longer suffer from their negative status.

PECs and Cox' proportional hazard model

The semi-parametric Cox proportional hazard model represents the gold standard for dealing with covariates in the TTE framework. It represents probably one of the most often used statistical approaches at all – at least in the medical context. The crucial assumption of this model is the proportionality of the hazard functions of the two groups ($X = 1$ and $X = 0$), i.e. their hazard ratio is independent of time. We will only consider the simple case of one binary covariate.

The gist of the difference between the Cox model and the PEC approach is easily stated: the Cox model looks at $P(T > t \mid X = 1)$, while the PEC features $P(X = 1 \mid T > t)$. PECs study the time development of the covariate distribution, while in the Cox model the covariate influence is independent of time. The influence of the covariate is parametrically quantified by modeling $P(T > t \mid X = 1)$ with proportional underlying hazards, without restriction of the functional form.

It is illustrative to investigate the form of a PEC if the assumptions of a Cox model hold, i.e $h_1(t) = \lambda \cdot h_0(t)$.
Then we get from (10):

$$(11) \quad \Psi'_X(t) = (1-\lambda) \cdot h_0(t) \cdot \Psi_X(t) \cdot (1 - \Psi_X(t)) .$$

This means that a PEC from a valid Cox model would always be monotone, for $\lambda > 1$ decreasing and for $\lambda < 1$ increasing. This comes as a necessary condition for the validity of the Cox model, but monotonicity of the PEC is not sufficient: from (10) we may have $h_1(t) > h_0(t)$ for all t and thus a monotone PEC, but $h_1(t) = \lambda(t) \cdot h_0(t)$ with a non-constant $\lambda(t)$, violating the proportionality assumption.

In the case of Figure 3 it is very doubtful if a simple Cox model can render reasonable results. Maybe a piecewise approach can be tried by considering the data between day 40 and 200 to be fitted by a Cox approach.

The PECs in Figures 1 and 2 show a perfect linear decline. How can this be interpreted, assuming that the Cox model is valid ? To this end, suppose that the PEC is really linear:

$$\Psi_X(t) = \alpha \cdot t + \pi_1 \quad ,$$

where the intercept is

$$\pi_1 = \Psi_X(0) = P(X = 1) .$$

Using (11) and putting $\pi_0 = 1 - \pi_1$ we find:

$$(12) \quad h_0(t) = \frac{1}{1-\lambda} \cdot \frac{\alpha}{(\alpha \cdot t + \pi_1) \cdot (-\alpha \cdot t + \pi_0)}$$

Thus, a simple parameterization of the PEC leads to an explicit functional form for Cox' baseline hazard. In order to provide a real example, the involved parameters were estimated from the data of the Figure 2 example. The results were as follows: $\pi_1 = 50.86\%$ (obtained from PEC), $\alpha = 4.88 \frac{\%}{1000d}$ (obtained by linear regression), and $\lambda = 3.24$ (obtained from Cox regression). The resulting hazard estimates according to (12) are displayed in Figure 4. The hazards are almost constant, the small embedded picture shows the baseline hazard with a magnified scale – demonstrating the slight increase of the hazard over the time span.

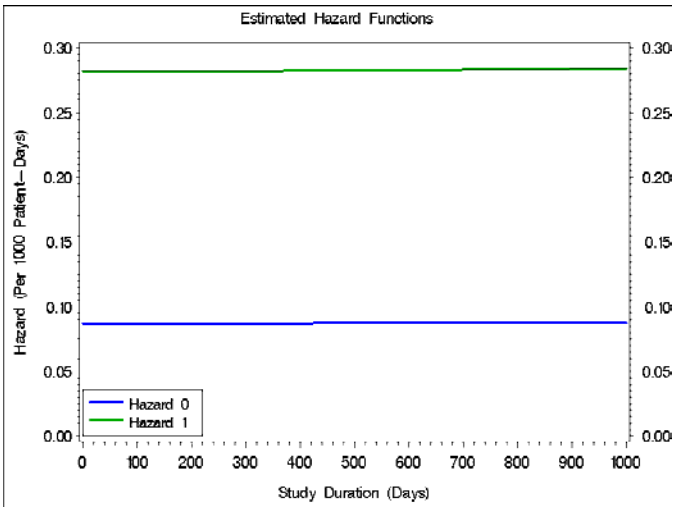


Figure 4: Estimated hazard functions according to (12) based on the data displayed in Figure 2. Blue trace: baseline hazard; green trace: hazard for biomarker-positive values. Embedded graph: baseline hazard on magnified scale.

Figure 5 displays survivor functions for the subgroups estimated by various approaches. The fit of the Cox model is reasonable (as seen from the estimated survivor functions based on $\lambda = 3.24$ (light colors) in comparison to the standard Kaplan Meier subgroup estimate. The almost exponential survivor functions estimated according to (12) provide a good compromise in these data.

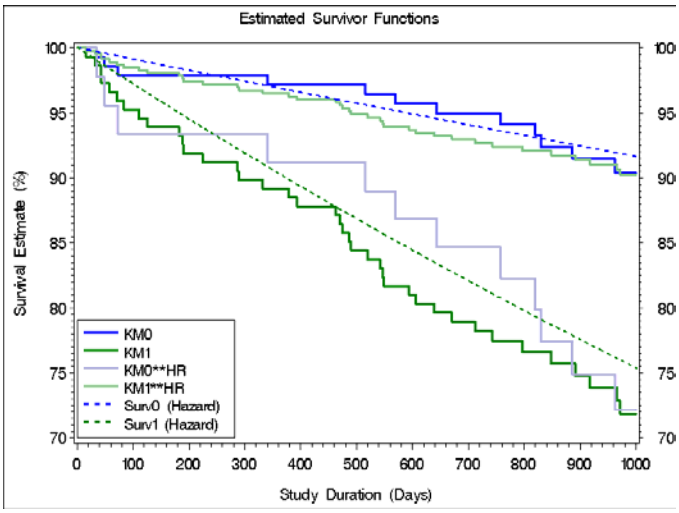


Figure 5: Estimated survivor functions based on the data displayed in Figure 2. Blue trace(KM0): Kaplan – Meier estimate for group of biomarker-negative values. Green trace(KM1): Kaplan – Meier estimate for group of biomarker positives. Light blue(): estimated survival of biomarker positives according to the Cox estimate of hazard ratio 3.24. Light green(): estimated survival of biomarker negatives according to the Cox estimate of hazard ratio 3.24. Dotted lines: estimated survivor function according to the hazard estimates of Figure 4.

Discussion

The basic idea of PECs was to consider TTE data as a selection process and to analyze the development over time of the observed cohort. This leads to quite interesting graphical displays which allow for an alternative view of the TTE data and covariates, in particular by detailing out the time pattern of the covariate influence.

The idea of PECs evolved in the year 1996, when the cardio-vascular safety of recombinant erythropoietin (EPO) for dialysis patients was analyzed in a large study pool with initially 3,111 patients from 22 clinical trials with variable follow-up of up to 4 years. The estimated course of the hazard for cardio-vascular-death showed a surprising decline over the follow-up period (see Moecks et al., 1997). Could this represent a long term benefit of the anti-anaemic therapy? Was the composition of the population starting the first year (n=3,111) comparable to those starting the fourth year (n=101) ? Cardiovascular (CV) death could select out those patients who are anyway prone to this risk, leav-

ing only patients healthy in comparison at later stages. The study pool contained studies with differing follow-up and different inclusion/exclusion criteria, with an unclear effect on the remainder cohort. Therefore, the PEC addressed the total selection process in particular for baseline covariates which implied a cardiovascular risk. The PECs revealed that there was no decrease in percentage of patients with pre-existing CV-disease or diabetes, rather these percentages even increased. The composition of patients with respect to CV risk was well comparable between first and fourth year and in between, thus supporting that the observed hazard decrease was a long term treatment benefit (Moecks, 2000).

The second practical instance where PECs proved useful came up in the context of bisphosphonate therapy for bone morbidity in cancer patients. The endpoints were bone events (bone pain, pending fractures, etc) which should be reduced by therapy compared to control. Here, premature dropouts presented issues since those with advanced morbidity under the (inefficient) control therapy showed a higher dropout rate with a downward bias in reported bone events. PECs revealed that dropout was selective, i.e. tended to select out patients with advanced morbidity. Moreover, in some studies, the dropout effect was different for active and control, showing that more morbid patients dropped out at a higher rate under control (Moecks et al, 2002). This underlines efficacy but gives standard approaches a hard time to show efficacy.

An important merit of the PEC approach in the context of clinical trials is the possibility to use this chart for censoring events: suppose that we find a covariate which clearly is associated with the target events, and in addition a PEC for censoring reveals a dependency as well. Then the censoring could violate the customary assumptions and exert a bias in the one or the other direction, e.g. as indicated in the bone event example.

The derivations of this article show that the PEC serves as a basic descriptive tool in the TTE methodology. For instance, equation (7) shows that the overall hazard and the subgroup hazards are linked through the PEC – a quite fundamental relation, linking these intuitive descriptors of the risk development.

The present article only dealt with the simplest case of a binary covariate. Displays similar to PECs can also be defined and used for metric covariates. Furthermore, a couple of more testing options exist in order to get a probabilistic evaluation of the PEC course. The focus of a PEC is on a single covariate, and the concept does not readily generalize to the multiple covariate situation. It is however possible to treat one binary covariate (e.g. treatment arms) as subgrouping variable and to compare PECs for a further covariate in one graphical display (e.g. a PEC for each treatment arm separately).

In the meantime, PECs have been applied in many data analyses, in particular in the context of diagnostic markers and biomarkers in oncology, providing many fruitful insights. A further development of this methodology appears promising for the TTE-field.

Computations of this paper were based on SAS code; a macro for the PEC-estimate $\hat{\Psi}_{II}(t)$ can be obtained from the authors upon request.

References

Kalbfleisch J.D., Prentice R.L.(1980): The Statistical Analysis of Failure Time Data. John Wiley &Sons: New York

Lee E.T.(1992): Statistical Methods for Survival Data Analysis – Second edition. John Wiley &Sons: New York

Moecks J., Franke W., Ehmer B., Quarder O. (1997): Analysis of Safety Database for Long-Term Epoetin-β Treatment: A Meta-Analysis Covering 3697 Patients. In: Koch & Stein (editors) "Pathogenetic and Therapeutic Aspects of Chronic Renal Failure". Marcel Dekker: New York, 163-179.

Moecks J. (2000): Cardiovascular mortality in haemodialysis patients treated with erythropoietin – a retrospective study; Nephron, 86; 455-462

Moecks J., Köhler W., Scott M., Maurer J., Budde M., Givens S.(2002): Dealing with Selective Dropout in Clinical Trials. Pharmaceutical Statistics, Vol 1: 119-130

Scientific Article

SCREENING FOR GENE EXPRESSION PROFILES AND EPISTASIS BETWEEN DIPLOTYPES WITH S-PLUS ON A GRID

Tingting Song, Cameron Coffran, Knut M. Wittkowski,
The Rockefeller University, New York.

Correspondence to: Knut Wittkowski
kmw@rockefeller.edu

Abstract

It is rare that a single gene is sufficient to represent all aspects of genomic activity. Similarly, most common diseases cannot be explained by a mutation at a single locus. Since biological systems tend to be non-linear and to have components that are of unknown relative importance, the assumptions of traditional (parametric) multivariate statistical methods can often not be justified on theoretical grounds. Non-parametric methods, on the other hand, tend to be computationally more demanding, so that less appropriate parametric methods are often used, even if the assumption they rely on are unlikely to be justified. This paper demonstrates how a grid of PCs can be used to enable the use of u-statistics as a non-parametric alternative for scoring multivariate ordinal data. Applications are screening for genomic profiles and genetic risk factors that correlate best with a complex phenotype.

1 Introduction

When applying statistical methods to complex phenomena, a single measure often does not appropriately reflect all relevant aspects to be considered, so that several measures need to be combined. Such problems may arise in many applications, although here we focus on analysis of SNP and gene expression microarrays.

Most multivariate methods are based on the linear model. One scores each variable individually on a comparable scale and then defines a global score as a weighted average of these scores. While mathematically elegant and computationally efficient, this approach has shortcomings when applied to real world data. Since neither the variables' relative importance nor each variable's functional relationship with the immeasurable factor of interest are typically known, the weights and functions chosen are often not easily justified. The diversity of scoring systems used attests to their subjective nature.

Even though the assumptions of the linear model regarding the contribution to and the relationship with the underlying immeasurable factor are questionable, as in genetics and genomics, it is often reasonable to assume that the contribution of a locus and the expression of each gene have at least an 'orientation', i.e., that, if all other conditions are held constant, the presence of an additional mutation or the an increase in a gene's expression is either 'good' or 'bad'. The direction of this orientation can be known (hypothesis testing) or unknown (selection procedures). Order statistics (also known as rank statistics) are well suited for such problems, yet without the simplifying assumptions of the linear model the order among multivariate data may only be partial [1,2], i.e., for some pairs of observations, the ordering may ambiguous.

Originally, we used the marginal likelihood (MrgL) principle to handle partial orderings, e.g., to assess the overall risk of HIV infection based on different types of behavior[3]. More recently, we applied this approach to assessing immunogenicity in cancer patients[4]. In short, one determines all rankings compatible with the partial ordering of the observed multivariate data and then computes a vector of scores as the average across these rankings. While this approach does not rely on questionable assumptions, the computational effort required can be prohibitive even for moderately sized samples, let alone micro arrays with thousands of SNPs or genes.

2 Methods

2.1 U Statistics

U-statistics[5] are closely related to the MrgL approach. When MANN and WHITNEY[6], in 1947, proposed their version of what is now known as the WILCOXON/MANN-WHITNEY test, it was one of the first uses of u-statistics. HOEFFDING formalized this concept in 1948[5]. Originally, observations were allowed

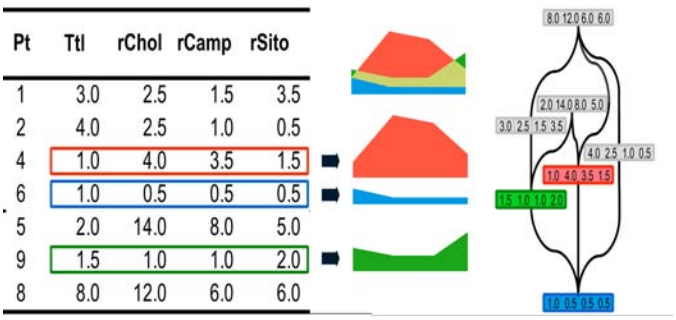
to be multivariate. When GEHAN[7], in 1965, applied u statistics to censored observations, however, he viewed them as univariate observations (x_{jk1}: time under study), accompanied by an indicator of precision (x_{jk2} = 1: event, x_{jk2} = 0: censoring), rather than as multivariate data (x_{jk1}/x_{jk2} earliest/latest time point). Thus, the potential of u-statistics for the analysis of multivariate data was not fully recognized, most likely because the computational effort to handle multivariate data was prohibitive, in general, and no algorithm was presented that would have allowed application of the method at least for small sample sizes.

U-statistics are closely related to the MrgL approach, yet are easier to compute. In either case, one does not need to make any assumptions regarding the functional relationships between variables and the latent factor, except that each variable has an orientation, i.e., that if all other variables are held constant, an increase in this variable is either always 'good' or always 'bad'.

Each subject is compared to every other subject in a pairwise manner. For stratified designs[1], these comparisons are made within each stratum only, e.g., within each sex. Even though not all pairwise orderings can be decided, all (most) subjects can be scored. With an indicator function I , one can assign a score to each subject by simply counting the number of subjects being inferior and subtracting the number of subjects being superior[8]

$$u(x_{jk}) = \sum_{j'k'} I(x_{j'k'} < x_{jk}) - \sum_{j'k'} I(x_{j'k'} > x_{jk})$$

Fig. 1 provides a graphical representation of creating a partial ordering for multivariate data. From the lattice (non-circular directed graph) on the right of Fig. 1, the main features of u-statistics are easily seen. (1) Only those pairs are linked, where the order is independent of any weights that could be assigned to the different variables. (2) Adding a highly correlated variable is unlikely to have any effect on the lattice structure. Relative importance and correlation do not even need to be constant, but may depend on the other variables.



Pt	Ttl	rChol	rCamp	rSito
1	3.0	2.5	1.5	3.5
2	4.0	2.5	1.0	0.5
4	1.0	4.0	3.5	1.5
6	1.0	0.5	0.5	0.5
5	2.0	14.0	8.0	5.0
9	1.5	1.0	1.0	2.0
8	8.0	12.0	6.0	6.0

Fig. 1. Generating a partial ordering from the blood cholesterol profiles (total cholesterol, cholestanol, camposterol, sitosterol) of seven patients. The profiles of patients 4, 6, and 9 are shown to the right of the data. The overlay above these profiles shows that #4 and #9 have a higher profile than #6, but that the pairwise order between #4 and #9 cannot be decided without making assumptions about the relative importance of Ttl and rSito (higher in #9) vs rChol and rCamp (higher in #4). The complete partial ordering of the seven patients is depicted as a lattice, with lines indicating for which pairs the pairwise order can be decided. Patients #4, #6, and #9 are highlighted.

For univariate data, all pairs of observations can be decided, i.e., the resulting ordering is 'çomplete'. For multivariate data, however, the ordering is only 'partial', in general, because for some pairs of outcome profiles the order may be undetermined. This is the case, for instance, if the first outcome is higher in one subject, but that of the second outcome is higher in the other.

2.2 *Grid architecture*

'Screening' thousands of expression profiles or epistatic sets to find the profile or set whose scores correlate best with the scores of a complex phenotype, can easily become impractical even on a fast computer or a traditional 'beowulf' style cluster. Thus, we have formed a grid of the PC work stations at The Rockefeller University Hospital. Data is uploaded to a Web front-end and then passed to a dispatcher (Grid Server) that splits the job in dozens of work units, which are then sent to the work stations to be analyzed in parallel (Fig. 2). This service, which for the first time makes u-statistics for multivariate data more widely available, can be accessed through muStat.rockefeller.edu .

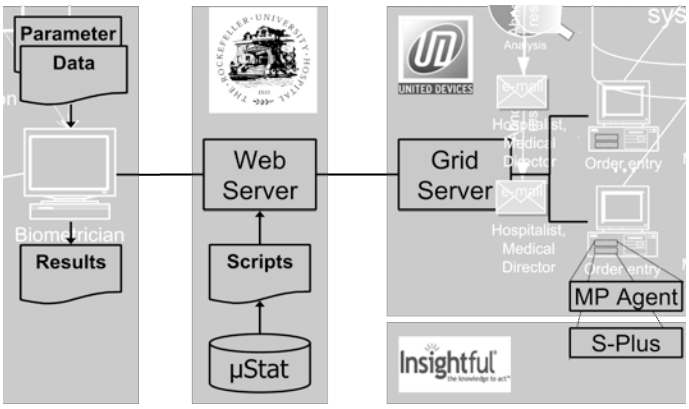


Fig. 2. The muStat grid environment. A user uploads parameters and data, which are executed by an S-PLUS script utilizing the µStat library on a grid of PCs running S-PLUS under a node agent.

The grid is controlled by a Linux server running an IBM DB2 database, a secure Apache web server for a management console, and the grid software from United Devices (www.ud.com). The client nodes consist of mixed x86 Microsoft Windows workstations. Users willing to contribute capacity to the grid can download and install the agent from the muStat Web site. The grid server then installs S-PLUS on the node and moves the node into the queue of active nodes, which process work units at low priority whenever the workstation is not executing other tasks. When a user uploads a job, the server dispatches work units to the nodes, verifies the integrity of the returned results, and notifies the requester when the job is done. As the agent is centrally customized to include an installation of the S-PLUS application, it suffices for the work units to include the data to be analyzed and the S-PLUS script to do the analysis.

2.3 Software

For the grid architecture to be effective, a method had to be found that was not np-hard. Moving from the marginal likelihood principle to u statistics, while foregoing some second order information, formed the basis for developing algorithms that were computationally efficient.

The u-test, except for a missing term in the variance formula, was originally published by THOMAS DEUCHLER[9] in 1914, 33 years before MANN AND WHITNEY. Unfortunately, DEUCHLER presented his ideas using a somewhat confusing terminology, which made the results difficult to understand. On the other hand, being a psychologist, he laid out a scheme for computations that, had it been more widely known, might have given u-statistics an equal footing with methods based on the linear model. Based on his work, we developed algorithms[8] that, besides growing with the square of the number of subjects only, are easily implemented Fig. 3.

	2 1 2	0 0 1	1 1 1	1 1	
	1 2 0	0 0 0	1 1 1	0 0	
	2 0 0	0 0 0	1 0 0	0 0	
2 1 2	0 0	1 1	1 1	1 1	5
1 2 0	0 0	0 0	1 1	1 1	3
2 0 0	-1 0	0 0	1 1	1 1	1
0 0 1	-1 0	0 0	0 0	1 0	0
0 1 0	-1 -1	0 0	0 0	1 -1	-1
1 0 0	-1 -1	-1 0	0 0	1 -2	-2
0 0 0	-1 -1	-1 -1	-1 -1	-1 0	-6

Fig. 3. Computation of u-scores from a data set with the lattice structure of Fig. 1. Each profile is written on both axes. Cells are filled as follows: If the row profile is higher then the column profile, enter "1", if it is smaller "-1", otherwise "0". Once all cells are filled, the u-scores are obtained as the row sums of the array.

With this approach, individual analyses can often be performed even using spreadsheet software. For small samples, spreadsheets for different partial orderings can be downloaded from our Web site (muStat.rockefeller.edu). While clearly not the preferred implementation for routine applications, the availability of spreadsheets demonstrates the ease and computational simplicity of the method.

The power of S-PLUS (www.insightful.com) as a bioinformatics tool lies in the ease with which statistical concepts can be translated into code. Since the matrix of pairwise orderings (Fig. 3) is always symmetric, it is computationally more efficient to work from the non-symmetric matrix of GE (greater or equal) indicators[10;11]. The computation of µ-scores can then be simplified by computing a GE matrix individually for each variable (or pair of variables) and combining these uni-/or bi-variate GE matrices to a profile GE matrix, from which the scores are computed. The muStat package (available from csan.insightful.com/PackageDetails.aspx?Package=mustat), separates the process of computing µ-scores into three functions (error checking and extensions, such as ways to handle the topological structure of 'neighboring' SNPs are omitted here):

```
NAtoZer <- function(x) { x[is.na(x)] <- 0; x }
sq.array <- function(x)
  array(x, dim=c(rep(sqrt(dim(x)[1]),2), dim(x)[2]))
sq.matrix <- function(x) matrix(x, sqrt(length(x)))

mu.GE <- function(x, y=x) {
  x <- as.matrix(x); y <- as.matrix(y)
  if (length(y)>1)
```

```
  apply(rbind(x,y),2,mu.GE,numRows(x)) else
  as.numeric(NAtoZer(outer(x[1:y],x[-(1:y)],">=")))}

mu.AND <- function(GE) { GE <- sq.array(GE)
 AND <- apply(GE, 1:2, any)
 nNA <- AND[,1]*0
 for (i in 1:dim(GE)[3]) {
  nNA <- nNA + diag(GEi <- GE[,,i])
  AND <- AND * (GEi + (1-GEi)*(1-t(GEi))) }
 return(as.numeric(AND * ((c(nNA)%o%c(nNA))>0))) }

mu.Sums <- function(GE) {
 ICW <- function(GE) {
  wgt <- colSums(GE|t(GE))
  sqrt(wgt*(wgt>1)/numRows(GE))}
 GE  <- sq.matrix(GE)
 wght <- ICW(GE)
 list (
  score = ifelse(wght,rowSums(GE)-colSums(GE),NA),
  weight = wght) }
```

By separating the computation of μ-scores into three functions, the analysis process can be separated into an initial step, where data is prepared on the grid server (mu.GE) and then distributed to the nodes for the more computationally intensive steps (mu.AND, mu.Sums, ...).

A downside of having a conceptually simple language for implementing methods can be the lack of computational efficiency. Thus, the most computationally demanding tasks to be performed were implemented in C. In essence, the function mu.GE converts each variable into a bit string, where each bit represents a pairwise ordering. For 2800 subjects, for instance, each variable's GE matrix is stored (and processed) as a 1 MB bit string (before compression). The advantage of using an efficient representation of information sufficient for non-parametric tests is particularly striking with quantitative data, such as gene expression profiles, where the size of the GE matrix does not depend on the accuracy of the measurements.

One of the advantages of screening jobs is that, once the GE matrices have been computed, each statistical test is performed independently, so that parallelization can be done at a high level using standard S-PLUS language without requiring overhead or complicating the numerical calculations. In short, the number of parallel work units (NPWU) is determined up front and each work unit (WU) performs only one of every NPWU calculations

```
for (row in (1:nrows)) {
 if (((row <- row+1) %% NPWU) != WU) next
 # compute results from GE matrices
 cat(c(row,":", results, append = T)
}
```

The grid server then merges the results returned by the work units using the row number preceding them as a key. Fig. 4 shows a typical distribution of work units across a heterogenous grid.[12]

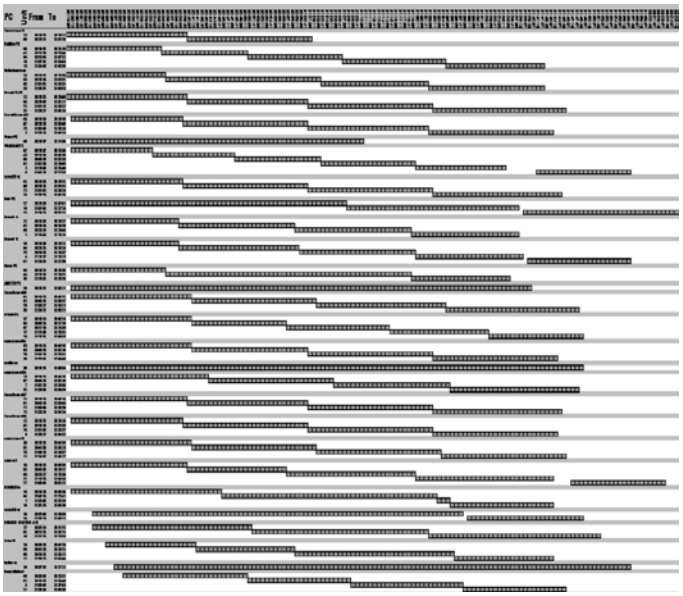


Fig. 4. Distribution of work units across a grid. Rows are work units by node (28 nodes) columns are time points (minutes). Depending on processor speed, concurrent applications, etc., the first node processed two work units, the second node five, and so on.

3 Results

The grid has already been widely used for the analysis of gene expression data[8;13]. Here, we demonstrate its use for genetic information[14]. In this (small) example, we analyzed 179 SNPs in 90 male mice to identify loci coding genes involved in plant sterol metabolism. Through the analysis of individual SNPs, a locus on chromosome 10 is easily identified, with the first four loci having the highest phenotype/genotype correla-

tion. There was also some evidence for association with the first six loci on chromosome 19 (Fig. 5).

As the number of available SNPs increases, several adjacent SNPs may be in linkage disequilibrium with a diseases locus. With µScores, information from adjacent SNPs can be analyzed in a comprehensive fashion, thereby taking advantage of linkage disequilibrium on either side of the disease locus. In short, one computes GE matrices of intervals first and then combines the intervals to diplotypes.

As the "polarity" of the alleles on either side is unknown *a-priori,* the number of diplotypes to be analyzed (times polarity) is 179 (×1), 159 (×2), 139 (×4), 119 (×8) ... . In this example, the diplotypes with the highest phenotype correlation among the 774 diplotypes analyzed, were consistently those including the first SNP on chromosome 10, suggesting that a disease locus at the very end of this chromosome. The results for chromosome 19 did not show a clear preference for a particular locus within the first six SNPs.

With common diseases, a phenotype may be associated with an 'epistatic set' of diplotypes several markers apart, or even on different chromosomes[15], i.e., some loci may contribute mainly by increasing the risk conferred by other loci. With muStat, epistasis is easily investigated, albeit at the price of a substantial increase in computational effort: the above 774 diplotypes can be arranged as 244,378 non-overlapping epistatic pairs. Fig. 5 shows a visualization of this plethora of data.

4 Discussion

As the amount of data and the complexity of the problems to be addressed increases, so does the demand for both bioinformatics and, in particular, biostatistics. In fact, the increase in demand for computational capacity for more complex biostatistical models to be applied typically increases at a higher rate than the demand for data storage and transfer.

Grid computing provides substantial computational resources utilizing unused capacity available in virtually every commercial or academic institution. Here, we demonstrate how harnessing a combination of low-tech solutions can enable grid technology for researchers in genetics and genomics.

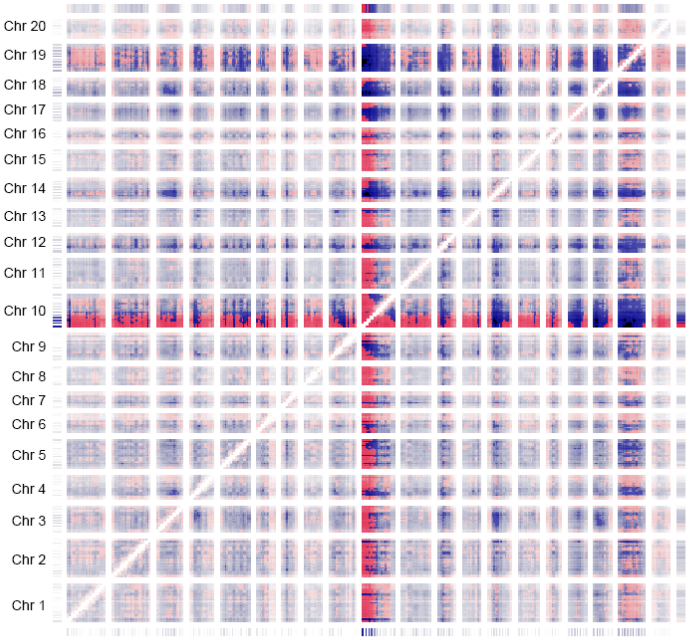


Fig. 5. Epistasis 'heatmap' for genetic information from mice (using S-PLUS 'levelplot'), confirming known risk loci on chr 10 and 19, but also providing evidence for a locus on chr 14, which acts primarily through interaction with other risk loci (blue indicates higher correlation with the phenotype for the bivariate than for either of the univariate µ-scores). **Legend:** Bottom and left margin represent individual SNPs; top and right margin represent diplotypes (no epistasis). The shade in the 375×375 cells shows the highest correlation among all pairs of diplotypes centered at this locus (SNP or interval mid-point). The SNPs on the x-axis are labeled A–E, the SNPs on the y-axis 1:5 and 1:6. With diplotypes of length 1–6, the number of pairs per cell is 25.

To make any computational architecture worth the effort, we moved from the np-hard the marginal likelihood principle to the closely related u-statistics, where the computational effort raises with the n^2, rather than $n!$ so that spreadsheets are now available from muStat.rockefeller.edu for simple problems and as tutorial tools. We then optimized and modularized the algorithm to improve flexibility of the models, e.g., for epistatic sets of diplotypes vs. gene expression pathways. The S-PLUS library 'muStat' is available from CSAN.insightful.com.

We used C for computationally intensive subroutines improved computational efficiency and compression of data for exchange over a grid. Using C code on the grid increased speed approximately 10-fold.

Devising a high-level mechanism for splitting jobs into work units helped to avoid the need for inter-process communication and, thus guarantees that the speed increases linearly with the number of nodes on the grid. With 10,000 MFlops assigned to the job, it took 10 hours to compute the results of Fig. 5. Users can upload data to the grid via muStat.rockefeller.edu. A major advantage of the grid architecture is that every upgrade of the participating workstations also upgrades the grid as a whole.

Finally, extensive number crunching is of limited use if the results are not presented in a form easy to comprehend. For decades, statistical "presentation" graphics (e.g., bar charts) had only minimal information content. Microarray 'heatmaps' were an early step towards presenting massive univariate data. As the grid technology allows us to move towards multivariate data, new challenges arise to let users extract information from the results of massively parallel computations. In the last part of our paper, we suggested a graphical representation for the information contained in millions of analyses in a fashion that – as we believe – can be easily interpreted and communicated. Undoubtedly, the capability of software to produce highly informative graphics will be a key requirement for screening analyses of multidimensional data (epistatic sets, and genomic pathways) to become more widely accepted.

References

1. Wittkowski, KM (1988) Friedman-type statistics and consistent multiple comparisons for unbalanced designs. *Journal of the American Statistical Association* **83**: 1163-70
2. Wittkowski, KM (1992) An extension to Wittkowski. *Journal of the American Statistical Association* **87**: 258
3. Susser, E; Desvarieux, M*; et al.* (1998) Reporting sexual risk behavior for HIV: a practical risk index and a method for improving risk indices. *American Journal of Public Health* **88**: 671-4
4. Banchereau, J; Palucka, AK*; et al.* (2001) Immune and clinical responses after vaccination of patients with metastatic melanoma with CD34+ hematopoietic progenitor-derived dendritic cells. *Cancer Research* **61**: 6451-8
5. Hoeffding, W (1948) A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**: 293-325
6. Mann, HB; Whitney, DR (1947) On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**: 50-60
7. Gehan, EA (1965) A generalised Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* **52**: 203-23
8. Wittkowski, KM; Lee, E*; et al.* (2004) Combining several ordinal measures in clinical studies. *Statistics in Medicine* **23**: 1579-92
9. Deuchler, G (1914) Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. *Z pädagog Psychol* **15**: 114-31, 45-59, 229-42
10. Cherchye, L; Vermeulen, F (2007) Acknowledgement of Priority. *Journal of Sports Economics* **8**: 557-
11. Cherchye, L; Vermeulen, F (2006) Robust Rankings of Multidimensional Performances: An Application to Tour de France Racing Cyclists. *Journal of Sports Economics* **7**: 359-73
12. Wittkowski, K; Haider, A*; et al.* (2006) Bioinformatics Tools Enabling U-Statistics for Microarrays. *Conf Proc IEEE Eng Med Biol Soc* **1**: 3464-9
13. Spangler, R; Wittkowski, KM*; et al.* (2004) Opiate-like Effects of Sugar on Gene Expression in Reward Areas of the Rat Brain. *Molecular Brain Research* **124**: 134-42
14. Sehayek, E; Yu, HJ*; et al.* (2004) Genetics of cholesterol absorption and plasma plant sterol levels on the Pacific island of Kosrae. *Perfusion* **17**: 2
15. Gambis, A; Sehayek, E*; et al.* (2003) It's not Broadway - Visualizing Epistatic Interaction. *TIGR's 15th Annual Genome Sequencing and Analysis Conference (GSAC XV),* 2003-08-03..07. Savannah, GA

Graphics at JSM 2007

The Statistical Graphics Section had an exciting program at JSM 2007 with two invited sessions, three topic-contributed sessions, a regular contributed session, several roundtable luncheons and a continuing education class. Stat Graphics also co-sponsored 19 other sessions. The invited sessions were:

- Exploring Models Interactively, organized by Antony Unwin
- Scagnostics, organized by Leland Wilkinson

The topic contributed sessions were:

- Applications of Visualization for Web 2.0, organized by Dan Rope
- Statistical Graphics for Everyday Use, organized by Martin Theus
- Statistical Graphics for Analysis of Drug Safety and Efficacy, organized by Michael O'Connell

These 5 sessions were consistently high quality, pushing on new frontiers of information extraction, visualization and communication. In the Scagnostics session (short for scatter plot diagnostics) we saw a demo where any text or data file could be dragged into an application and variables with high scagnostics scores (e.g. linearity, lumpiness, ...) were rapidly extracted and graphically summarized.

In the Web 2.0 session we saw statistical graphics on the iPhone and a lots of fascinating graphics mashups. One mashup that was shown live included Bob Fosse choreography and an Andre 3000 rap soundtrack 'Walk it Out' with graffiti annotation!

In the Drug Safety and Efficacy session we saw a wide variety of informative graphics for clinical reports and interactive clinical review. These graphics were designed for rapid detection of potential safety and efficacy responses; and clear communication of results among clinical study teams.

The continuing education course was Graphics of Large Datasets given by Antony Unwin and Heike Hoffman. This was based on the terrific new book by Antony, Martin Theus and Heike, Graphics of Large Datasets. Springer (2006), ISBN: 978-0-387-32906-2. Several folks were spotted buying and reading this book during the meeting.

Kudos to Program Chair Simon Urbanek and all of the session organizers, chairs, and speakers.

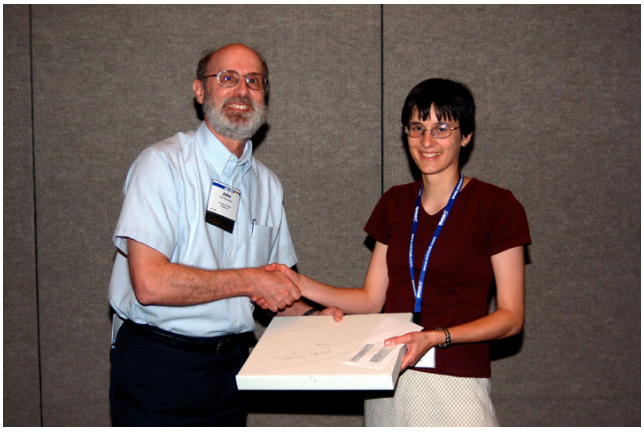
Some of the authors from the Drug Safety and Efficacy session and the Web 2.0 session have made their presentations available on the new the new Statistical Graphics and Computing web site http://www.stat-computing.org/events/2007-jsm

We've included a few photos from the JSM mixer below. There are more on the website along with a link to a Flickr account with even more photos.

We hope that the presentations and photo areas of the website will grow, both with contributions from 2007 and going forward in to 2008. We encourage folks to send their talks, photos, related materials or URL links to Hadley Wickham <h.wickham@gmail.com> for inclusion in the new web site.

- Michael O'Connell

Folks at the Statistical Computing and Graphics Mixer at JSM 2007



Heather Turner receives the Chambers Award from John Monahan at the JSM 2007 Mixer

Computing at JSM 2007

The Statistical Computing Section had a strong program at JSM 2007, with eight invited sessions, two topic-contributed sessions, twelve contributed paper sessions, and posters in three sessions.

A focal point this year was statistical learning and data mining, with four invited sessions organized by Ji Zhu, Yasmin H. Said and Ed Wegman, Hui Zou, and Yufeng Liu.

In contrast to recent JSMs, where the initials stood for "Jumble of Sessions on Microarrays", there were no sessions specifically on microarrays, though there were two on a variety of topics in computational biology, organized by Yasmin Said and Elizabeth Slate. Said and Wegman organized "Computationally Intensive Methods" (exploratory software, cross-validation, and clustering) and a session on networks.

Perhaps the weakest aspect of the program was the lack of topic-contributed sessions; there were only two, the annual Statistical Computing and Statistical Graphics Paper Competition organized by John Lockwood, one on Graphs and Networks organized by Sujay Datta.

Topic-contributed sessions provide a way for volunteers to create a coherent session on a topic they choose. They provide a more forum than the jammed-together short talks in a regular contributed session. I encourage people to consider organizing one of these sessions, either a traditional talks session or a panel.

Finally, poster sessions played a more prominent part in the program this year, a trend that will continue. I'll personally attest that posters provide a better opportunity for interaction and extended discussion than do talks, and I'm increasingly doing posters instead of talks.

Kudos to Program Chair Ed Wegman and all of the session organizers, chairs, and speakers.

The 2007 program can be searched at:

http://www.amstat.org/meetings/jsm/2007/onlineprogram/index.cfm?fuseaction=main

Speakers are invited to send URLs linking to their talks or related materials to Hadley Wickham <h.wickham@gmail.com> for inclusion in the new Statistical Graphics and Computing web site http://www.statcomputing.org.

- Tim Hesterberg

Computing at JSM 2008

OVERVIEW OF STAT COMP JSM 2008

Wolfgang Jank
Program Chair

The section on statistical computing has lined up an exciting program for JSM 2008. Our topics fill the entire spectrum between curriculum design, methodological development, and analysis of real-world problems. The individual sessions vary in format and include invited speakers and discussants from around the world, from academia, from research labs and from industry. The common theme among all of the sessions is the computational aspect. This could be in the form of teaching statistical computing to our students, developing new computational algorithms, or the use of computational methods to analyze challenging and large-scale data. In the following, I will provide a brief overview of the invited session program.

Our first invited session is on *Designing Courses on Statistical Computing*. By now, most statistics departments have recognized the need to make computation a central focus of their graduate and undergraduate education. This has been done in different ways, and the most common is a specifically designed course that is taught in the first or second semester of a student's career. The contents of such a course are not well-established, in part because there is no clearly defined common subject matter, and in part because there is no widely-used text that prescribes the content. This session presents several approaches to the problem of designing a broad and sufficient course on modern statistical computing to Ph.D. students in statistics.

Two of our invited sessions deal with the development of new statistical methodology. The first session is on *Global Maximization in EM-type Algorithms*. In likelihood-based modeling, finding the parameter estimates that correspond to the global maximum of the likelihood function is a challenging problem. This problem is especially serious in statistical models with missing data, where popular EM-type algorithms are used to obtain parameter estimates. Despite their hallmark stability and monotone convergence, EM-type algorithms (i.e. the EM algorithm and its variants, such as MM algorithms) frequently converge to local, sub-optimal solutions, especially when the dimension of the parameters space and/or the size of the data are large. This is a particular concern in e.g. finite mixtures and latent variable models. The problem has largely been ignored in statistics and only unprincipled approaches have been used to date (e.g. random multi-starts). Part of the problem is that no practically useful characterization of the global maximum is available. This session presents different solutions to global optimization problems in the context of the EM algorithm.

The second methodological session considers *Advances in Functional Data Analysis (FDA)*. The technological advancements in measurement, collection, and storage of data have led to more and more complex data-structures. Examples include measurements of individuals' behavior over time, digitized 2- or 3-dimensional images of the brain, and recordings of 3- or even 4-dimensional movements of objects traveling through space and time. Such data, although recorded in a discrete fashion, are usually thought of as continuous objects represented by functional relationships. This gives rise to functional data analysis (FDA), where the center of interest is a set of curves, shapes, objects, or, more generally, a set of *functional observations*. This is in contrast to classical statistics where the interest centers around a set of data vectors. FDA has experienced a rapid growth over the past few years, both in the range of applications for it techniques and in the development of theory for statistical inference. This session will bring together leading researchers in the field to discuss recent advances and applications of FDA. Topics addressed will range from theoretic properties and inference in functional linear regression to new regularization techniques and the application of functional models to the analysis of complex data structures.

Three of our invited sessions address computations challenges driven by large-scale and complex real-world applications. These applications include online advertising, social networks, and Wikipedia data. The Wikipedia is an important unintended consequence of the Internet. It demonstrates how valuable content

can be created by distributed volunteer effort. And it also provides statisticians with a rich data source on the connectivity between knowledge domains, the investment/growth trade-offs needed to bootstrap a distributed business, and the difficult problem of quality assurance in a climate of complex error sources. Also, from the standpoint of our professional society, Wikipedia offers a new model for service the ASA can provide to its members, by integrating a range of user-created code, data, discussion, and articles. The session *Analysis of Wikipedia Data* includes analyses of key Wikipedia data sets and a discussion of a possible Wikipedian future for ASA publications.

Another internet phenomenon that poses new and exciting challenges for the statistician are online social networks. Online social networks have become ubiquitous. Sites such as *mySpace* and *Facebook* create online communities where users define their own social networks; other sites such as blogs and online forums create de-facto networks as users link to each other's opinions. There has been a flurry of research on these networks on topics such as dynamic network evolution, community detection, discovering influential actors, etc. However, most of this work has been done *outside* of statistics, in the fields of machine learning, AI, and CS, and published at data mining conferences like KDD. In the session *Analysis of Massive Online Social Networks*, we will discuss a variety of topics about online social networks, blogs, and product review networks. We will also discuss why statisticians have been absent from this field to date, and what contributions our community can make.

Our last invited session deals with the ever-growing phenomenon of online advertising and its implications for the statistician. Online advertising is a multi-billion dollar industry as evident from the phenomenal success of companies like Google, Yahoo, or Microsoft, and it continues to grow at a rapid rate. With broadband access becoming ubiquitous, internet traffic continues to grow both in terms of volume and diversity, providing a rich supply of inventory to be monetized. Fortunately, the surge in supply has also been accompanied by increase in demand with more dollars being diverted to internet advertising relative to traditional advertising media like television, radio, or newspaper. Marketplace designs that maximize revenue by exploiting billions of advertising opportunities through efficient allocation of available inventory are the key to success. Due to the massive scale of the problem, the only feasible way to accomplish this is by learning the statistical behavior of the environment through massive amounts of data constantly flowing through the system. This gives rise to a series of challenging statistical problems which include prediction of rare events from extremely massive and high dimensional data, experimental designs to learn emerging trends, protecting advertisers by constantly monitoring traffic quality, ranking search engine results, modeling extremely large and sparse social network behavior. The session Statistical Challenges in Online Advertising and Search will discuss many of these challenges.

I hope that, by now, your are as excited about next year's program as I am, and I hope that I will see many of you at the sessions sponsored by *Stat Comp*. I would also like to remind you that it is not too late for you to contribute. There are still open slots for topic contributed sessions. If you have any questions, please don't hesitate to contact me at wjank@rhsmith.umd.edu.

Best wishes,
Wolfgang

Graphics at JSM 2008

OVERVIEW OF STATISTICAL GRAPHICS JSM 2008

... and call for topic-contributed sessions

David Hunter
Program Chair

The Section on Statistical Graphics is organizing four invited sessions at JSM 2008 in Denver. For those who aren't aware of how this number is determined, the various sections are allotted a fixed number of invited sessions (e.g., stat graphics currently gets two and

stat computing gets four), then each section was allowed this year to submit up to two additional entries into a competition, which is judged by all of the section program chairs and the overall JSM program chair, for the remaining spots. How is the number of allotted slots determined? Partly at least it is based on the number of topic-contributed sessions sponsored by a section – so if you have ideas for sessions that you'd like to organize, please consider doing so. The period from October-December is the best time to organize such sessions, so if you've been thinking about it, don't wait! If you have questions, email me directly at dhunter@stat.psu.edu.

The theme of JSM 2008 is "Communicating Statistics: Speaking out and reaching out", and the stat graphics sessions all address this theme in some way. After all, communication is arguably what statistical graphics is all about! These four sessions, in no particular order, are:

The emergence of social data analysis and its impact on the field of Statistics. The term "web 2.0" is commonly used to describe web sites/ applications that allow users to create their own content and network with other users who have similar interests. Commonly cited examples are YouTube (for videos), Flickr (for photos) and MySpace (for personal web sites). Recently, a number of web sites have used web 2.0 ideas to allow users to upload/analyze data, publish their results and discuss their findings with other users. The term coined for this activity is social data analysis. In this session, representatives from three of these sites (Many Eyes, Swivel and StatCrunch) will discuss their intentions and experiences with regard to social data analysis. They will also discuss the specific impact of social data analysis on the practice of statistics and broader impacts on the whole of society.

Political Science, Statistical Science, and Graphics. 2008 is of course an election year in the U.S., and this session addresses the theme of politics by bringing together several speakers from the quantitative political science community who will present their innovative graphical stastistical work. From a broadly applicable new R package for graphics to a discussion of new developments in cartograms to a case study on farm subsidies using novel statistical tools to uncover a political message, this session should lead to lively discussions not only about statistics but about geography and politics as well!

Visualizing Large Datasets. As interest in effective data-mining techniques explodes, demand for novel statistical graphical techniques for exploring and understanding large datasets has never been greater. This session describes several new approaches to visualizing the information in these datasets. The speakers, representing both industry and academia, will discuss cutting-edge techniques for dealing with streaming data, applying simple models to large dataset visualization, and diagnosing scatterplots (via "scagnostics") in high-dimensional space.

Statistics-Geography Mashups on the Web. This session comprises four presentations that show how geography and statistics have been closely linked on freely accessible web pages. The speakers will provide a general overview on geo-mashups using Google Earth for data exploration; demonstrate research focused on web-based geovisualization and geocollaboration; show how previously developed code for micromaps has been linked with a climatology data base on the web; and report on statistical lessons learned when considering cultural effects for the construction of online maps. This session should be of interest to everyone dealing with the web - as a user or as a developer. In addition to the disciplines directly related to these presentations, i.e., geography, medicine, and climatology, a variety of additional applications might benefit from the information presented here: governmental web sites and web sites providing access to medical or environmental data, to mention only a few.

My thanks to Web West, Anton Westveld, Heike Hofmann, and Juergen Symanzik for organizing this great lineup and for providing much of the text of the write-ups above. I hope you all enjoy your week at the foot of the Rocky Mountains!

- David Hunter

News & Awards

2008 CHAMBERS AWARD

The Statistical Computing Section of the American Statistical Association announces the competition for the John M. Chambers Statistical Software Award. In 1998 the Association for Computing Machinery presented its Software System Award to John Chambers for the design and development of S. Dr. Chambers generously donated his award to the Statistical Computing Section to endow an annual prize for statistical software written by an undergraduate or graduate student.

The prize carries with it a cash award of $1000, plus a substantial allowance for travel to the annual Joint Statistical Meetings where the award will be presented.

Teams of up to 3 people can participate in the competition, with the cash award being split among team members. The travel allowance will be given to just one individual in the team, who will be presented the award at JSM. To be eligible, the team must have designed and implemented a piece of statistical software. The individual within the team indicated to receive the travel allowance must have begun the development while a student, and must either currently be a student, or have completed all requirements for her/his last degree after January 1, 2006. To apply for the award, teams must provide the following materials:

- Current CV's of all team members.
- A letter from a faculty mentor at the academic institution of the individual indicated to receive the travel award. The letter should confirm that the individual had substantial participation in the development of the software, certify her/his student status when the software began to be developed (and either the current student status or the date of degree completion), and briefly discuss the importance of the software to statistical practice.
- A brief, one to two page description of the software, summarizing what it does, how it does it, and why it is an important contribution. If the team member competing for the travel allowance has continued developing the software after finishing her/his studies, the description should indicate what was developed when the individual was a student and what has been added since.

Access to the software by the award committee for their use on inputs of their choosing. Access to the software can consist of an executable file, Web-based access, macro code, or other appropriate form. Access should be accompanied by enough information to allow the judges to effectively use and evaluate the software (including its design considerations.) This information can be provided in a variety of ways, including but not limited to a user manual (paper or electronic), a paper, a URL, online help to the system, and source code. In particular, the entrant must be prepared to provide complete source code for inspection by the committee if requested.

All materials must be in English. We prefer that electronic text be submitted in Postscript or PDF. The entries will be judged on a variety of dimensions, including the importance and relevance for statistical practice of the tasks performed by the software, ease of use, clarity of description, elegance and availability for use by the statistical community. Preference will be given to those entries that are grounded in software design rather than calculation. The decision of the award committee is final.

All application materials must be received by 5:00pm EST, Monday, February 25, 2008 at the address below. The winner will be announced in May and the award will be given at the 2008 Joint Statistical Meetings.

Information on the competition can also be accessed on the website of the Statistical Computing Section (www.statcomputing.org or see the ASA website, www.amstat.org for a pointer), including the names and contributions of previous winners. Inquiries and application materials should be emailed or mailed to:

Chambers Software Award
c/o J.R. Lockwood
The RAND Corporation
4570 Fifth Avenue, Suite 600
Pittsburgh, PA 15213
lockwood@rand.org

- JR Lockwood

2007 AWARDS

The winner of the 2007 John M Chambers Award was:

Heather Turner and David Firth (University of Warwick Department of Statistics) for "gnm", an R package for fitting generalized nonlinear models

Heather was on hand at the JSM Mixer in Salt Lake City for presentation of the award. Heather and David wrote up a great article on this package in our last newsletter.

The winners of the 2007 Student Paper competition were:

•*Andrew Finley (advisors Sudipto Banerjee and Alan R. Ek), "spBayes: An R Package for Univariate and Multivariate Hierarchical Point-referenced Spatial Models"*

•*Alexander Pearson (advisor Derick R. Peterson), "A Flexible Model Selection Algorithm for the Cox Model with High-Dimensional Data"*

•*Sijian Wang (advisor Ji Zhu), "Improved Centroids Estimation for the Nearest Shrunken Centroid Classifier"*

•*Hadley Wickham (advisors Di Cook and Heike Hofmann), "Exploratory Model Analysis"*

NEW SECTION ON STATISTICAL LEARNING AND DATA MINING

At this point all the paperwork for creating our new ASA Section on Statistical Learning and Data Mining has been handed in to the ASA office. It is scheduled to be reviewed in February and voted on at the JSM in August. If all goes well, we become official as of January 1, 2009.

- Joe Verducci

NSF FUNDS FOR GRADUATE STUDENT ATTENDANCE AT ISAAC NEWTON INSTITUTE WORKSHOPS

The Isaac Newton Institute for Mathematical Sciences at Cambridge, U.K., is hosting a six-month research program on "Statistical Methods for Complex, High-Dimensional Data" from January through June, 2008. The program is attracting some of the leading researchers in statistics, computer science, and computational biology. See www.newton.cam.ac.uk/programmes/SCH/ for more details on the program and the participants.

As part of that program there will be a workshop on "High Dimensional Statistics in Biology" from March 31 through April 4, and a closing workhop on "Future Directions in High-Dimensional Data Analysis" from June 23-27. The National Science Foundation has provided funds to enable ten senior graduate students at U.S. institutions to attend one of those workshops.

Awards will be determined by a committee consisting of Sara van de Geer (chair), David Banks, Peter Bickel, and Ann Lee. The deadline for receiving applications is January 7, 2008. Interested students should send an application packet that contains:

1. A letter of recommendation from their advisor or department chair (that letter should indicate the applicant's gender and minority status).
2. The curriculum vitae of the applicant.
3. If the applicant would like to submit a poster for possible presentation at the workshop, please include a title and abstract.
4. An indication of which workshop the applicant would like to attend (this will not be used for selection).

All applications should be sent electronically to both Sara van de Geer (geer@stat.math.ethz.ch) and David Banks (banks@stat.duke.edu).

- David Banks

R and S-PLUS User Meeting Roundup

useR! 2007

Iowa State University hosted useR! 2007 was held in Ames Iowa, Aug 8-10 2007.

About 250 participants came from many different subject areas -- physics, business, ecology, agronomy, bioinformatics, statistics, mathematics -- attesting to the impact of R. John Chambers started the conference off with the keynote talk on the mission of R, in which he noted the similarities between programming with cooking.

There were many interesting talks and posters, including teaching with R, a comparison between the graphical models of lattice and ggplot and graphical user interfaces for R. A panel session on the use of R in clinical trials and industry-sponsored medical research discussed the progress towards FDA acceptance of R for reporting on clinical trials.

The event roughly coincided with the 10th anniversary of the forming of R core. To celebrate a huge blue cake in the shape of R was shared by all, and Thomas Lumley gave a short presentation at the conference dinner about the history of R Core.

There were two programming competitions. The first, announced in advance, to produce a package useful for large data sets. The team of Daniel Adler, Oleg Nenadic, Wlater Zucchini and Chistian Glaser from Gottingen won this with their *ff* package. The second competition was a series of short R programming tasks held over a two hour period during the meeting. This was won by Elaine McVey and Olivia Lau in a tie for first.

Copies of presentation slides, posters and details on the programming competitions are available at http://www.user2007.org.

The next useR! 2008 will be hosted by Dortmund University, August 12-14 2008. Information is available at http://www.R-project.org/useR-2008.

R-core folks look up from a 10th anniversary cake!

- Di Cook

INSIGHTFUL IMPACT 2007

A large crowd of statisticians from many commercial, government and academic settings celebrated 20 years of S-PLUS and 10 S-PLUS User Conferences with the Insightful Impact 2007 Meeting in Atlantic City, October 5-9, 2007.

Three parallel presentation tracks in financial services, life sciences and general customer analytics kept attendees engaged. Dr. Thomas Davenport, author of *Competing on Analytics, The New Science of Winning*, delivered the keynote address. The Life Sciences track included the following presentations:

- Harry Southworth (Astra Zeneca): *S-PLUS Data Review Tools for Clinical Study Data*
- José Pinheiro (Novartis): *MCPMod: A Library for the Design and Analysis of Dose Finding Trials*
- Simon Zhou (Wyeth): *Exposure-Response Based Trial Simulations to Assess Adaptive Designs in Phase II Dose Ranging Trials*
- Terry Therneau (Mayo Clinic): *Analysis of Proteomics Data*

- Drew Levy (Novartis): *S-PLUS Graphics and the Analysis of Outcomes Data*
- Michael Durante (GSK): *S-PLUS Clinical Graphics for Exploratory Analysis, Reporting, Submission and Presentation*
- Coen Bernaards (Genentech): *Using S+SeqTrial™ to Design Time-To-Event Trials When Interim Analyses are at Fixed Calendar Times*
- Michael O'Connell (Insightful): *Statistical and Graphical Analysis of Drug Safety Data in Clinical Trials*

There were also several training sessions on S-PLUS products and solutions including Terry Therneau (Mayo Clinic) on *Modeling Survival Data with S-PLUS*, and Eric Zivot (U Wa) on *Advanced Time Series Methods in S-PLUS® and S+FinMetrics® for Modeling Financial Data*. A panel session on *The Future of Statistical Computing*, discussed the future of statistical and data analysis computing as it relates to the finance, life science and customer analytic industries.

At the Insightful Impact Awards Dinner, four customers were recognized for Insightful Solutions that have made a significant impact on their organization's growth and/or profitability. These were:

- ImClone Systems, a biopharmaceutical company, has deployed S-PLUS® Enterprise Server to create a system that tracks its production metrics, enabling the company to produce products more efficiently.
- Macronix, a Taiwanese integrated device manufacturer, is providing engineering and business data analysis on demand across its production pipeline by implementing Insightful's S-PLUS Enterprise Server along with customer-developed Java applications.
- Time/Warner Retail Sales & Marketing, the largest publisher-owned magazine retail marketer in the US, has improved production and distribution and created a new business by deploying Insightful's S-PLUS Enterprise Server as its predictive analytics and reporting solution.
- Zurich Financial Services, an insurance-based financial services provider, utilized S-PLUS Enterprise Server to enable systems from around the world to share data, models and results within its Risk Modeling Platform.

S-PLUS Users celebrate novel uses of S-PLUS inside the Russian ice room at the Tropicana Hotel, after the Insightful Impact Awards

Everyone at the conference had a fun time. The Tropicana is a great hotel with many interesting restaurants and amenities. It is easy to see why the Deming Conference is held at this hotel each year.

- Kim Kelly

BASS XIV

More than 120 attended the fourteenth annual Biopharmaceutical Applied Statistics Symposium (BASS XIV), held 5-9 November 2007, at the Mulberry Inn Suites Hotel in historic Savannah, Georgia. Sixteen 1 hour tutorials, on diverse topics pertinent to the research, clinical development and regulation of pharmaceuticals, were presented 5-7 November, by speakers from academia, the pharmaceutical industry and the Food and Drug Administration (FDA). Three parallel, two-day short courses: *Statistical Methods for Biomarker Discovery* by Mark van der Laan and Cathy Tuglis; *Hot Topics in Clinical Trials* by Louise Ryan, Scott Evans & Tianxi Cai ; and *Statistical Graphics, Covariance Analysis and Bayesian Methods in RCTs* by Frank Harrell, were presented 7-9 November.

BASS XIV Student Award Winners

Mitch Gail delivers the keynote address at BASS XIV

Popular features of BASS XIV were the keynote address and the FDA/ Industry session. The keynote address *Absolute Risk: Clinical Applications and Controversies* was delivered by Dr. Mitchell Gail of the NCI. BASS is a nonprofit entity founded in 1994 by Dr. Karl E. Peace and currently sponsored by the Jiann-Ping Hsu College of Public Health at Georgia Southern University and the Department of Biostatistics, Medical College of Virginia Campus, Virginia Commonwealth University, for the purpose of generating funding for graduate fellowships in biostatistics.

BASS XV will be held 3-7 November 2008 at the Mulberry Inn Suites Hotel in Savannah, Georgia. For further information, please contact the BASS registrar at 912-486-7904, 912-486-7907 fax, email: bass@georgiasouthern.edu or Dr. Laura Gunn, BASS Co-Chair, 912-486-7422, email address lgunn@georgiasouthern.edu .

The BASS Webpage may be viewed at http://BASS.georgiasouthern.edu .

- Laura Gunn and Karl Peace

Puzzles

The following puzzle comes from the Insightful Impact Meeting, contributed by Stephen Kaluzny. Please submit your answers to Michael O'Connell (moconnell@insightful.com) for publication in the next newsletter. FYI, this puzzle was solved by an S-PLUS user at the Insightful Impact Meeting. Their solution will be published in the next newsletter.

Question

Given a square piece of property of unit side you wish to build fences so that it is impossible to see through the property, i.e. there is no sightline connecting two points outside the property and passing through the property that does not intersect a fence. The fences do not have to be connected and several fences can come together at a point.

The fences can be placed in the interior of the property, they aren't restricted to the boundary. What is the minimum total length of fencing required and how is it arranged. For example you could place fencing along all four sides. This would have total length 4 but is not the best possible.

Hint: You can do better than 2*sqrt(2)

Statistical COMPUTING & GRAPHICS

The Statistical Computing & Statistical Graphics Newsletter is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. All communications regarding the publication should be addressed to:

Michael O'Connell, Editor,
Statistical Computing Section
moconnell@insightful.com

Andreas Krause, Editor
Statistical Graphics Section
akrause@Pharsight.com

All communications regarding ASA membership and the Statistical Computing and Statistical Graphics Section, including change of address, should be sent to:

American Statistical Association,
1429 Duke Street
Alexandria, VA 22314-3402
USA
phone: (703) 684-1221,
fax: (703) 684-2036
asainfo@amstat.org