SVD of Weighted or Missing Data

By Brian Holland

Matrix decomposition techniques such as singular value decomposition (SVD), shed light on data's underlying structure. My article in the last newsletter describes SVD and described an example with unemployment rates. Unemployment rates were laid out in a grid, by county and month. The main features were quickly identified and then clustering techniques were applied among time periods—confirming the narrative that we already know about unemployment rates since 1990.

Insurance data, however, are rarely laid out in such a clean fashion. What happens when there are missing values? What happens when some values, like decrements, are based on little exposure and therefore volatile? These problems are not unique to insurance but actually quite widespread. Techniques are being developed and refined to address similar applications. This note gives a quick overview of approaches and points to further sources.

NONINSURANCE APPLICATIONS

There are a few incomplete data sets that we all encounter. Recommender systems are based on user ratings or purchases. None of us has watched—or at least rated—all the offerings on Netflix, with the possible exception of a few procrastinating actuarial students. The matrix of ratings by user and movie is mostly holes, i.e., it is a sparse matrix where most of the cells in any particular row or column are empty. So, how do you decompose that? For another example, images are matrices of pixels. What if many pixels are missing? How can objects be identified or images cleaned or completed? There is such a huge case for estimating who wants what, it is no wonder there are so many papers on the topic.

ACTUARIAL APPLICATIONS

There are several cases where clustering incomplete data would be highly useful. How many x-factor sets are needed? Or more generally, how many sets of assumptions are needed? The data are incomplete because the population is still running out. Much of actuarial work involves estimating the incomplete parts of a matrix: experience that has not yet evolved by policy duration or certain attained ages, for example. These matrix completion efforts could inform the actuary's decision on estimates for areas of thin exposure.

MISSING DATA: FILLING THE GAP

There are several approaches to estimating the missing data. The paper *Methods for Large Scale SVD with Missing Values* (Kurucz, Benczur, and Chalogany) focuses on estimating missing Netflix ratings. The dataset is much larger than actuaries are likely to encounter in practice. The authors conclude that a modified Lanczos algorithm is somewhat better than the power algorithm. The expectation maximization (EM) algorithm used proceeds with two steps after initializing the blanks: SVD is performed, blanks are re-estimated from the SVD results. The process repeats until adequate convergence.

Three initializations are tested for the missing values: either zeros, averages, or an item-based recommender using an adjusted cosine distance. Initialization with zeros or averages led to slow, if any, convergence. As for the algorithm itself, the authors used SVD as implemented by the Lanczos algorithm in *svdpack*, an R package, with some modifications. They also described and tested a modified power iteration method, but found that it over-fit the data.

Actuaries need to deal with weights in addition to missing data, and also include prior knowledge. In *Fast Regularized Low Rank Approximation of Weighted Data Sets*, Saptarshi Das and Arnold Neumaier present just such an approach. Regularization is applied to impose prior knowledge, in this case smoothness of a series, and also to avoid over-fitting. Imposing smoothness requirements is reminiscent of familiar actuarial techniques like graduation. The usual SVD formula is modified to apply a matrix of weights for the SVD error terms and also regularization terms for the left and right singular vectors. The example on which the method is tested is noise removal from astronomical images.

A more direct link to statistical models is shown in *Forecasting Time Series of Inhomogeneous Poisson Processes with Application to Call Center Workforce Management* by Haipeng Shen and Jianhua Huang. Calls into call centers are

CONTINUED ON PAGE 24

SVD OF WEIGHTED OR MISSING DATA | FROM PAGE 23

treated as a Poisson. They vary by time of day. The calls by time of day (bucketed) and day form a matrix, the expected value of which is the Poisson parameter. This parameter is transformed via a link function and decomposed to find latent parameters by maximizing likelihood. The Poisson factor model is fit to reduce the dimensions of the matrix. The ultimate goal of this paper is to fit a time series onto the factor score series. There are clear similarities to many insurance processes. The link function and probability-oriented approach could fit well within an actuarial context.

Brian Holland

Brian D. Holland, FSA, MAAA, is director and actuary, of Life and Individual A&H Experience Studies at AIG. He also serves as Vice-Chair of the Forecasting and Futurism Section Council. He can be reached at brian.holland@aig.com.

SVD's related technique, principal component analysis (PCA), is reformulated as a maximum likelihood solution of a latent variable model in *Pattern Recognition and Machine Learning* by Christopher Bishop. This reformulation is dubbed "probabilistic PCA." In this Bayesian treatment there are prior and posterior distributions of the latent principal components. Placing PCA in the familiar Bayesian framework brings all the familiar Bayesian advantages. For example, explicit expression of opinion and updating the model for new events.

All of these techniques hold promise for actuaries. Financial models and experience contain a vast array of decrements. Linking these decrements with dimension reduction techniques, and linking those further with Bayesian techniques, can yield a communicable overview of both existing models and also actual experience. ▼