Conference on
missing values and matrix completion
June, 17-19, 2015
AGROCAMPUS OUEST, 65 rue de Saint Brieuc, Rennes, France
**Introduction**

First of all, we would like to welcome you to the 1st MissData conference.

The objective of this conference is to spotlight the latest research in missing values and matrix completion. Our aim is to bring together theoretical and applied researchers from different fields of application such as social-science, genetics, medical analysis, survey, collaborative filtering, image reconstruction, etc.

We do believe that the different communities with their different background, vocabulary will gain a lot by discussing and hope that you can confront your practice, bring complementary skills and different views on the same subject.

We hope that you will have a terrific time at the conference with fruitful and stimulating exchanges and that you will enjoy your stay in Rennes.

Vincent Audigier, Rémi Gribonval and Julie Josse, principal investigators.

**Organization**

MissData is organized by the statistic department of Agrocampus and is supported by the group Data Mining and Learning group of the French Statistical Society.

The *organising committee* for missData 2015 consists of the following members:

- The applied mathematics department of Agrocampus Ouest head by David Causeur, with the secretaries Karine Bagory and Elisabeth Lenaud
- Agrocampus Ouest with Stephane Crespel and Anne Bourdeau (Logistic), Valérie Le Sabazec and Vinciane Marchais (Communication)
- The French Statistical Society SfDS with Nathalie Villa-Vialaneix, Dominique Moccatti and the Data Mining group

The *scientific committee* consists of:

- Vincent Audigier - Agrocampus Ouest, France
- Guillaume Chauvet - Ensai, France
- Remi Gribonval - INRIA, France
- Julie Josse (President) - Agrocampus Ouest, France
- Ken Kleinman - Harvard Medical School, US
- Rahul Mazunder - Columbia University, US
- Cattram Nguyen - Murdoch Children Research, AUS
- Angelina Wood - Cambridge Public Health, UK

**Posters - Reception**

The posters will be exposed at Masure (where the coffee breaks will be held) during the meeting.

The Reception will be held on Thursday 18 June 7pm at Masure. It will be an “Apéritif-dinatoire” with many finger foods, cheese (and wine) and some desserts... It is a social event with the opportunity to discuss one-on-one with the presenters of the talks and the posters! Please don’t forget to vote for the best posters!!! Awards will be given for the three best.
Sponsors

The organizers would like to thank the following organisms which sponsored the event and contribute to the success hopefully of this first conference on missing values and matrix completion!

• Among the academic partners:
  
  – **Agrocampus-Ouest**: the ‘elite public university-level college’, for life and environmental sciences and technologies where the conference takes place.
  
  – **Sfds**: the French Statistical Society is an association that aims to promote the use and understanding of statistics, and to foster its methodological development
  
  – **IRMAR**: the Mathematics research institute of Rennes
  
  – **CHL**: the Lebesgue Center of Mathematics is a research center for mathematics in the western part of France
  
  – **Rennes metropole**: the city of Rennes
  
  – **UEB**: the European University of Brittany
  
  – **Crest-Ensai**: the center for research in economics and statistics
  
  – **Inria**: A public science and technology institution dedicated to computational sciences

• Among the sponsoring companies:
  
  – **DIANA** Pet Food, the world’s leading partner committed to improving pet food performance, develops innovative solutions to increase the well-being of cats and dogs, and pet owner satisfaction. Member of the Symrise Group, the Pet Food Business unit relies on a strong establishment in 5 continents with 670 employees, 17 industrial sites, 17 sales offices, 1 research center, 4 local development and technical support centers and 3 cats and dogs expert panels.
  
  – **Mediametrie** is the French industry leader in media research, which observes, measures and analyses audience behaviour and market trends. Created in 1985, Médiamétrie is expanding its activities in television, radio, the internet, film, telecom and the cross-media sector in France and abroad. Médiamétrie generated a turnover of €82,4 million in 2014. Website: www.mediametrie.fr Twitter : www.twitter.com/ Mediametrie Facebook: www.facebook.com/Mediametrie
  
  – **Kuzulia** is a company that provides expertise in applied mathematics to several fields: biology, genetics, ecology, agriculture, sensory studies. The company offers both consulting and adult professional training using R software.
  
  – **XLSTAT** is the richest tool for the data analysis and the statistical treatment with MS Excel. This is available in 7 languages, through 100 countries and has been chosen by more than 50000 users. The Addinsoft company, that develops and distributes the product, is located in Bordeaux (France). This company is growing significantly, especially in USA where the product has been chosen in a lot of universities for teaching.
  
  – **L’Oreal** is the French cosmetics and beauty company the most famous in the world.
  
  – **Rte** is the electricity transmission system operator of France.
  
  – **Ritme** brings scientific softwares’ ease-of-use to companies, laboratories and universities in France, Benelux, Switzerland and Italy. Ritme brings solutions in answer to your needs for greater performance : from guidance in softwares’ selection to products’ & cross-functional trainings or tailored consultancy needs. Our mission is to facilitate yours.
  
  – **Rstudio** is an integrated development environment for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial editions.
  
  – **Revolution** R software and services suite, is the innovative leader in Big Data, enterprise analytics. Powered by the R language, Revolution R is used by enterprises with massive data, performance and multi-platform requirements who need to drive down the cost of Big Data
• **Among the book publishers:**

  – **Wiley** is a publisher of books, journals, books, databases, reference works and laboratory protocols in every major academic, scientific and professional field. The range of statistics and mathematics books can be viewed online at www.wiley.com/statistics. The website www.StatisticsViews.com can be visited for the latest news, features, books and research articles in Statistics.

  – **Chapman & Hall/CRC** is a premier publisher of science books, with a world-leading program in statistics. The core areas of our program are statistical theory and methods, biostatistics, and computational statistics. Visit https://www.crcpress.com/statistics to browse our catalog, and use the promotion code provided in the wallet for a 20% discount.
Conference Program

Wednesday June 17 (Pre-conference tutorials) ________________

- **14h00 - 16h00**: Handling missing data in R with MICE. Gerko Vink and Stefan van Buuren
- **16h00 - 16h30**: Coffee break
- **16h30 - 18h30**: Model-based clustering/imputation with missing/binned/mixed data using the new software MixComp. Christophe Biernacki

Thursday June 18 ________________

- **10h00 - 10h10**: Welcome
- **10h10 - 11h00**: Multiple imputation
  (Chairman: Francois Husson)
  Stef van Buuren, *Fully conditional specification, past, present and beyond*
- **11h00 - 11h30**: Coffee break - posters
- **11h30 - 13h10**: Low-rank matrix completion
  (Chairman: Rémi Gribonval)
  Olga Klopp, *Robust matrix completion*
  Pierre Vandergheynst, *Matrix completion on graphs*
- **13h10 - 13h30**: Lightening talks: dimension reduction, matrix completion, and inpainting
  (Chairman: Mathieu Emily)

  Lunch

- **14h45 - 16h25**: Multiple imputation in biostatistics
  (Chairman: Gilbert MacKenzie)
  Mouna Akacha, *Missing Data in Clinical Trials: Challenges and Recent Developments*
- **16h25 - 16h45**: Lightning talks: inference with missing values
  (Chairman: Francois Husson)
- **16h45 - 17h15**: Coffee break - posters
- **17h15 - 18h05**: Inference and mechanism of missing values
  (Chairman: Gilbert Saporta)
  Roderick Little, *Partially missing at random and ignorable inferences for parameter subsets with missing data*
- **18h05 - 18h25**: Lightening talks: inference and imputation
  (Chairman: Benjamin Guedj)
- **18h25 - 18h45**: Lightening talks: multiple imputation
  (Chairman: Benjamin Guedj)
- **18h45**: Posters Session - Reception
Friday June 19

- **9h30 - 11h10**: Multiple imputation in biostatistics  
  (Chairman: Matthias Templ)  
  Shaun Seaman, *Handling missing values in Matched Case-Control Studies using Multiple Imputation.*  
  Recai Yucel, *Gaussian-based routines for imputing categorical variables in complex data*

- **11h10 - 11h40**: Coffee break - posters

- **11h40 - 13h20**: Survey  
  (Chairman: Guillaume Chauvet)  
  David Haziza, *Triply robust inference in the presence of missing survey data*

  Lunch

- **14h45 - 16h25**: Images  
  (Chairman: Teddy Furon)  
  Patrick Perez, *From Image to Video Inpainting, with Patches*  
  Sandrine Pires, *Missing values in Astrophysics*

- **16h25 - 16h45**: Closing session poster awards
Handling missing data in R with MICE

Gerko Vink and Stefan van Buuren
Dept of Methodology and Statistics, FSS, University of Utrecht
Netherlands Organisation for Applied Scientific Research TNO

Multiple imputation (Rubin 1987, 1996) is a recommended method for complex incomplete data problems. Two general approaches for imputing multivariate data have emerged: joint modeling (JM) and fully conditional specification (FCS) (van Buuren 2007). Multivariate Imputation by Chained Equations (MICE) is the name of software for imputing incomplete multivariate data by FCS which will be presented in this tutorial.

Topics will include: concise theory on multiple imputation - a description of how the algorithm in MICE works - specification of the imputation model - sensitivity analysis under MNAR - interacting with other software.

Prerequisites: elementary knowledge of general statistical concepts and (linear) statistical models is assumed. Moreover, basic programming in R is useful.
Model-based clustering/imputation with missing/binned/mixed data using the new software MixComp

Christophe Biernacki
Lille 1 University, France, INRIA

The “Big Data” paradigm involves large and complex data sets. Complexity includes both variety (mixed data: continuous and/or categorical and/or ordinal and/or functional...) and missing, or partially missing (binned), items. Clustering is a suitable response for volume but it needs also to deal with complexity, especially as volume promotes complexity emergence.

Model-based clustering has demonstrated many theoretical and practical successes (McLachlan 2000), including multivariate mixed data with conditional (Biernacki 2013) or without conditional independence (Marbac et al. 2014). In addition, this full generative design allows to straightforwardly handle missing or binned data (McLachlan 2000; Biernacki 2007). Model estimation can also be performed by simple EM-like algorithms, as the SEM one (Celeux and Diebolt 1985).

MixComp is a new R software, written in C++, implementing model-based clustering for multivariate missing/binned/mixed data under the conditional independence assumption (Goodman 1974). Current implemented mixed data are continuous (Gaussian), categorical (multinomial) and integer (Poisson) ones. However, architecture of MixComp is designed for incremental insertion of new kinds of data (ordinal, ranks, functional...) and related models.

Currently, MixComp is not freely available as an R package but will be soon freely available through a specific web interface. Beyond its clustering task, it allows also to perform imputation of missing/binned data (with associated confidence intervals) by using the mixture model ability for density estimation as well.

Topics will include: mixture models - conditional independence - SEM algorithm - model selection criteria.

Prerequisites: elementary knowledge of general statistical concepts, of mixture models, of EM algorithm and of standard model selection criteria is assumed. Moreover, basic programming in R is useful.
Fully Conditional Specification: Past, present and beyond

Stef van Buuren
Dept of Methodology and Statistics, FSS, University of Utrecht
Netherlands Organisation for Applied Scientific Research TNO

Two traditions have emerged for drawing multiple imputations in multivariate missing data. The joint modeling approach specifies a multivariate distribution of the missing data, whereas fully conditional specification (FCS, or chained equations, sequential regression) requires a specification of each conditional distribution. The lecture highlights how FCS combines algorithmic elements of the NORM method with the philosophy of the French tradition of L’analyse de Donnees. The result is a highly flexible methodology for creating imputations that stay close to the data, but that also lead to valid statistical inferences for a wide variety of complete - data estimates. I will discuss strong and weak points of the approach, open problems in current FCS, and future applications, such as estimating the individual causal effect (ICE), i.e., the answer to the question “What will this intervention do for this person at this time?”
Robust Matrix Completion

Olga Klopp
Paris Ouest University, CNRS

We consider the problem of recovery of a low-rank matrix in the situation when most of its entries are not observed and a fraction of observed entries are corrupted. The observations are noisy realizations of the sum of a low rank matrix, which we wish to recover, with a second matrix having a complementary sparse structure such as element-wise or column-wise sparsity. We analyse a class of estimators obtained by solving a constrained convex optimization problem that combines the nuclear norm and a convex relaxation for a sparse constraint. Our results are obtained for the simultaneous presence of random and deterministic patterns in the sampling scheme. We provide guarantees for recovery of low-rank and sparse components from partial and corrupted observations in the presence of noise and show that the obtained rates of convergence are minimax optimal.

This is a joint work with K. Lounici and A. Tsybakov.
Matrix Completion on Graphs

Pierre Vandergheynst
EPFL, Lausanne, Switzerland

The problem of finding the missing values of a matrix given a few of its entries, called matrix completion, has gathered a lot of attention in the recent years. Although the problem is NP-hard, Candès and Recht showed that it can be exactly relaxed if the matrix is low-rank and the number of observed entries is sufficiently large. In this work, we introduce a novel matrix completion model that makes use of proximity information about rows and columns by assuming they form communities. This assumption makes sense in several real-world problems like in recommender systems, where there are communities of people sharing preferences, while products form clusters that receive similar ratings. Our main goal is thus to find a low-rank solution that is structured by the proximities of rows and columns encoded by graphs. We borrow ideas from manifold learning and signal processing to constrain our solution to be smooth on these graphs, in order to implicitly force row and column proximities. Our matrix recovery model is formulated as a convex non-smooth optimization problem, for which a well-posed iterative scheme is provided. We study and evaluate the proposed matrix completion on synthetic and real data, showing that the proposed structured low-rank recovery model outperforms the standard matrix completion model in many situations.

This is joint work with Vassilis Kalofolias, Michael Bronstein and Xavier Bresson.
Dimension reduction, matrix completion, and inpainting

- *Low Rank Matrix Completion with Exponential Family Noise*, Jean Lafond [Talk and poster]
- *Non-parametric methods for the interpolation of noisy space-time dynamics*, Redouane Lguensat, Pierre Tandeo, Pierre Aillot, Ronan Fablet [Talk and poster]
- *Audio Inpainting, Source Separation, Audio Compression All with a Unified Framework Based on NTF Model*, Cagdas Bilen, Alexey Ozerov and Patrick Perez [Talk and poster]
- *A general low complexity recovery framework and its application to structured sparsity*, Yann Traonmilin, Rémi Gribonval [Talk and poster]
- *Sketching techniques for very large matrix factorization*, Raghavendran Balu, Teddy Furon, Laurent Amsaleg [Poster]
- *Linear embeddings of low-dimensional subsets of a Hilbert space to \( \mathbb{R}^m \)*, Gilles Puy, Mike Davies, Rémi Gribonval [Poster]
The matrix completion problem consists in reconstructing a matrix from a sample of entries, possibly observed with noise. A popular class of estimator, known as nuclear norm penalized estimators, are based on minimizing the sum of a data fitting term and a nuclear norm penalization. Here, we investigate the case where the noise distribution belongs to the exponential family and is sub-exponential. Our framework allows for a general sampling scheme. We first consider an estimator defined as the minimizer of the sum of a log-likelihood term and a nuclear norm penalization and prove an upper bound on the Frobenius prediction risk. The rate obtained improves on previous works on matrix completion for exponential family. When the sampling distribution is known, we propose another estimator and prove an oracle inequality w.r.t. the Kullback-Leibler prediction risk, which translates immediately into an upper bound on the Frobenius prediction risk. Finally, we show that all the rates obtained are minimax optimal up to a logarithmic factor.

References


Non-parametric methods for the interpolation of noisy space-time dynamics

Redouane Lguensat\(^1,\star\), Pierre Tandeo\(^1\), Pierre Ailliot\(^1,2\), Ronan Fablet\(^1\)

1. Institut TELECOM ; TELECOM Bretagne ; UMR 6258 LabSTICC ; Technopôle Brest Iroise CS 83818 29238 Brest Cedex 3
2. University of Western Brittany Brest ; UMR 6205 LMBA ; 6 Avenue Le Gorgeu CS 93837 29238 Brest Cedex 3

\*Contact author: redouane.lguensat@telecom-bretagne.eu

**Keywords:** Stochastic filtering, Inpainting, Analog method, Hidden Markov Model, Data mining.

In this work we present a non parametric method to interpolate noisy multivariate time series and image series with missing data using a data-driven stochastic filtering scheme.

Let \(x(t_k)\) be the dynamic time serie to infer and suppose we have \(y(t_k)\) non-complete noisy measurements of \(x(t_k)\) at each time \(k\). This inverse problem can be formulated by the following equation:

\[
y(t_k) = Hx(t_k) + \epsilon(t_k)
\]

where \(H\) is an observation operator representing a mask (0 for missing data, 1 for presence) and \(\epsilon\) a random perturbation which takes into consideration observation modeling error, noising and other sources of uncertainty. This can be the case where we have missing pixels in an image sequence, or occulted parts in a time serie, depending on the considered application setting (e.g., missing data due to clouds for satellite observations). These measurements are one source of information about the sequence \(x(t)\).

A second source of information about this sequence is provided by the dynamical model that controls the evolution of the dynamic system in time:

\[
x(t_{k+1}) = M(x(t_k)) + \eta(t_k)
\]

where \(M\) denotes the dynamical model and \(\eta\) a random noise representing the uncertainty about the model between two consecutive times. The mathematical resolution of the state-space model (1)-(2) refers to data assimilation in the geoscience community or stochastic filtering more commonly.

To address this issue classical stochastic filtering techniques, such as ensemble Kalman [1] [6] or particle filters and smoothers are used [7]. They proceed by an online evaluation of the physical model in order to provide a forecast for the state. And so it is crucial to have a well defined physical model which is not always the case, thus, the lack of consistency of the model with respect to the observed data and modeling uncertainties are severe limitations of this classical framework.

In our work, we investigate a data-driven methodology as introduced in [2]. The core idea consists in combining machine learning and stochastic filtering methods by using historical datasets which contains examples of the time evolution of the state as such, it exploits an implicit data-driven model. We investigate different stochastic filters, including a discrete reformulation of the proposed framework using Hidden Markov Models. To deal with high-dimensional data (including images), we exploit a PCA-based representation of the data to implement the proposed non-parametric stochastic filtering in a lower-dimensional space. This approach can then be regarded as a dynamic matrix completion setting, which constrains the time consistency of the matrix completion from a database of previously observed data series. Fig. 1 shows a time step of the filtering algorithm. We report numerical experiments for the classical Lorenz-63 chaotic dynamics as well as image time series to demonstrate the relevance of our data-driven approach for the interpolation of missing data in multidimensional time series.

For a lightning talk submission and a poster submission
Audio Inpainting, Source Separation, Audio Compression
All with a Unified Framework Based on NTF Model

Cagdas Bilen, Alexey Ozerov and Patrick Perez

Technicolor
cagdas.bilen, alexey.ozerov, patrick.perez@technicolor.com

Keywords: Audio Inpainting, Source Separation, Informed Source Separation (ISS), Compressive Sampling, Nonnegative Tensor Factorization (NTF)

We propose a general signal recovery algorithm, that can recover $J$ audio source signals in time, $s_{jt}, j \in \{1, J\}, t \in \{1, T\}$, given the quantized source samples, $y_{jt}$, each observed on the support $\Omega^T_j$ and the quantized mixture samples, $x^T_t = \sum_{j=1}^J s^T_{jt} + a_t$, observed on $\Xi_t$ with quantization noise $a_t$. The sources are modelled in the short time Fourier transform (STFT) domain with a normal distribution $s_{jt} \sim \mathcal{N}(0, v_{jt})$ where the variance tensor $V = [v_{jt}]_{j,f,n}$ has the following low-rank non-negative tensor factorization (NTF) structure [7], $v_{jt} = \sum_{k=1}^K q_{jk} w_{fk} h_{nk}$. This model is parametrized by $\theta = \{Q, W, H\}$ with $Q[q_{jk}]_{j,k} \in \mathbb{R}_+^{J \times K}$, $W = [w_{fk}]_{f,k} \in \mathbb{R}_+^{F \times K}$ and $H = [h_{nk}]_{n,k} \in \mathbb{R}_+^{N \times K}$.

We propose to recover the source signals with a generalized expectation-maximization (GEM) algorithm [4] based on multiplicative update (MU) rules [5]. The Algorithm is briefly described in Algorithm 1. Using the proposed approach, it is possible to solve a number of existing and new problems in audio signal processing:

- **Audio inpainting**: It is possible to recover arbitrary time domain losses in audio signals for applications such as signal declipping. NTF model is used for the first time for the recovery of arbitrary time domain losses [3].

- **Joint audio inpainting and source separation**: It is possible to jointly perform audio inpainting and source separation to improve the performance of both tasks. Audio inpainting and source separation are performed jointly for the first time [1].

- **Compressed sampling-based informed source separation**: It is possible to recover the sources from their random samples and the mixture via compressive sampling-based informed source separation. This new ISS scheme uses a simple encoder that has properties of distributed coding [8, 6] and it competes with traditional ISS. The concept of distributed coding and of compressive sampling based scheme is introduced for the first time in the informed source separation problem [2].

The presentation and the poster will include various new results for the proposed algorithm with comparisons to state of the art methods in the different problems discussed above.

---

* The first and second authors have contributed equally for this work.

This work was partially supported by ANR JCJC program MAD (ANR-14-CE27-0002).

This work is submitted for a lightning talk and a poster.

Algorithm 1 GEM algorithm for NTF model estimation. Details can be found in [1, 2, 3]

1: procedure ESTIMATESOURCES-NTF($\Xi^T, x^T, \{y^T_{jt}\}_{j=1}^J, \{\Omega_{jt}\}_{j=1}^J$)
2: Initialize non-negative $Q$, $W$, $H$ randomly and apply any known prior on $Q$, $W$, $H$
3: repeat
4: Estimate sources ($s_{jt}$), given $Q$, $W$, $H$, $x^T$, $\Xi^T$, $\{y^T_{jt}\}_{j=1}^J$, $\{\Omega_{jt}\}_{j=1}^J$ with Wiener filtering
5: Apply constraints in the time domain, and estimate posterior power spectra of the sources ($\hat{P}$)
6: Update $Q$, $W$, $H$ given $\hat{P}$ using MU rules
7: until convergence criteria met
8: end procedure
A general low complexity recovery framework and its application to structured sparsity

Yann Traonmilin1*, Rémi Gribonval1

1. INRIA, Rennes
*Contact author: yann.traonmilin@inria.fr

Keywords: sparse recovery, compressed sensing, restricted isometry, structured sparsity

To estimate data from underdetermined observations, a regularity assumption is needed. Sparse recovery theory gives a powerful framework to perform this estimation. Recent advances in this domain show that this estimation is possible with general low complexity regularity models which are suitable to model many kinds of data. Consider underdetermined observations of a signal \( x \in \mathcal{H} \) (\( \mathcal{H} \) is a Hilbert space) by a linear operator \( M \). The regularity of \( x \) is defined by the fact that \( x \in \Sigma \), where \( \Sigma \) is the regularity model set. Usually, \( \Sigma \) enforces low complexity on the signal \( x \): few non-zero coefficients for sparse vectors, few non-zero singular values for low rank matrices, etc. Let \( f \) be a regularization function. It was proven that the program

\[
\arg\min_{y \in E} f(y) \quad s.t. \quad My = Mx
\]

(1)
is able to recover \( x \) for a selection of \( M \), \( \Sigma \) and convex norms \( f \) [6]. We consider the following question: Given a regularity model set \( \Sigma \), what regularizer \( f \) permits the recovery of low complexity signals in \( \Sigma \)?

To tackle this question, we look for a (preferably convex) regularizer \( f \) for which recovery can be guaranteed for a large set of operators \( M \). It is known that operators \( M \) with a Restricted Isometry Property (RIP) allow for a stable uniform recovery of all \( x \in \Sigma \) when \( f \) is the \( \ell^1 \) norm and \( \Sigma \) is the set of \( K \)-sparse vectors [5]. Many RIP results are available for different pairs of model set \( \Sigma \) and norm \( f \). Each of these aiming at either generalizing the area of application or the sharpness of hypotheses. For example, Cai and Zhang [4] gave a sharp RIP result valid for sparse vector and low rank matrix recovery. In fact, Bourrier et al. [3] showed that a lower RIP is a necessary condition for the robust recovery of signals with arbitrary model set \( \Sigma \). This suggests that the RIP is a good tool to study low complexity recovery.

In this contribution, we show that under some generic geometric conditions on the regularizer \( f \) and the model set \( \Sigma \), we can provide a uniform recovery theorem on \( \Sigma \) under RIP conditions for \( M \). This result gives a framework to check if a given regularizer is a good candidate to perform low complexity recovery, and get a compressed sensing result when combining with the work of Puy et al. [7], where it was shown that the RIP holds on very generic model sets, even in infinite dimension. Given a model set \( \Sigma \) and a function \( f \), we can get a uniform recovery result for random observations of signals in \( \Sigma \) by checking the RIP of random observations on the secant set \( \Sigma - \Sigma \) and verifying our geometric conditions.

Application: structured sparsity in an infinite dimensional space. We use this general framework to give sharp and uniform recovery results in an infinite dimensional context with structured sparsity inducing norms without overlap (or group norms). Ayaz et al. [2] gave an uniform recovery result for structured compressed sensing under RIP hypothesis in a finite dimensional setting. In this case, the regularizer is a mixed \( \ell^1 - \ell^2 \) norm (group norm) and a RIP constant of \( \sqrt{2} - 1 \) for vectors in \( \Sigma - \Sigma \) guarantees recovery of vectors from \( \Sigma \). In [1], Adecoek and Hansen propose a generalized sampling strategy to recover sparse signals in an infinite dimensional setting. With our new general framework, we show that the RIP constant on group sparse vectors of Ayaz et al. can be improved to the sharp \( \frac{1}{\sqrt{2}} \). We generalize these results to non-overlapping group norms in an infinite dimensional context which encompass the generalized sampling of [1]. This setting is important for problems were the data lives in an infinite dimensional space. For example, in magnetic resonance imaging (MRI), Fourier coefficients of the image are observed, this image being an element of the function of finite energy \( L^2 \) which has infinite dimension.
Sketching techniques for very large matrix factorization

Raghavendran Balu, Teddy Furon, Laurent Amsaleg

Inria, Rennes

Contact author: raghavendran.balu@inria.fr

Keywords: Count sketch, online algorithms, matrix factorization, recommender system, collaborative filtering

Matrix factorization is a prominent technique for approximate matrix reconstruction and noise reduction. Its appeal is attributed to its space efficiency and its ability to generalize with missing information. Our focus is matrix factorization technique applied to collaborative filtering [2], where the given data is a sparse matrix \( R \) with non zero entries \( r_{u,i} \) representing the rating provided by the user \( u \) for a given item \( i \).

Factorization of the matrix \( R \) maps both users \( u \in U \) and items \( i \in I \) in a joint vector space \( \{ p_u \in \mathbb{R}^d, q_i \in \mathbb{R}^d \} \) in such a way that their inner product \( p_u^\top q_i \) approximates the observed value \( r_{u,i} \) [2]. The learned latent factors are thus used to predict (by inner product) the missing entries and thereby to fill the incomplete user-item matrix. The objective is to minimize the error function \( \sum_{\text{observed}} (r_{u,i} - p_u^\top q_i)^2 + \lambda (\|p_u\|^2 + \|q_i\|^2) \).

In collaborative filtering systems, the sparse matrix \( R \) is too huge to be stored with the number of users and items growing to millions. The observed elements in \( R \) are not available as a static data beforehand, but as a stream of tuples \( <u, i, r_{u,i}> \). One has to learn online and store the latent factors alone.

In regular matrix factorization, latent vectors \( \{ p_u \}_{u \in U} \) and \( \{ q_i \}_{i \in I} \) are stored as contiguous dense arrays in memory. Yet, the number of latent factors is linear with the number of users and items. This also raises problem in a dynamic environment when new users and items are added every now and then. We propose replacing this matrix representation with count sketch to improve and control the storage efficiency. Count sketch is a probabilistic data structure to maintain frequency information of items approximately but with sub linear space complexity [1]. More surprisingly, although user and item vectors carry different semantic, their underlying representations are the same and we store both of them in the same count sketch structure designed for \( d(|U| + |I|) \) elements.

The sketch based online factorization is not much different from regular online factorization [2] other than the latent factor retrieval and gradient updates merging. When a new tuple \( <u, i, r_{u,i}> \) arrives, the count sketch is queried to reconstruct user and item vectors \( \tilde{p}_u \) and \( \tilde{q}_i \). The rating estimation \( \tilde{p}_u^\top \tilde{q}_i \) is compared with the actual rating \( r_{u,i} \) to get the error and gradient updates for all the \( 2d \) values of the user and item vectors. Then for each component of \( \tilde{p}_u \) as well as \( \tilde{q}_i \), the count sketch structure \( C \) is updated with their corresponding sign corrected gradients.

We show that this simple approach is indeed equivalent to the online minimization of a functional combining the prediction error and the estimation noise. This estimation noise inherent to count sketch has some regularization properties exhibited by experiments on large scale systems (up to 100 M of ratings).

References


Linear embeddings of low-dimensional subsets of a Hilbert space to $\mathbb{R}^m$

Gilles Puy $^{1,*}$, Mike Davies $^2$, Rémi Gribonval $^1$

$^1$. INRIA Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France. $^2$. The University of Edinburgh, Edinburgh EH9 3JL, UK.

*Contact author: gilles.puy@inria.fr

Keywords: Compressed sensing, Restricted isometry property, Box-counting dimension.

A particular case of problems with missing data are inverse problems that appear in compressed sensing (CS). In this field, the restricted isometry property (RIP) characterises the possibility to reconstruct sparse signals from their measurements. In this work, we are interested in extending known CS results in finite dimensional ambient space to infinite dimensional ambient space. These developments are important, e.g., in CS to extend the theory to an analog setting [1] and explore connections with the sampling of signals with finite rate of innovation [7], and also in machine learning to develop efficient methods to compute information-preserving sketches of probability distributions [6, 2].

We consider the problem of embedding a low-dimensional set, $\mathcal{M}$, from an infinite-dimensional Hilbert space, $\mathcal{H}$, to $\mathbb{R}^m$. Defining appropriate random linear projections, we construct a linear map which has the RIP on the secant set of $\mathcal{M}$, with high probability for a number of projections essentially proportional to the intrinsic dimension of $\mathcal{M}$. In this abstract, $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathcal{H}$ and $\|\cdot\|$ is the associated norm. Our goal is to construct a linear map $L: \mathcal{H} \to \mathbb{R}^m$ that satisfies

$$
(1 - \delta) \|x_1 - x_2\| \leq \|L(x_1 - x_2)\|_2 \leq (1 + \delta) \|x_1 - x_2\|
$$

for all pairs of vectors $x_1, x_2 \in \mathcal{M}$. This is equivalent to showing that $\sup_{y \in S(M)} |\|L(y)\|_2 - 1| \leq \delta$ where $S(M)$ is the normalized secant set of $\mathcal{M}$: $S(M) := \{ y = x_1 - x_2/\|x_1 - x_2\| \mid x_1 \neq x_2 \in \mathcal{M} \}$.

In this work, we assume that this normalized secant set has upper box-counting dimension strictly bounded by $s > 0$. Therefore, there exists $\epsilon_0 > 0$ such that, for all $\xi \in (0, \epsilon_0)$, $S(M)$ can be covered by at most $\xi^{-s}$ balls of radius $\xi$. We now fix a value $\epsilon$ in $(0, \epsilon_0)$, choose a cover of $S(M)$ by balls of radius $\epsilon$, and denote by $T(\epsilon)$ the set of centers of these balls. We construct our linear map as follows.

Let $V \subset \mathcal{H}$ be the finite-dimensional linear subspace spanned by $T(\epsilon)$ and $(a_1, \ldots, a_d)$ be an orthonormal basis for $V$. We draw $m$ independent centered isotropic random gaussian vectors $a_1, \ldots, a_d \in \mathbb{R}^d$, set $l_i := \sum_{k=1}^d \alpha_i^k a_k / \sqrt{m}$, for all $\forall i \in \{1, \ldots, m\}$, where $\alpha_i^k$ is the $k^{th}$ entry of $\alpha_i$, and define

$$
L_\epsilon: \mathcal{H} \to \mathbb{R}^m
\quad
x \mapsto L_\epsilon(x) = (\langle l_1, x \rangle, \ldots, \langle l_m, x \rangle)^T.
$$

**Theorem 1.** There exist absolute constants $D_1, D_2, D_3 > 0$ with $D_1 < 1$ such that if $S(M)$ has upper box-counting dimension strictly bounded by $s$, then for any $0 < \delta < \min(D_1, \epsilon_0)$ and $\rho \in (0, 1)$, if $\epsilon \leq \delta/2$ and

$$
m \geq D_2 \delta^{-2} \max \left\{ s \log \left(D_3 / \delta\right), \log \left(6 / \rho\right) \right\},
$$

then $L_\epsilon: \mathcal{H} \to \mathbb{R}^m$ satisfies the RIP (1) with constant at most $\delta$ and probability at least $1 - \rho$.

The proof is based on a chaining argument which is a powerful technique to obtain sharp bounds for the supremum of random processes [5, 3]. A detailed proof can be found in [4].
Missing Data in Clinical Trials: Challenges and Recent Developments

Mouna Akacha
NOVARTIS, Basel, Switzerland

Health authorities are increasingly concerned about the impact of missing data on clinical trial results. In 2010, both US and European regulators released or commissioned documents that emphasize the prevention of missing data, request continued data collection after study treatment discontinuation and discuss the need for a clearly defined primary objective (‘estimand’) of the trial.

Indeed, defining the primary objective of a clinical trial in the presence of non-compliance or non-adherence to the assigned treatment is crucial for the choice of design, the statistical analyses - including the handling of missing data - and the interpretation of the results.

At first glance this seems obvious, however, primary objectives stated in clinical trial protocols often fail to give a precise definition of the measure of intervention effect. In particular, the impact of missing data is frequently not taken into account when defining the intervention effect of interest. However, different ways of dealing with missing data may well imply different measures of effect.

We will review these aspects and discuss why “missing at random” based analyses may be unsuitable in certain clinical trial settings. We illustrate our ideas based on a case study and share recent health authority interactions on this topic.
Advances in Research on Participant Attrition in Prevention Intervention Studies

John W. Graham, Lauren E. Connell, and Michael L. Hecht
Penn State University

This paper describes a promising approach to dealing with participant attrition, a methodological challenge common in prevention intervention research. Attrition undermines the internal validity of studies evaluating the efficacy/effectiveness of preventive interventions. Reducing the impact of attrition is key to these essential evaluations.

Graham (2012) suggested several strategies for studying attrition, and outlined a taxonomy of eight cases of attrition that consider all possible combinations of the treatment (vs. control: T), the dependent variable (Y), and the interaction between them (TY) as causes of missingness. Attrition is not a problem in two cases (case 1 = not T, not Y, not TY as causes; case 2 = T only as cause), because missingness is either MCAR or MAR. Previous research (Collins, Schafer, & Kam, 2001; Graham et al., 2008; Graham, 2012) has shown minimal attrition bias in a third case for the regression coefficient for T predicting Y ($\beta_Y T$) (Case 3: Y only as cause). The remaining five cases have not been studied.

This paper describes results of a Monte Carlo simulation that extends the work of Collins, Schafer, & Kam (2001) to cover the five cases of the taxonomy that have not been studied previously in depth. Our simulation focuses on case 8, in which T, Y, and the TY interaction are all responsible, to varying degrees, for the observed missingness on Y. Our work focused on what Collins et al. (2001) referred to as MNAR-Linear missingness, and we created missingness using the same strategy they used. We studied several key factors relating to missingness: overall % missing; the T effect (difference in % missing between T and C groups); the overall Y effect (using a quantity we call “range”: the difference in the probability of Y missing between the 4th and 1st quartiles of Z); and the TY effect (difference in “range” between T and C groups). We also examined two values for the quantity $\rho_{ZY}$, the correlation between the cause of missingness, Z, and the main dependent variable, Y. Our work makes use the definition suggested by Collins, Schafer, & Kam (2001) for missing data bias that is not significant in a practical sense; we judged bias not to be of practical significance as long as the absolute value of the standardized bias was less than 40. Results of our simulation show clear patterns of demarcation between combinations of factors that do and do not show bias that is significant in a practical sense. My presentation will present these results.

Estimating the various quantities (% missing; T effect; Y effect, TY effect; $\rho_{ZY}$) in empirical data is key to estimating the impact of bias due to attrition. The % missing, T effect, and $\rho_{ZY}$ quantities are all easily estimated in empirical studies with longitudinal follow-up data. The Y and TY effects are more difficult to estimate, but with longitudinal follow-up data, plausible estimates of these quantities are possible. Making judicious use of the estimates of all these quantities in empirical data, in conjunction with careful sensitivity analyses, one can make judgements in a particular empirical study about the likelihood that attrition has caused bias that is significant in a practical sense.

Graham (2012) also suggested data collection methods for reducing the impact of attrition. One strategy involves asking attrition relevant questions (e.g., whether the participant will be available for the next wave of measurement). The benefits of this strategy have not been studied previously in empirical research. Using our simulation results, and our strategies for estimating the various quantities in empirical data, we demonstrate the benefits of using the attrition-relevant measures to reduce attrition bias.
**Inference**

- *Longitudinal data analysis with auxiliary item information to deal with missing questionnaire data*, Iris Eekhout, Craig K. Enders, Jos W.R. Twisk, Michiel R. de Boer, Henrica C.W. de Vet, Martijn W. Heymans [Talk and poster]

- *Variable selection using the integrated complete-data likelihood to cluster dataset with missing values*, Matthieu Marbac, Mohammed Sedki [Talk and poster]

- *Estimation of the AIDS incidence rate in Portugal via the backcalculation method*, Julio Silva, Joaquim Costa, A. Rita Gaie [Talk and poster]

- *Conditional moments and data missing at random*, Marian Hristache, Valentin Patilea [Talk and poster]

- *Systematically missing data in meta-analysis: a graphical model approach*, Jelena Kovacic, Veda Marija Varnai [Poster]
Longitudinal data analysis with auxiliary item information to deal with missing questionnaire data.

Iris Eekhout¹²*, Craig K. Enders³, Jos W.R. Twisk², Michiel R. de Boer⁴, Henrica C.W. de Vet², Martijn W. Heymans²⁴

1. Research center – Military Mental Health, Ministry of Defense, Utrecht, the Netherlands; Brain Center Rudolf Magnus, Department of Psychiatry, University Medical Center Utrecht, Utrecht, the Netherlands.
2. VU University medical center, Amsterdam, the Netherlands and EMGO Institute for Health and Care Research, Amsterdam, the Netherlands.
3. Arizona State University, Tempe, AZ.
4. VU University, Amsterdam, the Netherlands.
*Contact author: i.eekhout@umcutrecht.nl

Keywords: structural equation modeling, questionnaires, longitudinal data, full information maximum likelihood, auxiliary variables

Study outcomes are often measured by multi-item questionnaires. Missing item scores can impair total score calculation and result in incomplete scale scores. Even though scale scores are of interest, it is most desirable to handle missing items at the item-level because that improves power [1,2]. Many questionnaire manuals advice to average the available items or impute the person mean score, however these methods can introduce bias. A valid way to handle missing data without imputation is by using a structural equation model with maximum likelihood estimation. For longitudinal data this could be a linear growth model. However, these kinds of models do not necessarily include the item scores. The aim of this study was to propose two novel methods for dealing with incomplete scale scores that result from missing item responses by including the item information in the model as auxiliary variables [3]. These methods are advantageous because they incorporate the observed item-level data while maintaining the sum scores as the outcome of the analysis. The auxiliary variables improve missing data handling by using the item information to improve the precision of model estimates.

The two methods that were evaluated included either the item scores themselves or a summary of a parcel of item scores (e.g., the average of the available item scores) as auxiliary variables, while treating the scale scores missing in a latent growth model. The performance of the novel methods was examined in a simulation study in several longitudinal data conditions and analyzed through bias, mean square error, and coverage.

Results showed that including the item information as auxiliary variables results in rather dramatic power gains compared to not including the item information under varying conditions. Including the observed item information as auxiliary variables is the most optimal way to regain power, refine parameter estimates and decrease bias when scale scores are incomplete.

In general, we recommend including a summary of the items in the auxiliary part of the latent growth model when incomplete scale scores result from missing item scores. This study shows that the summary of the items improved the precision of the estimates over not including auxiliary information. Furthermore, the inclusion of a summary is an efficient method that does not over-complicate model estimations.

References

For a lightning talk submission and a poster submission
Variable selection in cluster analysis is important yet challenging. Indeed, in many cases, the partition among the individuals may be best explained by only a subset of the observed variables. So, by performing the variable selection in the cluster analysis, both model fitting and result interpretation are facilitated. This objective can be achieved by regularization methods like the sparse K-means algorithm. This approach has been proposed by Witten and Tibshirani [4] to cluster continuous data sets. It realizes a trade-off between the clustering accuracy and the number of selected variables by adding a lasso-type penalty to the classical K-means method. Unfortunately, the calibration of the penalty term can suffer from criticisms. Moreover, the management of missing values by regularization methods stays a difficult task.

In this context, model selection methods are an efficient alternative. When the variables are continuous, clustering can be achieved by variable selection in a Gaussian mixture model. However, it needs a difficult optimization of an information criterion which involves combinatorial problems. Raftery and Dean [3] also propose to find the set of relevant variables maximizing the BIC criterion by using a greedy search algorithm. However, their algorithm is based on a suboptimal procedure (e.g. stepwise method) and needs multiple calls of EM algorithms for each model comparison that involve a long-time procedure for the model estimation.

Recently, The MICL criterion (Maximum Integrated Compete-data Likelihood) has been introduced by Marbac, M. and Sedki, M. [2] to facilitate the variable selection in a Gaussian mixture model. The authors assume conditional independence between variables to implement a simple procedure for assessing the model maximizing the MICL criterion. The main advantage of this procedure is to avoid the maximum likelihood inference for each competing model. Thus, parameter inference is performed only for the unique selected model. This talk presents an extension of [2] for dealing data sets with values missing at random. The poster presents the management of data sets with continuous, counting, categorical and mixed-type data. A real data set of life science is used to illustrate the twofold benefit of the variable selection: easier interpretation of the components and better partitioning accuracy. The proposed method is implemented in the R package VarSelLCM [1].

References


For a lightning talk submission and a poster submission
Estimation of the AIDS incidence rate in Portugal via the backcalculation method

Júlio Silva¹, Joaquim Costa¹,², A. Rita Gaio¹,²

1. Departamento de Matemática, Faculdade de Ciências da Universidade do Porto
2. Centro de Matemática da Universidade do Porto

Contact author: juliolisboa14@gmail.com

Keywords: HIV/AIDS, Backcalculation, EM algorithm, Newton-Raphson method, Poisson distribution

The Human Immunodeficiency Virus (HIV) can evolve to severe illness with the late symptoms referred to as Acquire Immune Deficiency Syndrome (AIDS). This infection can have a major impact on socio–demographic and economic features of the affected countries. In Portugal, as in most countries, the epidemic is monitored by a surveillance system based on cases notification by physicians. It can take several months until the diagnosed cases are notified. Because HIV infection is usually not immediately detected, good data on incubation times are only available for specific cohorts, in principle not representative of the general population. The purpose of this work is to estimate the 2000-2012 Portuguese incidence rate of AIDS, from notification data and taking into account the problems mentioned above. Public Health policies tell us that all efforts should focus primarily on the improvement of HIV prevention. The knowledge of the incidence rate will be crucial to evaluate current and previous prevention campaigns and, if necessary, can lead to new programs of action.

The estimation of the incidence rate was addressed by the backcalculation method. For a fixed month, the number of newly infected individuals in the previous months is unknown. It was estimated from the number of individuals diagnosed at that month and from the distribution of the incubation time (taken to be the difference between the diagnosis date and the infection date). The latter was assumed to follow a Weibull or a Gamma distribution, see [1]. The number of new diagnoses and new infected cases in any given month was assumed to follow a Poisson distribution. The method was based on a convolution equation adapted for the discrete case, see [2]. The likelihood function was considered for the complete data, including the mean number of new infections per month as unknown parameters, which are to be estimated given the number of new diagnoses per month. Maximization of the complete likelihood used a combination of the expectation-maximization algorithm [3] and linear regression, in order to ensure smoothing. Results from the methodology were tested on simulated and real data.

Correction for the reporting delays was obtained by conditional likelihood estimation for count data assuming a Poisson distribution, as previous work in [4] and [5].

References


For a lightning talk submission and a poster submission
Conditional moments and data missing at random

Marian Hristache\textsuperscript{1,*}, Valentin Patilea\textsuperscript{1}

\textsuperscript{1} CREST (ENSAI-UEB)

\textsuperscript{*} Contact author: marian.hristache@ensai.fr

\textbf{Keywords:} Double-robustness, Inverse probability weighting, MAR assumption, Moment restrictions models, Semiparametric regression.

We consider a general statistical model defined by moment restrictions when data are missing at random. Using the inverse probability weighting, we show that such a model is equivalent with a model for the observed variables only, augmented by a moment condition defined by the missing mechanism. We therefore extend a result stated by Graham (2011) for the equivalence of the corresponding efficiency bounds. In particular, our framework covers parametric and semiparametric mean regressions and quantile regressions. We allow for missing responses, missing covariates and any combination of them. The equivalence result is obtained under minimal technical conditions and sheds new light on various aspects of interest in the missing data literature, as for instance the double-robustness, imputation, the restricted estimators and the puzzling efficiency of the procedures based on the estimated, rather than known, selection probabilities. It also provides a general framework for building (efficient) estimators. Moreover, as an immediate consequence of our general theory, we derive the efficiency of the complete cases analysis in a general semiparametric regression model with responses missing at random. One can see Hristache and Patilea (2014) for more details.

\textbf{References}


Systematically missing data in meta-analysis: a graphical model approach

Jelena Kovačić1,⋆, Veda Marija Varnai1

1. Institute for Medical Research and Occupational Health, Zagreb

⋆Contact author: jkovacic@imi.hr

Keywords: meta-analysis, systematically missing data, graphical models

Systematically missing data in meta-analysis arise when studies measure different sets of adjusting variables. Common approaches to this problem can cause bias (using only variables measured in all studies) or loss of precision (using only studies with all variables; complete-case approach). A previous study showed that exploiting correlation between partially and fully adjusted regression coefficient estimates in a bivariate meta-analysis increases precision [5]. We propose a graphical model that uses these correlations and, additionally, hypothesized conditional independence between differently adjusted estimates to calculate fully adjusted estimate. Unlike previously proposed approaches [5, 3, 2], full data is not required; only estimated coefficients with standard errors. Moreover, even studies with only one reported coefficient can be included in the analysis.

The proposed graphical model is illustrated in Fig.1. Node $X_i$ in a graph corresponds to regression coefficient adjusted for the set of variables $S_i$, and edge from $X_i$ to $X_j$ indicates that $S_j \subset S_i$. Joint density is derived from graph structure (Fig.1) and nodes’ distributions, which we specify as $X_i \sim N(\beta_i, \sigma_i^2 + \tau_i^2)$ if $X_i$ has no parents in the graph, or $X_i|X_j \sim N(\beta_i + \rho_{ij} \sqrt{\sigma_i^2 + \tau_i^2}/(\sigma_j^2 + \tau_j^2)(X_j - \beta_j), (\sigma_i^2 + \tau_i^2)(1-\rho_{ij}^2))$ if its parent is $X_j$. Here $\sigma_i^2$ and $\tau_i^2$ are within- and between-study variances, and $\rho_{ij}$ correlation coefficient between $X_i$ and $X_j$. Parameters are then estimated either by Markov Chain Monte Carlo sampling or maximum likelihood, both in a standard way.

The method was tested using data for 11 countries from the Survey of Health, Ageing and Retirement in Europe [1]. As our goal was solely to develop methodology, we repeat the analysis by Siegriest et al. [4]. The analysis aimed to estimate the effect of effort-reward imbalance on intended early retirement, adjusted for 9 covariates representing sex, age, socio-economic status, physical and mental health parameters, and work conditions (3 binary, 3 3-level, and 3 continuous variables; Fig.1). We considered two missing data patterns, one monotone (Fig.1 a) and the other non-monotone (Fig.1 b). For each pattern, countries were randomly assigned to missing data categories for 3 times. The proposed graphical model (G-MA) was compared to meta-analysis of all 11 fully-adjusted estimates ("gold standard") and complete-case meta-analysis with normal distribution-based confidence intervals (CC-MA; conventional approach). Graphical model estimates were based on 50 000 iterations from Markov Chain Monte Carlo analysis, with uninformative priors for coefficients’ means ($\beta_i$) and between-study variances.

Meta-analysis of all 11 fully-adjusted estimates yielded an odds ratio estimate of 1.87, indicating that higher effort-reward imbalance increases odds of retiring early. In both missing pattern settings (Fig 1.a), G-MA showed less bias than CC-MA (Table 1). In one example with simple missing pattern 95% G-MA credible interval had the same width as the 95% CC-MA confidence interval, while in all other cases CC-MA intervals were wider, by 16% to 55%. None of the G-MA models indicated any lack of convergence.

To conclude, the proposed approach showed less bias and more precision than the conventional approach in our examples, while at the same time requiring less data than alternative approaches. It is easily extended to more complex, non-monotone missing data settings, all types of regression models with any kind of variables, and fixed-effects meta-analysis, while also allowing for incorporation of prior knowledge. All of this makes it suitable for a variety of applications.

For a poster submission
Partially Missing At Random and Ignorable Inferences for Parameter Subsets with Missing Data

Roderick Little
University of Michigan

For likelihood-based inferences from data with missing values, models are generally needed for both the data and the missing-data mechanism. Modeling the mechanism can be difficult, because its parameters are often poorly identified and results are often sensitive to model misspecification. Rubin (1976) showed that sufficient conditions for ignoring the missing data mechanism are that (a) the missing data are missing at random (MAR), and (b) the parameter of the data model and the missing-data mechanism are distinct. These conditions are not always necessary, and they relate to the complete set of parameters in the model. We propose definitions of partial MAR and ignorability for a subvector of parameters of particular substantive interest. We apply these definitions to regression with missing covariates, where an ignorable likelihood method is applied to a subsample of observations that are complete on a set of variables, but possibly incomplete on others. This method gives valid estimates for mechanisms where both complete-case analysis and full ignorable likelihood methods are biased.
Imputation – inferences

- **A case study of completion of cats’ consumption missing data on preference mapping**, Julien Rogues [Talk and poster]

- **Linear regression with missing data and unknown colored noise: Application to a space mission in fundamental physics**, Quentin Baghi [Talk and poster]

- **The Effect of Missing Visits on GEE, A Simulation Study**, Julia Geronimi, Gilbert Saporta [Talk and poster]

- **Recovering missing individual block information in a multiblock Multiple Factor Analysis**, Valentin Voillet, Ignacio Gonzalez, Laurence Liaubet, Phillipe Besse, Magali Sancristobal [Talk and poster]

- **QRILC: a quantile regression approach to left-censored missing data imputation in quantitative high-throughput proteomics**, Cosmin Lazar, Myriam Ferro, Christophe Bruley, Thomas Burger [Poster]

- **Integrating Data Deficient species in evolutionary history loss analyses**, Simon Veron, Sandrine Pavoine, Philippe Clergeau [Poster]

- **Analyzing multi-group multivariate data with non-ignorable missingness**, Myrsini Katsikatsou, Jouni Kuha, Irini Moustaki [Poster]
Dealing with missing data in palatability measurement: a case study with external preference mapping

Julien Rogues

1. DIANA Pet Food, ZA du Gohélis, 56250 Elven, France
Contact author: jrogues@diana-petfood.com

Key words: missing data, palatability, preference mapping, pet food, cats

Ensuring that pet foods are attractive to companion animals and their owners is essential. Classical methods to assess palatability performance include preference trials conducted with dogs and cats expert, semi-expert or in-home panels. A huge amount of information such as animal daily consumption data can be collected from these tests, nevertheless it recurrently happens that some data are not available and/or reliable due to veterinary cares or unusual animal behaviors. A simple way to solve this issue is to calculate the overall mean of the database and replace the missing values with the calculated mean. However several alternative and more precise methods exist. The goal here was to find out which completion method in the literature fitted the best to these consumption data.

A database was set up in order to compare the performance of 60 cat foods representative of the market. Products palatability was assessed using Liking tests conducted on a forty adult cat panel at Panelis, DIANA Pet food expert center in palatability measurement. In addition, all the diets were characterized according to their physical and chemical specificities. A multivariate analysis was performed with the physico chemical data in order to create a bi-dimensional map of products. Palatability data, illustrated by the cat consumption levels, were then used to create an external preference mapping. On the total number of 2400 cat’s consumption values expected, 7 values from 4 out of the 40 cats of the panel were missing in the database (0.3% of missing values). Since preference mapping method requires removing from the global analysis every cat having at least one missing value, only 36 cats could thus be represented in the map (10% “missing cats”).

In order to evaluate different completion methods, a new database with no missing values was created from the original one by taking into account the cats with no missing values: 60 x 36 = 2160 data available. 17 levels of missing values, from 0.2% to 60% of the total number of recorded values, were simulated on this complete database using 7 completion methods selected from the literature: the mean of the global database, the last observation carried forward (LOCF), the nearest neighbors (kNN) [1], the local regression (LOESS) [2], the singular value decomposition (SVD) [2], the random forests (missforest)[3] and the impute PCA [4]. The mean of absolute differences between real data and reconstituted data - the absolute error of imputation - was the criterion considered to compare the efficiency of the methods. 1000 different missing values simulations were performed with each method and for each level of missing values, in order to obtain a reliable value for this criterion.

ImputePCA, missforest and kNN were the most relevant methods to replace missing data in the database considered (Figure 1). The 7 missing values of the original database were successfully replaced using ImputePCA and allowed performing the external preference mapping on the forty initial cats, thus recovering the most of all the information gathered.

These completion methods could be interestingly tested to recover various type of palatability missing data obtained from classical two-bowls palatability test, during a Bradley-Terry model or even with some specific measurement methods developed at Panelis (specific behavioral recording, kinetics of consumption, treats trials…). They could possibly be used when using incomplete block design.

Poster and lightning talk
Linear regression with missing data and unknown colored noise: Application to a space mission in fundamental physics

Quentin Baghi1,∗, Gilles Métiris2, Joël Bergé1, Bruno Christophe1, Manuel Rodrigues1

1. ONERA - The French Aerospace Lab, 29 avenue de la Division Leclerc, 92320 Chatillon, France
2. Observatoire de la Côte d’Azur, Bt 4, 250 rue Albert Einstein, Les Lecules 1, Sophia Antipolis, 06560 Valbonne, France
∗Contact author: quentin.baghi@onera.fr

Keywords: Linear regression, Colored noise, Data imputation, Spaceborne instruments, Experimental tests of gravitational theories.

The analysis of measurements in experimental physics is often subject to interruptions in the acquired data, especially for spaceborne experiments where long integration time may be needed. We show that the uncertainty of a regression analysis on incomplete stationary time series can be increased by several orders of magnitude compared to the complete data case, when the residuals are highly correlated and unmodeled. We present a linear regression method that constructs an approximate general least squares estimator by first fitting an autoregressive (AR) model to the noise [1]. The use of an AR model is efficient to estimate arbitrary noise spectra in spite of the missing data, in particular when using Burg’s algorithm [3]. The state-space representation of the noise provided by the AR model allows us to orthogonalize the data with a Kalman filter. We then estimate the parameters of interest using the data decorrelated in this way, ensuring that the precision is close to the minimal variance bound. While the direct maximization of the likelihood [4] or even Expectation-Maximization procedures [6] can be computationally expensive for long time series, the proposed approach requires very few iterations. In addition, the outputs of the developed algorithm can be used to estimate the missing values using Gaussian conditional expectations. The reconstruction is precise compared to methods based on sparse decomposition [2] but relies on a more restrictive model for the deterministic part of the signal. We demonstrate the performance of the method with simulated time series of the MICROSCOPE space mission [5], an in-orbit experiment scheduled for launch in 2016 that aims to test the equivalence principle. We show that the precision of the fit is 60 times better than ordinary least squares methods.

References


For a lightning talk submission and a poster submission
The Effect of Missing Visits on GEE, A Simulation Study

Julia Geronimi\textsuperscript{1,2,⋆}, Gilbert Saporta\textsuperscript{1}

1. Cedric-Cnam, 292 rue Saint Martin 75141 Paris Cedex 03
2. Institut de Recherches Internationales SERVIER, 50 rue Carnot 92150 Suresnes

Contact author: geronimi.julia@gmail.com

Keywords: Longitudinal data, repeated correlated data, correlation, missing data, simulations, Generalized Estimating Equation

Clinical research is often interested in longitudinal follow-up over several visits. All scheduled visits are not carried out and it is not unusual to have a different number of visits by patient. The Generalized Estimating Equations can handle continuous or discrete autocorrelated response. The method allows a different number of visits by patients. The GEE are robust to missing completely at random data, but when the last visits are fewer, the estimator may be biased. We propose a simulation study to investigate the impact of missing visits on the estimators of the model parameters under different missing data patterns. Different types of responses are studied with an exchangeable or autoregressive of order one structure. The number of subjects affected by the missing data and the number of visits removed, vary in order to assess the impact of the missing data. Our simulations show that the estimators obtained by GEE are resistant to a certain rate of missing data. The results are homogeneous regardless to the imposed missing data structure.

References

Recovering missing individual block information in a multiblock Multiple Factor Analysis

Valentin Voillet\(^1\), Magali SanCristobal\(^1,2\), Laurence Liaubet\(^1\), Phillipe Besse\(^2\), Ignacio Gonzalez\(^3\)

1. INRA, UMR1388, Génétique, Physiologie et Systèmes d’Élevage (GenPhySE), F-31326 Castanet-Tolosan, France
2. Université de Toulouse INSA, UMR5219, Institut de Mathématiques, F-31077 Toulouse, France A
3. INRA, UMR875, Mathématiques et Informatiques Appliquées (MIAT), F-31326 Castanet-Tolosan, France

\(^{*}\)Contact author: valentin.voillet@toulouse.inra.fr

Keywords: Missing Individual, Multiblock, Imputation, Multiple Factor Analysis

Omics technologies enable the researcher to extract thousand of thousands of measures at the cellular scale: genomics, transcriptomics, proteomics, metabolomics, etc... Multiple tissues are also being analysed at the same time on a common set of individuals. As a matter of example, Omics studies have been performed to compare different conditions: different stages of fetal muscle development in the pig [7]. After an analysis of each omics table, the aim is to integrate the information on the multiple omics tables. In the presence of two tables, the partial least square (PLS) method [5] can be applied. With multiple tables, multivariate exploratory data analysis methods such as multiple factor analysis (MFA) have been used in the context of Omics data tables analysis [2]. MFA [3, 6] deals with datasets in which the same group of individuals is described by several tables of variables. These approaches are usable with no missing values. When few values are missing at random, a regularised iterative MFA approach [4] have been proposed. However in some datasets, some individuals may not be present in all sets of variables for lots of reasons (e.g. due to technical issues during the sampling). These individuals are thus often removed before applying MFA, resulting sometimes in the inability to perform this analysis. To avoid loosing useful information, we developed a method to handle these individuals.

First, let us briefly remind the basis of MFA. MFA is composed of two steps. Firstly, \( J \) separate principal components analyses (PCA) are executed on each set of variables \( K_j \), where \( K_j (j = 1, \ldots, J) \) is the \( j^{th} \) omic table. A global analysis is then performed weighting each set of variables \( K_j \) by the first eigenvalue \( 1/\lambda^1_j \).

In the merged dataset \( K = \{ K_1, \ldots, K_J \} \), it happens that one individual has missing values in an entire omics dataset \( K_j \). Our method randomly chooses and duplicates an individual in \( K_j \) to replace individuals not systematically present in all omic tables before performing MFA. The substitute has to belong to the same condition and the same omic table as the missing individual. The same approach is then performed several times and all MFAs are then summarized following the principle of the double PCA [1]. We consider \( T = \{ K^1, \ldots, K^M \} \) where \( M \) is the number of MFA and \( K^m \) is the \( m^{th} \) merged set \( K^m = \{ K^m_1, \ldots, K^m_J \} \). Each dataset \( K^m \) is weighted by the square root of the sum of the relevant eigenvalues \( q \) of its MFA (1). A final MFA is then performed on the center of gravity \( G \) of these weighted datasets.

\[
G = \frac{1}{M} \sum_{m=1}^{M} \frac{K^m}{\sqrt{\sum_{l=1}^{q} \lambda^m_l}}
\]  
(1)

Repeating the replacement of the missing individuals and aggregating the obtained datasets allows us to obtain a more robust estimation of the missing individuals.

We compared our method with the regularised iterative MFA approach [4] on a simulated dataset (3 merged omic data tables of 64 individuals and 20 variables) by measuring the correlation between the first components of the MFA with the complete dataset and of the MFA with the depleted then completed dataset. Our method was closer to the consensus configuration than the regularised iterative MFA even when a lot of individuals were missing, showing thus promising results (Figure 1).
QRILC: a quantile regression approach to left-censored missing data imputation in quantitative high-throughput proteomics

Cosmin Lazar\textsuperscript{1,3,4}, Myriam Ferro\textsuperscript{1,3,4}, Christophe Bruley\textsuperscript{1,3,4}, Thomas Burger\textsuperscript{1,2,3,4}\textsuperscript{*}

1. Univ. Grenoble Alpes, iRTSV-BGE, F-38000 Grenoble, France.
2. CNRS, iRTSV-BGE, F-38000 Grenoble, France.
3. CEA, iRTSV-BGE, F-38000 Grenoble, France.
4. INSERM, BGE, F-38000 Grenoble, France.
\textsuperscript{*}Contact author: thomas.burger@cea.fr

Keywords: quantitative proteomics; left-censored data; quantile regression.

Background: Missing data is a major concern for data scientists in many disciplines and in particular in mass spectrometry-based proteomics where high rates of non-responses are recorded during experiments. It is generally acknowledged that the major proportion of missing data is due to the detection limits of the spectrometers leading to left-censored measurements (see Fig. 1). Solutions to cope with this problem rely on the imputation of missing data and a number of strategies have been proposed for proteomics experiments. Despite the large amount of literature dedicated to the missing data imputation in statistics, the problem is still challenging in proteomics due to the specificity of the missingness mechanism.

Results: We present QRILC, a new approach to infer left-censored missing data from a truncated distribution with parameters estimated using quantile regression. We show that under the assumption of Gaussianity, the quantile regression provides with good estimates of the complete-data distribution parameters assuming that the missing data is left-censored; the new missing data imputation method was evaluated in a rigorous framework to test both for its goodness of fit performances as well as its influence on downstream analysis, and it was also compared with six other state-of-the-art methods for missing data imputation.

Conclusions: We performed simulated experiments both on artificial data as well as on three selected real datasets in a scenario where the percentage of missing data is controlled. The results obtained show that the proposed solution provides the best goodness of fit performances compared to the original data, and in the same time performs best in the context of differentially abundance analysis, among a number of six other different methods used for comparison.

R package: The QRILC imputation method is freely available in the \texttt{imputeLCMD} R package [1].

References


Poster submission.
Integrating Data Deficient species in evolutionary history loss analyses

S.Veron¹,⋆, P. Clergeau¹,⋆ and S. Pavoine¹,²,⋆

1.Centre d’Ecologie et des Sciences de la Conservation (CESCO UMR7204), Sorbonne Universities, MNHN, CNRS, UPMC, CP51, 55-61 rue Buffon, 75005, Paris, France
2. Mathematical Ecology Research Group, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

⋆Contact author: sveron@edu.mnhn.fr; pavoine@edu.mnhn.fr;clergeau@edu.mnhn.fr

Keywords: Conservation status, data deficient species, simple imputation methods, evolutionary

A major objective for biologists is to identify the most threatened species in order to define and prioritize conservation actions. In this aim, the Red List developed by the International Union for the Conservation of Nature (IUCN) enables to identify species at risk of extinction and has supported a large number of conservation programs (IUCN, 2015). Yet, among the species evaluated in the IUCN Red List, available information is not always sufficient to quantify threat status. Those species are classified as Data-Deficient (DD). Assessing the conservation status of DD species is a great challenge in conservation biology (Bland et al., 2014). Indeed, DD species brings high uncertainties in the proportion of threatened species in a group, and, as it was argued for amphibians, DD species could be more threatened than already assessed species (Howard & Bickford, 2014). DD species may then strongly impact the prioritization of conservation actions.

DD species are also a crucial issue to assess the loss of Phylogenetic Diversity (PD) and Evolutionary Distinctiveness (ED) (Isaac et al., 2007; Collen et al., 2011). PD and ED are highly valuable components of biodiversity which depict the evolutionary history of conservation interest (May, 1990; Faith, 1992). Several studies found that DD species greatly influenced the amount of evolutionary history predicted to be lost following species extinctions (Collen et al., 2011; Isaac et al., 2012, Jono & Pavoine, 2012). For example, if some clustered DD species are in high danger of extinction it may lead to the loss of deep branches and dramatically increase the loss of evolutionary history. So far, in a phylogenetic framework, none study tried to minimize the bias due to DD species which are generally excluded from the analysis.

Our first objective was to explore the bias brought by DD species on 4 widely used indices which measure ED and PD at risk: HEDGE, EDGE, PDloss and ExpectedPD loss indices (Faith, 1992; Isaac et al., 2007; Steel et al., 2007; Faith, 2008). Second we looked for an easy-to-use and multi-taxa imputation method that minimized the bias for those 4 indices. To do so we simulated DD species in three vertebrate groups: carnivores, amphibians and reptiles. DD species were simulated according to three parameters: the proportion of DD species, the time extinction scenario and the degree of clustering of DD species in the most evolutionary distinct and in the most range restricted species.

We then imputed the simulated DD species with four methods: first, DD species were all considered to be in critical danger of extinction (CR status), second they were all considered to be non-threatened (LC status), third DD species were assigned a threat status with a probability calculated from the proportion of each threat status in already assessed species, finally we imputed threat status according to the size of the geographic range of DD species. We then measured the difference between the real values of PD and ED indices with the values obtained from simulation and imputation of DD species. We found that DD species brought important bias in the assessment of all four indices in the three groups. The method which performed the best to minimize those bias independently of the parameters considered was the one based on geographic range size. CR and LC status imputation method enabled to assess the worst case and best case scenario of extinctions and PD and ED loss, respectively, but they brought large gaps with the real values of PD and ED loss. As for the probabilistic method, it performed second to minimize those biases in PD and ED loss. We thus showed that in a phylogenetic framework, an easy to use method based on geographic...
Analyzing multi-group multivariate data with non-ignorable missingness

Myrsini Katsikatsou1,⋆, Jouni Kuha1,2, Irini Moustaki1
1. Department of Statistics, London School of Economics, UK
2. Department of Methodology, London School of Economics, UK
*Contact author: m.katsikatsou@lse.ac.uk

Key words: item non-response, latent trait model, latent class model

The standard multigroup latent variable modeling framework is extended to the modeling of missing data mechanism.

Let \( Y_C = (Y_1, \ldots, Y_p)\top \) denote a vector of \( p \) observable variables for a single observational unit which measures a latent variable \( \eta \). Let \( Y_O \) and \( Y_{mis} \) denote the items in \( Y_C \) which are observed and missing, respectively. A non-empty \( Y_O \) provides information about \( \eta \). The information about the pattern of missing data for a respondent is incorporated in a vector \( R = (R_1, \ldots, R_p)\top \), where \( R_i = 1 \) if \( Y_i \) is observed, and 0 otherwise. The proposed modeling framework may be viewed in the general framework of maximum likelihood models for handling missing data. In particular:

\[
p(R, Y_C | X) = \int p(R|Y_C, \eta, X) p(Y_C|\eta, X) p(\eta|X) d\eta,
\]

where \( X \) is a vector of covariates, including the group variable, \( p(R|Y_C, \eta, X) \) is the non-response model, \( p(Y_C|\eta, X) \) is the measurement model, and \( p(\eta|X) \) is the structural model. Distinct parameters are assumed for the non-response model and the rest of the model (measurement and structural models).

We make the assumption that \( p(R|Y_C, \eta, X) = p(R|\eta, X) \), since \( \eta \) is measured by \( Y_C \). Adjusting the definitions provided in Rubin (1976) and Little and Rubin (2002) to our framework the data are considered missing at random (MAR) if \( p(R|\eta, X) = p(R|X) \). Note that conditioning on all \( X \) that affect both \( \eta \) and \( R \) is essential to avoid a confounding correlation between \( \eta \) and \( R \). In the case of non-ignorable missingness, a model is needed for \( R \). Unlike general non-ignorable models, \( p(R|\eta, X) \) is identifiable without further assumptions, because the multiple indicators in \( Y_C \) identify \( \eta \) and the non-response is ignorable given \( \eta \).

We propose a latent class model for \( R \); the associations among the \( R_j \)'s are explained by a discrete latent variable denoted by \( \xi \) with \( C \) latent classes, where \( C << 2^p \). The advantages of using a latent class model are firstly, that it will approximate the observed multinomial distribution of \( R \) as the number of latent classes increases. Thus, it provides a flexible way of modelling the associations among the \( R_j \)'s and avoids the risk of a confounding correlation between \( \eta \) and \( R \). Secondly, we can achieve a well specified and parsimonious model for the missingness mechanism. The proposed model is as follows:

\[
p(R | \eta, X) = \Pr (R = r | \eta, X) = \sum_{c=1}^{C} \Pr (R = r | X, \eta, \xi = c) P(\xi = c | X, \eta),
\]

where \( \Pr (R = r | X, \eta, \xi = c) \) is allowed to depend on \( X \) with the aim of specifying a well-fitting model for the missingness indicators within each group indicated by \( X \). The parameter that represents the association between \( \xi \) and \( \eta \) summarizes the type of missingness.

The proposed modeling framework is demonstrated by analyzing data from the European Social Survey (ESS) Round 4 (2008) that was conducted in 29 countries.

For a poster submission
Multiple Imputation

- *Analysis of incomplete data and imputation of missing values: the R package VIM*, Matthias Templ, Alexander Kowark [Talk]

- *Multiple imputation using multiple correspondence analysis*, Vincent Audigier, François Husson, Julie Josse [Talk]

- *PLS regression and Multiple Imputation*, Phillipe Bastien [Talk]

- *Using multiple imputation for a zero-inflated contingent valuation with potentially biased sampling*, Marouenne Mbarek, Damien Rousseliere, Julien Salanie [Talk and poster]

- *Multiple Imputation using Mahalanobis Distance*, Elaine M. Berkery, Kevin Hayes [Poster]


- *Proposition of a Multiple Imputation approach for MNAR mechanism using Heckman’s model*, Jacques-Emmanuel Galimard, Sylvie Chevret, Matthieu Resche-Rigon [Poster]
Analysis of incomplete data and imputation of missing values: the R package VIM

Matthias Templ\textsuperscript{1,2,*}, Alexander Kowarik\textsuperscript{2}

1. Vienna University of Technology
2. Statistics Austria
*Contact author: matthias.templ@tuwien.ac.at

Keywords: Missing Values, Visualization, Imputation, Robust Methods

The package VIM \cite{Templ:2013, Templ:2012} is developed to explore and analyze the structure of missing values in data using visualization methods, to impute these missing values with the built-in imputation methods and to verify the imputation process using visualization tools, as well as to produce high-quality graphics for publications. Univariate, bivariate, multiple and multivariate plot methods are implemented to analyse the structure of missing values, but they can also be used to evaluate imputed data. The most common imputation methods such as hotdeck, kNN and regression imputation are implemented in an efficient manner to also work for large data sets. Moreover, also more advanced methods based on robust statistics are implemented \cite{Templ:2011}. Most of the implemented method can deal with a mixture of continuous, semi-continuous, count, categorical and binary variables and allow also for multiple imputation.

A point- and click graphical user interface has been developed to give access to these methods and tools to users with limited R skills. It is available in the package VIMGUI \cite{Schopfhauser:2013} that is also prepared to work for objects produced by R’s survey package. All important methods are supported by the flexible point- and click-interface. The presentation describes the application of the methods available in the package VIM.

References


Multiple imputation for categorical variables with multiple correspondence analysis

Vincent Audigier$^1$,⋆, Francois Husson$^1$, Julie Josse$^1$

1. Agrocampus Ouest

⋆ Contact author: audigier@agrocampus-ouest.fr

Keywords: missing values, categorical data, multiple imputation, multiple correspondence analysis, bootstrap

Missing values are a key problem in statistical practice. Indeed most statistical methods cannot be applied directly on an incomplete data set. To handle this issue, one of the common approaches is to perform single imputation. This consists in imputing missing values by plausible values. It leads to a complete data set that can be analyzed by any standard statistical method. However, one drawback of single imputation is that it does not take into account the uncertainty associated with the prediction of missing values based on observed values. Thus, if we apply a statistical method on the completed data table, the variability of the estimators will be underestimated. To avoid this problem, a solution is to use multiple imputation [5, 4] which consists in predicting different plausible values for each missing entry leading to multiple imputed data set. The variability amongst the imputed values reflects the variability of prediction.

This presentation proposes a new method of multiple imputation dedicated for categorical variables based on multiple correspondence analysis (MCA) [2, 3]. This method imputes the missing entries using the principal components method dedicated to categorical data: multiple correspondence analysis. MCA is a principal components method to describe, summarise and visualise multidimensional matrices with categorical data. This powerful method allows us to understand the two-way associations between variables as well as the similarities between individuals. Imputation using MCA benefits of the relationships between variables and similarities between individuals. To reflect the uncertainty of the MCA parameters (i.e. the principals components and the loading) from one imputation to the next, we use a non-parametric bootstrap approach. Multiple imputation using MCA (MIMCA) requires estimating a small number of parameters due to the dimensionality reduction property of MCA. It allows the user to impute a large range of data sets. In particular, a high number of categories per variable, a high number of variables or a small the number of individuals are not an issue for MIMCA. Through a simulation study based on real data sets, the method is assessed and compared to the reference methods (multiple imputation using the loglinear model [6], multiple imputation by logistic regressions [9]) as well to the latest works on the topic (multiple imputation by random forests [1, 7] or by the Dirichlet process mixture of products of multinomial distributions model [8]). The proposed method shows good performances in terms of bias and coverage for an analysis model such as a main effects logistic regression model. In addition, MIMCA has the great advantage that it is substantially less time consuming on data sets of high dimensions than the other multiple imputation methods.

References


For a lightning talk submission
PLS Regression and Multiple Imputation

Philippe Bastien

1. L’Oreal Research & Innovation, 1 avenue Eugène Schueller, 93601 Aulnay-sous-Bois

*Contact author: pbastien@rd.oreal.com

Keywords: PLS regression, Nipals, EM, Multiple Imputation

There are various ways to deal with missing values. Among conventional methods for handling missing data "listewise deletion" could be considered as a safer or less biased approach. However if the amount of data that must be discarded under listwise deletion is dramatic other alternatives must be considered. A first strategy is the simple imputation which will substitute a value for each missing value. For example each missing value can be imputed by the average of the complete cases, the nearest neighbour [1], or by the predicted value of a regression model. These ad hoc methods are not without disadvantages, drawing in the first both cases variance-covariances towards zero and in the second case by increasing the observed correlations. Other strategy consist in using the EM algorithm [2] for the estimation of maximum likelihood with incomplete data or the direct Maximum likelihood estimation used in program for structural equation modelling [3]. Although ML represents a major advance over conventional approaches to missing data, it has its limitation. In particular the estimation of the Cox proportional hazard model or the ordered logistic regression model will not be obvious.

The PLS Regression in its standard form with the use of the NIPALS algorithm [4, 5] can deal with missing values. Treatment of missing values with PLS-NIPALS can be implicitly associated as a simple imputation method. The coefficients $w_{h,j}$ of PLS components $t_h$ are calculated as the slope of the least squares straight line passing through the origin of the set $(y_i, x_{h,i,j})$, with $x_{h,j}$ the residual of the regression of $x_j$ on $t_1, \ldots, t_{h-1}$. That implicitly amounts replacing the missing points by points located on the slope.

Simple imputation which treats the missing values in a deterministic way does not reflect uncertainty associated with their prediction. If the data are missing at random, and if the parameters of the model do not depend on the process generating the missing values, Rubin in the 1970's [6] proposed to replace each missing value with a set a plausible values that represent the uncertainty about the right value to impute. This process leads to valid statistical inferences that properly reflect the uncertainty due to missing values.

Multiple imputation inference involves three distinct phases:
- The missing data are filled in $m$ times to generate $m$ complete data sets.
- The $m$ complete data sets are analyses by using standard procedures.

The results from the $m$ complete data sets are combined for the inference. As for Bayesian inference, imputations are generated by simulation of the joint posterior distribution of $(\theta, y_{mis}/y_{obs})$. These quantities being unknown, their distribution can be simulated iteratively [7] by using Data Augmentation, a Monte Carlo Markov Chain method which is close in its formulation to the EM algorithm. The posterior distribution is simulated by alternatively drawing the missing values and the parameters.

We will present an application of Multiple Imputation in the framework of PLS logistic regression [8, 9] to predict quality of Bordeaux wine. Results from PLS-NIPALS and PLS MI will be compared.

For a lightning talk submission
Using multiple imputation for a zero-inflated contingent valuation with potentially biased sampling

Marouene Mbarek¹,², Damien Rousselière¹,²,⋆, Julien Salanié³,⁴

¹. AGROCAMPUS OUEST, Department of Economics, Management and Society, Angers, France
². GRANEM; AGROCAMPUS OUEST, Université d’Angers, Angers, France
³. Université Jean Monnet, Saint-Etienne, France
⁴. GATE Lyon-Saint-Etienne; Université de Lyon, France

⋆ Contact author: damien.rousseliere@agrocampus-ouest.fr

Keywords: Contingent valuation, Multiple Imputation, Sample and Non-response biases

There is a growing interest from social science researchers in multiple imputation (MI). Beyond the classical application to missing data, MI can also deal with measurement errors [2], comparability of international surveys [16], selective drop-out [4] or causal inference using the potential outcomes framework [15]. Building on an argument developed by S. Van Buuren [16], we generate synthetic records by MI in order to address a potentially biased sampling.

A biased-sample is a subject of matter in all observational studies [3]. This problem is specially important in contingent valuation analysis [17] as it distorts aggregate economic estimates. Therefore it has strong implications for public policies. Based on the access to auxiliary data, this approach assumes that the bias depends on known covariates which are fully observable. Therefore it meets the MAR (Missing At Random) assumption. We also assume that the distributions of some covariates are known. These assumptions are not different from those which are usually necessary for the classical method (post-stratification or weighting to known population totals)[7]. But as proposed by S. Van Buuren, “Imputation provides fine-grained control over the correction process”. MI can help us therefore to introduce sensitivity analysis [11]. This thought experiment approach can also be of interest to any pragmatist, institutionalist or reflexive economic methodology [5].

The sample is augmented by new artificial records using various hypotheses and priors. In spite of its known limits, our MI benchmark algorithm will be AMELIA [10] because informative Bayesian priors about individual missing data cells can be included. AMELIA has also an advantage in its flexibility and its efficiency. The incorporation of priors follows basic Bayesian analysis where the imputation turns out to be a weighted average of the model-based imputation and the prior mean, with weights depending of the relative strength of the data and prior [9]. These informative prior can come from the elicitation of expert belief or from the analysis of previous studies [6]. There are helpful for handling sparse data [12], which can appear if the sample is strongly biased, as more additional artificial records are needed.

Our case study is a contingent valuation of the WTP (Willingness to Pay) for accessing a MPA (Marine Protected Area) in a developing country (Tunisia). To our knowledge, previous studies did not correct for biased sampling [14]. In July and August 2012, 315 tourists were surveyed randomly during their visit to Kuriat islands. Thanks to an access to official statistics on the tourism in Tunisia, we generate 185 additional artificial records (with 100 replications) in order to correct for misrepresentation of gender and nationality. Additional variables were included in order to keep the imputation model and the substantive (or econometric) model congenial [13].

Our econometric models are the ZiOP (Zero-inflated ordered probit) which is a double-hurdle combination of a split probit model and an ordered probit model [8] and its extension the ZiOPC which assumes that the two errors terms are correlated [1]. These models address the problem of two distinct data generating processes for the zeros. One type corresponds to individuals who will always refuse to pay for MPA (because of ideological reasons), whereas the other type refers to a corner solution. Our robustness checks show that age, income and environment perceptions are the main determinants of WTP.
Multiple Imputation using Mahalanobis Distance

Elaine M. Berkery∗1, Kevin Hayes1

1. Deptartment of Mathematics and Statistics, University of Limerick *Contact author: elaine.m.berkery@ul.ie

Keywords: Multiple imputation, Mahalanobis distance, hot deck imputation.

The occurrence of missing data is ubiquitous in applied statistics and can occur in all types of data sets and for a variety of different reasons. A common way of dealing with the missing data is multiple imputation where several fully observed data sets are created, analysed and the results combined to obtain the best estimates for the parameters that are of interest. One such method is hot deck imputation whereby missing values are replaced by observed values in respondents (donors) that are similar to the non-responding case (recipients). Matching donors to recipients can be done in a number of different ways, one of which is to use the Mahalanobis distance, 

\[ D_{ij}^2 = (x_i - \bar{x}_j)'S^{-1}(x_i - \bar{x}_j) \]

Subsequently, donor pools are obtained for each recipient depending on the Mahalanobis distance calculated and multiple fully observed data sets created. This work investigates differences that occur in imputations as a result of variations in the calculation of the variance covariance matrix, how donor pools are created and the selection of imputed values from the donor pools for both continuous and categorical variables. Among the ways I propose to select from the donor pools is to use a probabilistic proportional sampling method.

One software package that specialises in missing data imputation, and uses the Mahalanobis distance as a matching method, is called SOLAS. Consider multivariate data in the rectangular layout

\[
\begin{bmatrix}
\begin{array}{ccc}
\bar{z}_{11} & \cdots & \bar{z}_{1q} \\
\vdots & \ddots & \vdots \\
\bar{z}_{n1} & \cdots & \bar{z}_{np} \\
\bar{x}_{11} & \cdots & \bar{x}_{1p} \\
\vdots & \ddots & \vdots \\
\bar{x}_{n1} & \cdots & \bar{x}_{np} \\
\end{array}
\end{bmatrix}
\begin{bmatrix}
\begin{array}{cc}
\bar{z}_{11} & \cdots & \bar{z}_{1q} \\
\vdots & \ddots & \vdots \\
\bar{z}_{m1} & \cdots & \bar{z}_{mq} \\
\vdots & \ddots & \vdots \\
\bar{x}_{11} & \cdots & \bar{x}_{1p} \\
\vdots & \ddots & \vdots \\
\bar{x}_{m1} & \cdots & \bar{x}_{mp} \\
\end{array}
\end{bmatrix}
= \begin{bmatrix}
\bar{Z}_{obs} & X \\
\bar{Z}_{mis} & \tilde{X}
\end{bmatrix}
\]

where the superscripts “obs” and “mis” on the matrices \(Z_{obs}\) and \(Z_{mis}\) indicate observed values and missing values. When investigating the process carried out by SOLAS, and outlined further in Grannell [3], it is not obvious whether the sample variance-covariance matrix \(S\) used in \(D_{ij}^2\) is based on: the \(n \times p\) data matrix \(X\); or the \((n - 1) \times p\) data matrix \(X_{-i}\), that is, \(X\) with row \(i\) deleted; or the augmented \((n + m - 2) \times p\) matrix \(\begin{bmatrix}
\begin{array}{c}
X_{-i} \\
\tilde{X}_{-j}
\end{array}
\end{bmatrix}\), where \(\tilde{X}_{-j}\) is matrix \(\tilde{X}\) with row \(j\) deleted. The various combinations that \(S\) may be calculated using, as outlined above, may have an effect on the robustness and bias of the imputations calculated. The proposed methods for selecting from donor pools and variations in the calculations of \(S\) are then applied to both existing and simulated data and the results analysed.

References


For a poster submission
Multiple Imputation of Missing Categorical Data using Latent Class Models: State of Art

Davide Vidotto¹* , Jeroen K. Vermunt¹

1. Department of Methodology and Statistics, Tilburg University
   *Contact author: d.vidotto@tilburguniversity.edu

Keywords: latent class models, missing data, mixture models, multiple imputation.

This work offers a review of the use of latent class models for multiple imputation of missing categorical data in large-scale analyses. Multiple imputation requires models that are able to describe all the possible associations among the items of a dataset to accurately draw imputing values from the posterior distribution of the missing data. Latent class (or finite mixture) models, besides being a well-known clustering tool, are good density-estimators of nominal variables. Unlike several other imputation models for categorical data that have been proposed in the literature, latent class models can pick up high-order relations in high-dimensional datasets, thus supplying a powerful imputation engine for large-scale studies. In this work an overview of both frequentist and Bayesian implementations, along with analogies and differences of such methods, is provided.

References


For a poster submission
Proposition of a multiple imputation approach for MNAR mechanism using Heckman’s model

Jacques-Emmanuel Galimard¹,²,⋆, Sylvie Chevret¹,²,³, Matthieu Resche-Rigon¹,²,³

1. INSERM U1153 ECSTRA team
2. Paris Diderot University
3. AP-HP, Saint-Louis Hospital

⋆ Contact author: jacques-emmanuel.galimard@inserm.fr

Keywords: Missing Not At Random (MNAR), Multiple Imputation: chained equation, Heckman model

Standard implementations of multiple imputation (MI) approaches provide unbiased inferences assuming underlying MAR mechanisms. However, in presence of missing data generated though MNAR mechanisms, although the MAR assumptions can be approached by collecting more explanatory variables, MI is not satisfactory and difficult to validate. Coming from econometric statistics, the Heckman’s method, also called the sample selection method, deals with selected sample using two joined linear equations, namely the selection equation and the outcome equation, respectively [1, 2]. The Heckman’s method has been successfully applied to missing outcomes in presence of MNAR mechanism. Nevertheless, such a method deals with missing outcomes only, and this is a strong limitation in clinical epidemiology settings where covariates are also often missing.

We propose to extend the validity of MI to MNAR mechanisms by using the Heckman’s model as the imputation model, using a two-step estimation process. This will provide a solution that can be used in a MI by chained equation framework to impute missing variables (outcomes or covariates) resulting either from a MNAR or a MAR mechanism.

This approach will be validated by a simulation study. It will be evaluated in presence of missing outcome and missing covariates either under MNAR or MAR mechanisms.

References


Handling Missing Data in Matched Case-Control Studies using Multiple Imputation

Shaun Seaman
Medical Research Council Biostatistics Unit, UK

The analysis of matched case-control studies is often complicated by missing data on covariates. One solution is to restrict the analysis to individuals with complete data, but this is inefficient and may be biased. We instead propose two multiple imputation methods. The first uses a model for the data on an individual and includes matching variables; the second uses a model for the data on a whole matched set and avoids the need to model the matching variables. Both methods allow for multiple partially-observed covariates, non-monotone missingness patterns and multiple controls per case. They can be easily applied in standard statistical software and valid variance estimates obtained using Rubin’s Rules. We demonstrate, in a simulation study, approximate unbiasedness of log-odds ratio estimators and good coverage of confidence intervals, and compare the efficiency of the two methods. Data on colorectal cancer and fibre intake from the EPIC-Norfolk study are used to illustrate the methods, in particular showing how efficiency is gained relative to restricting the analysis to individuals with complete data, and the choice of method is discussed.
Gaussian-based routines for imputing categorical variables in complex data

Murat Recai Yucel
University of Albany, United States

This presentation focuses on computationally feasible ideas on developing imputation models. Our main motivation is to employ these ideas in data structures with unique complexities such as large number of incompletely-observed categorical variables, multilevel data and/or clustered data with multiple membership. Our computational routines modify computationally advantageous Gaussian-based methods that use Markov Chain Monte Carlo techniques to sample from posterior predictive distribution of missing data. Specifically, we propose rounding rules to be used with these existing imputation methods, allowing practitioners to obtain usable imputation with small biases. These rules are calibrated in the sense that values re-imputed for observed data have distributions similar to those of the observed data. The methodology is demonstrated using a sample data from the NewYork Cancer Registry database.
Multiple Imputation for Complex Surveys: An Overview of the State of the Art

Joseph L. Schafer
Census Bureau, Washington, United States

Multiple imputation (MI) was proposed by Donald Rubin in 1977 as a general paradigm for handling nonresponse in surveys, censuses and administrative databases. Over the last decade, it has gained widespread acceptance in many fields of quantitative research, and nearly every major statistical software package now performs MI in some fashion. In some cases, MI has been applied to complex surveys, but in other cases survey practitioners still find MI too daunting to implement. In this presentation, I review the history of MI, describe the statistical and nonstatistical challenges it poses in the survey world, and look ahead to future developments in this area. Statistical issues include: mismatch between the imputation model and subsequent analyses; the challenges of jointly imputing large numbers of survey variables of mixed types; the need to preserve complex interactions; the need to reflect important features of the sample design; hierarchical or multilevel structures with missing values at multiple levels; preservation of skip patterns and logical constraints. Nonstatistical issues include: the availability and utility of software; perceived and actual difficulty of carrying out MI in a production environment; perceived and actual difficulty of explaining MI to data users and clients; and organizational culture and priorities.
Triply robust inference in the presence of missing survey data

David Haziza\textsuperscript{1}, Valéry Dongmo Jiongo\textsuperscript{2}, Pierre Duchesne\textsuperscript{1}

\textsuperscript{1} Département de mathématiques et de statistique, Université de Montréal, Montréal, Québec, Canada
\textsuperscript{2} Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada

Item nonresponse is typically treated by some form of single imputation in statistical agencies. For example, deterministic regression imputation that includes ratio and mean imputation within classes as special cases, is widely used in surveys. Recently, there has been an interest in doubly robust imputation procedures. An imputation procedure is said to be doubly robust when the resulting imputed estimator is consistent if either the imputation model (also called the outcome regression model) or the nonresponse model is correctly specified. However, in the presence of influential units, the resulting imputed estimator may be potentially very unstable. We propose a robust version of the doubly robust imputed estimator based on the concept of conditional bias of a unit. Implementation of the proposed method via a calibrated imputation procedure will be discussed. Finally, the results from an empirical study will be shown.
Visual inpainting, the task of completing an image region, is an important component of image editing and enhancement, from home photo touch-up to professional movie post-production. After 15+ years of research (which we will briefly summarize), the problem remains challenging, especially for videos. As an attempt to progress on that front, we propose a novel automatic video inpainting algorithm which relies on the optimization of a global, patch-based functional. Our algorithm is able to deal with a variety of challenging situations which naturally arise in video inpainting, such as the correct reconstruction of dynamic textures, multiple moving objects, and moving background. Furthermore, we achieve this in an order of magnitude less execution time with respect to the state-of-the-art. We are also able to achieve good quality results on high-definition videos. The resulting algorithm requires no segmentation or manual input other than the definition of the inpainting mask and can deal with a wider variety of situations than is handled by previous work.
Missing data in astrophysics

Sandrine Pirez
CEA Saclay

Astrophysics deals with the physics of the universe and covers many different areas including cosmology, the study of astrophysical objects such as stars, galaxies, cosmic microwave background, ... The observation of these astrophysical objects from ground and space produces every day a more and more important quantity of data. To capitalize on these increasingly large data sets and study the Universe with ever-greater precision, we have to be able to correct the data from all the observational bias.

A problem that is frequent during astrophysical observations is the occurrence of missing data. Although missing data are very common in astrophysics, they can severely bias the data analysis. Various methods are currently used to correct from missing data. They depend both on the analysis to be performed and on the characteristics of the signal.

Recently, we have proposed to replace these methods by a more general method: a sparse inpainting method that I will describe. The introduction of the sparse inpainting to correct from missing data in astrophysics already had some success on several areas. A comparison with the methods currently used will be presented, for these applications.