Summary of research contributions 2023 - Julie Josse

I conduct research in statistical methodology and computational statics. My research practice consists of a balance between theory and applications, with multidisciplinary projects at the heart of my work. To conduct innovative research, I strongly believe in the contributions of different cultures both within and across disciplines. Hence, I nurture many national and international collaborations. I detail below my main research topics: 1) The challenge of handling missing data which is ubiquitous in statistical methods, machine learning and application domains; 2) Causality with multi-sources data, which offers new opportunities to better understand many processes but come with many open problems requiring statistical and computational breakthroughs and 3) low-rank matrix approximation and multi-view data analysis.

1 Missing data

The problem of missing values exists since the earliest attempts of exploiting data as a source of knowledge as it lies intrinsically in the process of collecting, recording, and preparing the data itself. It is all the more unavoidable as vast amounts of data are currently collected from different sources[1]. The literature on the subject is therefore abundant [14] and in a recent survey, we reviewed more than 150 implementations available to handle missing data [17].

Most methods have been developed for inferential purposes *i.e.,* to estimate parameters and their variance in the presence of missing values. When the inference is carried out by maximum likelihood, one can resort to **Expectation-Maximization (EM)** algorithms to get point estimates. However, the expectation steps are often intractable and few implementations are available. For example, no such solution existed for logistic regression with missing data, before we derived a stochastic EM approximation (SAEM) algorithm [9], implemented in an R package. Variables selection with missing values is also under-explored, so we proposed Adaptive Bayesian SLOPE (**jointly w. G. Bogdan**) which controls the false discovery rate.

The major drawback of these methods is that a new algorithm has to be rederived for each statistical method. **(Single) imputation** techniques which consist of replacing missing values with plausible values, are very appealing as they allow to both get a guess for the missing entries as well as to perform downstream statistical methods on the completed data. Iterative singular value decomposition methods that I developed (linked to **matrix completion**) were proven excellent to handle high dimensional heterogeneous data (categorical, quantitative, etc.) [1] and data with a multi-level structure [8] (**jointly w. B. Narasimhan**). In addition, there is a duality between estimation and imputation and these techniques also allow for methods such as Principal Component Analyses (PCA) in the presence of missing data (implemented in [12]). To tackle shortcomings of existing techniques, I have also suggested non-parametric imputation methods, either based on optimal transport (**jointly w. M. Cuturi**) to preserve the data distribution [22], or on statistical depth [21] to ensure robustness to both outliers and heavy-tailed distribution (**post-doctorate supervision: P. Mozharovsky**).

Multiple imputation reflects the uncertainty associated with the prediction of missing entries and thus leads to confidence intervals with good coverage properties incorporating the additional variability due to missing values. To tackle high dimensional settings, and categorical data with rare categories, I have developed such methods based on PCA [2].

Theory and practice mostly rely on the **Missing At Random** (MAR) assumption: the probability of being missing only depends on observed values. Solutions are limited (to a single variable with missing data or linear models) for **Missing Not At Random** settings which are however omnipresent in practice. After showing identifiability, we have proposed solutions for estimating parameters of low-rank models [23, 24] (**jointly w. C. Boyer**) either by modeling the missing-values mechanism or using *missingness graph*. We are currently investigating extensions to mixture models. Open challenges include handling different types of missing values mechanisms within a dataset or distributional shifts in the missing values mechanism.

Finally, with **G. Varoquaux** and **E. Scornet**, we have obtained the **first theoretical results on supervised learning with missing data**. Our motivation was to predict an outcome given incomplete covariates. We have shown Bayes consistency of *impute and regress* strategies [19] for all missing values mechanisms. It implies that the widespread mean imputation is consistent for prediction [13] despite its drawbacks for estimation (distribution distortion). I expect these

[1] *"One of the ironies of Big Data is that missing data plays an increasingly important role"*. Zhu, et al. 2019. High-dimensional PCA with heterogeneous missingness.

contributions to have a strong impact in many fields because the proposed solutions are easy to implement for the end users. However, the results are asymptotic and there is a need for refined finite sample guidance. We have also studied solutions for random forests and developed the first theoretically justified **neural networks architecture** [20, 16] with missing entries. These works have repercussions in the field of causal inference as it allows to estimate a treatment effect, for example with double robust methods and missing data in the confounders [18] (**jointly w. S. Wager**). To know what confidence could be given to the predictions obtained from initially incomplete data, we have suggested with **Y. Romano**, **conformal prediction** [26] with missing data and have obtained coverage guarantees conditional on the missing data pattern.

2 Causal inference for multi-sources data

Launching a drug without randomized control trials? Modern *evidence-based* medicine puts Randomized Controlled Trial (RCT) at the core of clinical evidence. In practice, almost all new drugs are authorized through such trials (after a pre-clinical study). Indeed, *randomization* enables to estimate the Average Treatment Effect by avoiding confounding effects of spurious or undesirable associated factors. In other words, RCTs are the current gold-standard to empirically measure a causal effect of a given intervention on an outcome. But more recently, concerns have been raised on the limited scope of RCTs: stringent eligibility criteria, unrealistic real-world compliance, short timeframe, limited sample size, etc. Such limitations threaten the external validity of RCT studies to other situations or populations. The usage of complementary non-randomized data, referred to as *observational* or from *real world*, brings promises as additional sources of evidence, in particular combined to trials. **Transportability** (also known as **generalization**, *recoverability from sampling bias*, or *data-fusion*) allows to generalize or transport the trial findings toward a target population of interest, potentially subject to a covariate **distributional shift**, see a review in [7].

RCT and observational data are seldom acquired as part of a homogeneous effort. As a result, they come with different covariates. Restricting the analysis to the shared covariates raises the risk of omitting an important one leading to identifiabilities issues. This problem is reminiscent of unobserved confounding in causal inference with one observational data. In [4], we suggest a **sensitivity analysis** to handle cases where such covariates (namely treatment effect modifiers that are shifted between the two sets when studying risk difference) are missing in one or both sets. We also completed proofs on the consistency of generalization estimators that use either **weighting (Inverse Propensity of Sampling Weighting, IPSW)**, **outcome modeling** , or combine the two in **doubly robust approaches with Augmented IPSW (AIPSW)**.

We further analysed the IPSW estimator, which consists of re-weighting the trial so that it resembles the observational sample, in [5]. In particular, we established **finite sample bias and variance** (the literature mostly focuses on asymptotic results) and upper bound on the risk of different versions of the estimator: oracle, semi-oracle, etc. This work can lead to practical recommendations in terms of data collection (*e.g.,* doubling the size of the observational data leads to a smaller asymptotic variance than doubling the size of the trial). In addition, we studied the impact of the choice of variables: how including covariates that are not necessary for identifiability of the causal effect may impact the asymptotic variance.

There exists no estimators to generalize other measures than the risk difference such as the **risk ratio, survival ratio**, number needed to treat, odds ratio etc, no theoritical guaranteed on their behavior and no assessment of their empirical behavior. Such causal measures are however dominant in the medical literature and more appropriate for binary outcome for instance. In [6], we highlight that some measure are easier than other to generalize as they require few variables (for instance only treatment effect modifiers and not both modifiers and prognostic variables). This work focuses on identifiability conditions and do not provide neither study any estimators to generalize a risk ratio from an RCT to a target population.

Same questions arise when estimating optimal Individual Treatment Regime (ITR): the covariate shift between the source and target population, renders the source-optimal ITR not necessarily optimal for the target population. In [27], with **S. Yang**, we suggest an efficient and robust transfer learning framework for estimating optimal ITR with right-censored survival.

I have many ongoing works on the topic. With **A. Chambaz** we design **policy learning** methods adapting super-learner methods with missing values. I have also worked on variable importance for causal forests [3], etc. To push methodological innovation up to the stakeholders and to lead to actionable solutions, I continue my efforts toward open source implementation. I have recently federated colleagues to create a taskview on causal inference, to efficiently organise codes for users.

3 Low rank matrix estimation

Consider a model where a data matrix X with n rows and p columns is generated from some distribution $\mathcal{L}(\mu)$ with $\mathbb{E}_\mu[X] = \mu$: $X \in \mathcal{R}^{n\times p} \sim \mathcal{L}(\mu)$ with μ of low rank k. The statistical aim is to recover the signal $\mu \in \mathbb{R}^{n\times p}$ from the noisy data. The low-rank assumption arises naturally in several different settings and is a powerful way to address the problem of recommender systems. A classical approach is to estimate the signal μ as the best rank-k approximation to X, for an adaptively chosen k:

$$\hat{\mu} \in \underset{\mu\in\mathbb{R}^{n\times p}, \text{rank}(\mu)\leq k}{\arg\min} \|X-\mu\|_2^2 \ .$$

The solution is the truncated singular value decomposition (SVD) of the matrix $X = UDV^\top$ at the order k, namely $\hat{\mu}_k = \sum_{l=1}^k d_l\, u_l\, v_l^\top$, where d_l are the singular values organized in decreasing order.

Shrinking and thresholding the singular values Procedures beyond the truncated SVD have been proposed, including the nuclear norm regularized method which involves soft-thresholding singular values. However, with **S. Sardy**, **F. Husson**, and **S. Wager**, we [15, 11] demonstrate that while the soft-thresholding estimator has a small MSE for recovering μ in low SNR scenarios, it struggles in other cases. For greater flexibility, we proposed the **adaptive trace norm estimator** (ATN) in [15], which uses two regularization parameters (λ, γ) to threshold and shrink the singular values as: $\sum_{l=1}^{\min\{n,p\}} d_l \max\left(1-\frac{\lambda^\gamma}{d_l^\gamma}, 0\right) u_l\, v_l^\top$.

This estimator denoted $\hat{\mu}_{(\lambda,\gamma)}$ is a versatile estimator that includes both soft and hard thresholding as special cases. Notably, when $\gamma > 1$, it shrinks the smallest singular values more than the largest, which is beneficial since the smallest singular values are responsible for instability. The parameters can be chosen via cross-validation, but to address its computational cost, we proposed a **Stein unbiased estimator of the risk** in [15] to even handle missing values. Since this requires knowledge of the noise scale, we also derived a generalized SURE [15], inspired by generalized cross-validation.

In a specific asymptotic framework, considering n and p as fixed, but letting the noise variance tend to zero, we derive in [25] the MSE-minimizing estimator where each singular value is multiplied by the ratio of the signal variance over the total variance. To estimate the rank k we use the (generalized) cross-validation that I suggested in [10]. In experiments, this estimator and the ATN one have shown excellent recovery properties, especially when the SNR is moderate to high.

Bootstrap-based regularization All previous estimators are rooted in the SVD. In [11] we suggested an alternative regularization based on the bootstrap: for an observed matrix $X \sim \mathcal{L}(\mu)$, the optimal rank k linear estimator for μ would be

$$\hat{\mu}^{(k)^*} = XB^{(k)^*} \text{ where } B^{(k)*} = \underset{B}{\arg\min} \left\{ \mathbb{E}_{X\sim\mathcal{L}(\mu)}[\|\mu - XB\|_2^2] : \text{rank}\,(B) \leq k \right\}.$$

Yet, this estimator is infeasible: it involves computing an integral over an unknown data distribution. Hence, we propose replacing the draws $X \sim \mathcal{L}\,(\mu)$ from the unknown distribution with bootstrap draws $\widetilde{X} \sim \tilde{\mathcal{L}}_\delta\,(X)$, resulting in an estimator that we call the ***stable autoencoder***. In the simplest case (isotropic noise model with Gaussian parametric bootstrap), our method is equivalent to a classical SVD estimator. For non-isotropic cases (e.g., Poisson noise), the method does not reduce to singular value shrinkage, and yields new estimators performing well in practice. By iterating our stable autoencoding scheme, we automatically generate low-rank estimates without specifying the target rank as a tuning parameter.

I used all these works to derive extensions for missing values to carry out factor analyzes (PCA, MCA for categorical data) with missing data. We also have extended our results to regularize correspondence analysis (CA), e.g., for contingency tables such as text-word data. Implementation are available in the R packages **missMDA (250 download/day; > 400,000 in total; paper cited > 700 times)** & **FactoMineR (4 500 download/day; > 5 million in total, paper cited > 7000 times)**.

[1] Vincent Audigier, François Husson, and Julie Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10(1):5–26, 2016.

[2] Vincent Audigier, François Husson, and Julie Josse. Mimca: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and computing*, 27(2):501–518, 2017.

[3] Clément Bénard and Julie Josse. Variable importance for causal forests: breaking down the heterogeneity of treatment effects. *arXiv preprint arXiv:2308.03369*, 2023.

[4] Bénédicte Colnet, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Generalizing a causal effect: sensitivity analysis and missing covariates. *Journal of Causal Inference*, 2021.

[5] Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Reweighting the rct for generalization: finite sample error and variable selection. 2022.

[6] Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*, 2023.

[7] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gael Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review, 2023.

[8] Francois Husson, Julie Josse, Balasubramanian Narasimhan, and Genevieve Robin. Imputation of mixed data with multilevel singular value decomposition. *Journal of Computational and Graphical Statistics*, 28(3):552–566, 2019.

[9] Wei Jiang, Julie Josse, and Marc Lavielle. Logistic regression with missing covariates: Parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics and Data Analysis*, 145:106907, 2020.

[10] J. Josse and F. Husson. Selecting the number of components in pca using cross-validation approximations. *Computational Statististics and Data Analysis*, 56(6):1869–1879, 2011.

[11] J. Josse and S. Wager. Bootstrap-based regularization for low-rank matrix estimation. *Journal of Machine Learning Research*, 17(124):1–29, 2016.

[12] Julie Josse and Fran¸cois Husson. missmda: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):10–31, 2016.

[13] Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint*, 2019.

[14] Julie Josse and Jerome P. Reiter. Introduction to the special section on missing data. *Statist. Sci.*, 33(2):139–141, 05 2018.

[15] Julie Josse and Sylvain Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, 26(3):715–724, 2015.

[16] Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gael Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5980–5990. Curran Associates, Inc., 2020.

[17] Imke Mayer, Aude Sportisse, Julie Josse, Nicholas Tierney, and Nathalie Vialaneix. R-miss-tastic: a unified platform for missing values methods and workflows, 2021.

[18] Imke Mayer, Erik Sverdrup, Tobias Gauss, Jean-Denis Moyer, Stefan Wager, and Julie Josse. Doubly robust treatment effect estimation with missing attributes. *Ann. Appl. Statist.*, 14(3):1409–1431, 2020.

[19] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gael Varoquaux. What's a good imputation to predict with missing values? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[20] Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gael Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3165–3174. PMLR, 26–28 Aug 2020.

[21] Pavlo Mozharovskyi, Julie Josse, and Francois Husson. Nonparametric imputation by data depth. *Journal of the American Statistical Association*, 115(529):241–253, 2020.

[22] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In Hal Dauma III and Aarti Singh, editors, *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7130–7140. PMLR, 13–18 Jul 2020.

[23] Aude Sportisse, Claire Boyer, and Julie Josse. Estimation and imputation in probabilistic principal component analysis with missing not at random data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7067–7077. Curran Associates, Inc., 2020.

[24] Aude Sportisse, Claire Boyer, and Julie Josse. Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629?1643, nov 2020.

[25] M. Verbanck, F. Husson, and J. Josse. Regularized PCA to denoise and visualize data. *Statistics and Computing*, 25 (2):471–486, 2015.

[26] Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. *ICML*, 2023.

[27] Pan Zhao, Julie Josse, and Shu Yang. Efficient and robust transfer learning of optimal individualized treatment regimes with right-censored survival data. *arXiv preprint arXiv:2301.05491*, 2023.