# Factor screening in non-regular two-level designs based on mean squared error

**Yngvild Hamre**[1] **and John Tyssedal**[1]

[1] *Norwegian University of Science and Technology, Norway,*
*yngvild.hamre@ntnu.no/john.tyssedal@ntnu.no*

In this talk we would like to introduce a simple, yet efficient method for finding candidate sets when screening for active factors in non-regular two-level designs. In the first stage of an experiment, a large number of factors may have to be considered as potentially active. At that point, the main goal is to identify the ones that really influence the response. In most cases, the subspace of active factors is considerably smaller than the space of all factors. Factors not identified to have an impact on the response are normally not considered afterwards. Good and reliable methods for determining which factors are influential are therefore crucial.

The traditional choice of screening design has been two-level fractional factorial, also called regular designs. They have orthogonal columns and exist for $\frac{1}{2^p}, p = 1, 2, ...k-1$ fractions of $2^k$ factorial designs, where $k$ is the number of factors included in the design. The drawback of these designs is that effects may be fully aliased, making it difficult to separate the active effects from the rest. Non-regular two-level designs have therefore become increasingly popular. They project well onto lower dimensions and are far more flexible with regards to run sizes than regular designs. The alias structure may be complex, but the aliasing is often partial, making it possible to separate effects from each other. However, the partial aliasing between effects make traditional analysis methods fall short, as they usually rely on the ability to totally separate contrasts from each other. Thus there is a need for other methods for factor screening when using non-regular designs.

The method introduced in this talk starts by fitting the full projection models for all subsets of factors of a given size. Then the terms corresponding to the smallest coefficients are eliminated, and the reduced models with the remaining terms are fitted. Finally, the factors corresponding to the resulting models with the lowest mean squared errors (MSE) are chosen as the new candidates. This method allows for testing a wide range of designs and models. The main focus will be results found for 12- and 16-run non-regular designs with good projection properties for 3 and 4 active factors. The method is demonstrated by applying it on simulated models with varying characteristics and levels of noise. The number of candidate subsets of active factors is shown to be drastically reduced, while often preserving at least a 95% chance that the active factors are found among the remaining candidates. The performance when trying to identify 4 active factors is particularly interesting, as identifying 4 factors in 12 or 16 runs is considered challenging. The amount of noise that can be added before the performance decreases is also interesting. If the practitioner has an idea about the level of noise present in the investigated process, the results may yield an idea of which design should be used.

# Adoption spreading on partially observed social graphs

**Riccardo Parviero[1], Ida Scheel[2], Kristoffer H. Helton[3], Geoffrey Canright[4]**

[1] *University of Oslo – BigInsight, Norway, riccarpa@math.uio.no,*
[2] *University of Oslo, Norway,*
[3] *Norwegian Computing Center, Norway,*
[4] *Telenor, Norway*

Early prediction of the long-term behaviour of the adoption process of a new product in a market has important economic implications. Adoption processes can be driven by peer-to-peer (viral) influence mechanisms or by external factors, such as marketing campaigns. We propose an individual based method that is able to disentangle between these two sources of adoption pressure for customers whose relations can be described by a social graph.

When the true social graph is observed, our method provides an understanding on the relative importance of the forces driving the adoption process. The method estimates two sets of parameters $\boldsymbol{\alpha}$, associated to viral sources, and $\boldsymbol{\beta}$, associated to external ones.

When the true social graph is only partially observed, the presence of estimation error in $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ is inevitable. We investigate the situation in which not every interaction between consumers is observable, and provide bounds for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. Studying how these bounds behave also gives important insights on the portion of the true social graph that must be observed in order to have reliable estimates in this framework.

# Comparison of Ensemble-Based Data Assimilation Methods for Drift Trajectory Forecasting

**Florian Beiser[1,2,*], Håvard Heitlo Holm[1], and Jo Eidsvik[2]**

[1] *SINTEF Digital, Norway*
[2] *Norwegian University of Science and Technology, Norway*
[*] *florian.beiser@sintef.no*

Short-term prediction of drift trajectories is essential for several oceanic applications. In search-and-rescue operations at sea, for instance, it is crucial to efficiently define search areas that reflect both the forecasted trajectory and the associated uncertainties. To this end, we consider large ensembles of simplified ocean models consisting of non-linear shallow water equations that capture the relevant short-term physics. To reduce the uncertainty, we assimilate in situ buoy observations, which are typically very sparse compared to the high-dimensional state space.

We compare two state-of-the-art ensemble-based data assimilation methods for applications such as forecasting of drift trajectories.

The first method is a version of the ensemble-transform Kalman filter (ETKF), which is herein modified to be efficient for sparse point observations. The ETKF is very popular in numerical weather prediction, but is prone to introduce spurious correlations which are undesired from the statistical as well as physical perspective [1]. Therefore, we present a localisation scheme for the ETKF that centers low-dimensional analyses around the observation sites, making this a well-suited method for buoy observations in simplified ocean models.

The second method is the implicit equal-weights particle filter (IEWPF), a recently proposed method for non-linear filtering. Particle filters have traditionally been considered infeasible for geophysical applications due to the curse of dimensionality. This method, however, uses an implicit proposal density that avoids filter degeneracy, and has earlier been shown to be efficient for the relevant application [2].

We provide a comparison of the two data assimilation methods applied to a simplified ocean model, where we oppose the statistical properties of the state estimation and of the drift trajectory forecasts.

## References

[1] Hunt, B., Kostelich, E., Szunyogh, I. (2007). Efficient Data Assimilation for Spatiotemporal Chaos: A Local Ensemble Transform Kalman Filter. *Physica D: Nonlinear Phenomena*, **230**

[2] Holm, H., Sætra, M., and van Leeuwen, P. (2020). Massively parallel implicit equal-weights particle filter for ocean drift trajectory forecasting. *Journal of Computational Physics: X*, **6**, 100053.

# Quantification of the dating uncertainties in Greenland ice core records

**Eirik Myrvoll-Nilsen[1], Niklas Boers[1], Keno Riechers[1] and Martin Rypdal[2]**

[1] *Potsdam Institute for Climate Impact Research, Germany, myrvoll@pik-potsdam.de*
[2] *The University of Tromsø - The Arctic University of Norway, Norway*

Most layer-counting based paleoclimate proxy records have non-negligible uncertainties that arise from both the proxy measurement and the dating processes. Proper knowledge of the dating uncertainties in paleoclimatic ice core records is important for a rigorous propagation to further analyses; for example for identification and dating of stadial-interstadial transitions during glacial intervals, for model-data comparisons in general, or to provide a complete uncertainty quantification of early warning signals. We develop a statistical model that incorporates the dating uncertainties of the Greenland Ice Core Chronology 2005 (GICC05), which includes the uncertainty associated with layer counting. We express the number of layers per depth interval as the sum of a structural component that represents both underlying physical processes and biases in layer counting, described by a linear regression model, and a noise component that represents the internal variation of the underlying physical processes, as well as residual counting errors. We find that the dating uncertainties can be described by a multivariate Gaussian process that exhibits the Markov property, which grants a substantial gain in computational efficiency. Moreover, we investigate how tie-points from other proxy records can be used to match the GICC05 time scale to other chronologies and how this affects the dating uncertainty of Greenland interstadial transitions.