Technometrics, ISSN 0040-1706
Volume 62, number 3 (august 2020)

Abstract
Computer simulations often involve both qualitative and numerical inputs. Existing Gaussian process (GP) methods for handling this mainly assume a different response surface for each combination of levels of the qualitative factors and relate them via a multiresponse cross-covariance matrix. We introduce a substantially different approach that maps each qualitative factor to underlying numerical latent variables (LVs), with the mapped values estimated similarly to the other correlation parameters, and then uses any standard GP covariance function for numerical variables. This provides a parsimonious GP parameterization that treats qualitative factors the same as numerical variables and views them as affecting the response via similar physical mechanisms. This has strong physical justification, as the effects of a qualitative factor in any physics-based simulation model must *always* be due to some underlying numerical variables. Even when the underlying variables are many, sufficient dimension reduction arguments imply that their effects can be represented by a low-dimensional LV. This conjecture is supported by the superior predictive performance observed across a variety of examples. Moreover, the mapped LVs provide substantial insight into the nature and effects of the qualitative factors.

Abstract
Expensive black box systems arise in many engineering applications but can be difficult to optimize because their output functions may be complex, multi-modal, and difficult to understand. The task becomes even more challenging when the optimization is subject to multiple constraints and no derivative information is available. In this article, we combine response surface modeling and filter methods in order to solve problems of this nature. In employing a filter algorithm for solving constrained optimization problems, we establish a novel probabilistic metric for guiding the filter. Overall, this hybridization of statistical modeling and nonlinear programming efficiently utilizes both global and local search in order to quickly converge to a global solution to the constrained optimization problem. To demonstrate the effectiveness of the proposed methods, we perform numerical tests on a synthetic test problem, a problem from the literature, and a real-world hydrology computer experiment optimization problem.

Abstract
We develop a Bayesian nonparametric joint mixture model for clustering spatially correlated time series based on both spatial and temporal similarities. In the temporal perspective, the pattern of a time series is flexibly modeled as

a mixture of Gaussian processes, with a Dirichlet process (DP) prior over mixture components. In the spatial perspective, the spatial location is incorporated as a feature for clustering, like a time series being incorporated as a feature. Namely, we model the spatial distribution of each cluster as a DP Gaussian mixture density. For the proposed model, the number of clusters does not need to be specified in advance, but rather is automatically determined during the clustering procedure. Moreover, the spatial distribution of each cluster can be flexibly modeled with multiple modes, without determining the number of modes or specifying spatial neighborhood structures in advance. Variational inference is employed for the efficient posterior computation of the proposed model. We validate the proposed model using simulated and real-data examples.

Split Regularized Regression P. 330-338

Anthony-Alexander Christidis, Laks Lakshmanan, Ezequiel Smucler & Ruben Zamar

Abstract

We propose an approach for fitting linear regression models that splits the set of covariates into groups. The optimal split of the variables into groups and the regularized estimation of the regression coefficients are performed by minimizing an objective function that encourages sparsity within each group and diversity among them. The estimated coefficients are then pooled together to form the final fit. Our procedure works on top of a given penalized linear regression estimator (e.g., Lasso, elastic net) by fitting it to possibly overlapping groups of features, encouraging diversity among these groups to reduce the correlation of the corresponding predictions. For the case of two groups, elastic net penalty and orthogonal predictors, we give a closed form solution for the regression coefficients in each group. We establish the consistency of our method with the number of predictors possibly increasing with the sample size. An extensive simulation study and real-data applications show that in general the proposed method improves the prediction accuracy of the base estimator used in the procedure. Possible extensions to GLMs and other models are discussed. The supplemental material for this article, available online, contains the proofs of our theoretical results and the full results of our simulation study.

A Unified Approach to Sparse Tweedie Modeling of Multisource Insurance Claim Data P. 339-356

Simon Fontaine, Yi Yang, Wei Qian, Yuwen Gu & Bo Fan

Abstract

Actuarial practitioners now have access to multiple sources of insurance data corresponding to various situations: multiple business lines, umbrella coverage, multiple hazards, and so on. Despite the wide use and simple nature of single-target approaches, modeling these types of data may benefit from an approach performing variable selection jointly across the sources. We propose a unified algorithm to perform sparse learning of such fused insurance data under the Tweedie (compound Poisson) model. By integrating ideas from multitask sparse learning and sparse Tweedie modeling, our algorithm produces flexible regularization that balances predictor sparsity and between-sources sparsity. When applied to simulated and real data, our approach clearly outperforms single-target modeling in both prediction and selection accuracy, notably when the sources do not have exactly the same set of predictors. An efficient implementation of the proposed algorithm is provided in our R package MStweedie, which is available at https://github.com/fontaine618/MStweedie. Supplementary materials for this article are available online.

Model Misspecification of Generalized Gamma Distribution for Accelerated Lifetime-Censored Data P. 357-370

Marzieh Khakifirooz, Sheng Tsaing Tseng & Mahdi Fathi

Abstract

The performance of reliability inference strongly depends on the modeling of the product's lifetime distribution. Many products have complex lifetime distributions whose optimal settings are not easily found. Practitioners prefer to use simpler lifetime distribution to facilitate the data modeling process while knowing the true distribution. Therefore, the

effects of model mis-specification on the product's lifetime prediction is an interesting research area. This article presents some results on the behavior of the relative bias (*RB*) and relative variability (*RV*) of *p*th quantile of the accelerated lifetime (ALT) experiment when the generalized Gamma (GG_3) distribution is incorrectly specified as Lognormal or Weibull distribution. Both complete and censored ALT models are analyzed. At first, the analytical expressions for the expected log-likelihood function of the misspecified model with respect to the true model is derived. Consequently, the best parameter for the incorrect model is obtained directly via a numerical optimization to achieve a higher accuracy model than the wrong one for the end-goal task. The results demonstrate that the tail quantiles are significantly overestimated (underestimated) when data are wrongly fitted by Lognormal (Weibull) distribution. Moreover, the variability of the tail quantiles is significantly enlarged when the model is incorrectly specified as Lognormal or Weibull distribution. Precisely, the effect on the tail quantiles is more significant when the sample size and censoring ratio are not large enough. Supplementary materials for this article are available online.

Abstract
This article proposes a sampling technique that delivers robust designs, that is, point sets selected from a design domain in the shape of a unit hypercube. The designs are guaranteed to provide a *statistically uniform* point distribution, meaning that every location has the same probability of being selected. Moreover, the designs are *sample uniform*, meaning that each individual design has its points spread evenly throughout the domain. The *sample uniformity* (often measured via a *discrepancy* criterion) is achieved using distance-based criteria (ϕp or Maximin), that is, criteria normally used in space-filling designs. We show that the standard intersite metrics employed in distance-based criteria (Maximin and ϕp (phi)) do *not* deliver statistically uniform designs. Similarly, designs optimized via centered L_2 discrepancy or support points are also not statistically uniform. When these designs (after optimization based on intersite distances) are used for Monte Carlo type of integration, their statistical nonuniformity is a serious problem as it may lead to a systematic bias. This article proposes using a periodic metric to guarantee the statistical uniformity of the family of distance-based designs. The presented designs used as benchmarks in the article are only taken from the class of Latin hypercube designs, which forces univariate projections to be uniform and improves accuracy in Monte Carlo integration of some functions. Supplementary materials for this article are available online.

Abstract
Multivariate testing is a popular method to improve websites, mobile apps, and email campaigns. A unique aspect of testing in the online space is that it needs to be conducted across multiple platforms such as a desktop and a smartphone. The existing experimental design literature does not offer precise guidance for such a multi-platform context. In this article, we introduce a multi-platform design framework that allows us to measure the effect of the design factors for each platform and the interaction effect of the design factors with platforms. Substantively, the resulting designs are of great importance for testing digital campaigns across platforms. We illustrate this in an empirical email application to maximize engagement for a digital magazine. We introduce a novel "sliced effect hierarchy principle" and develop design criteria to generate factorial designs for multi-platform experiments. To help construct such designs, we prove a theorem that connects the proposed designs to the well-known minimum aberration designs. We find that experimental versions made for one platform should be similar to other platforms. From the standpoint of real-world application, such homogeneous subdesigns are cheaper to implement. To assist practitioners, we provide an algorithm to construct the designs that we propose.

Construction, Properties, and Analysis of Group-Orthogonal Supersaturated Designs P. 403-414

Bradley Jones, Ryan Lekivetz, Dibyen Majumdar, Christopher J. Nachtsheim & Jonathan W. Stallrich

Abstract

In this article, we propose a new method for constructing supersaturated designs that is based on the Kronecker product of two carefully chosen matrices. The construction method leads to a partitioning of the factors of the design such that the factors within a group are correlated to the others within the same group, but are orthogonal to any factor in any other group. We refer to the resulting designs as *group-orthogonal supersaturated designs*. We leverage this group structure to obtain an unbiased estimate of the error variance, and to develop an effective, design-based model selection procedure. Simulation results show that the use of these designs, in conjunction with our model selection procedure enables the identification of larger numbers of active main effects than have previously been reported for supersaturated designs. The designs can also be used in group screening; however, unlike previous group-screening procedures, with our designs, main effects in a group are not confounded. Supplementary materials for this article are available online.