PEOPLE
MARIE CURIE ACTIONS
International Outgoing Fellowships (IOF)
Call: FP7-PEOPLE-IOF-2008

Project Title: **Control Variates for Markov Chain Monte Carlo Variance Reduction "CVMCMC"**

Research Fellow: **Ioannis Kontoyiannis** yiannis@aueb.gr
Host Supervisor: **Petros Dellaportas** petros@aueb.gr
Project Number: **235837**

Project website: www.cvmcmc.eu

SUMMARY OF PROJECT ACTIVITIES

I. Objectives. Research in numerous scientific fields, including many of the fundamental research frontiers in science and engineering, has produced large empirical data sets with highly complex structure. There, the search for efficient ways of detecting and evaluating relevant information is currently one of the dominant problems, and the use of sophisticated statistical methods has been advanced as a necessary and central part of the analysis. In particular, recent advances in Markov chain Monte Carlo (MCMC) methods have revolutionized statistical analysis, vastly increasing its impact. The ability of MCMC algorithms to simulate from high-dimensional and potentially awkward distributions has made MCMC a major reason for – and an indispensable part of – the spectacular spread of statistical modeling to virtually every quantitative scientific area.

But in many applications, MCMC fails. On certain complex, high-dimensional problems, existing MCMC algorithms take too long to converge. And generally, the slow convergence of MCMC methods is one of the main limitations – perhaps *the* main limitation – of their applicability in statistics.

The *main objective* of this project is to develop a new family of efficient methodologies based on the method of control variates, for overcoming this obstacle in the context of statistical estimation. Our effort toward achieving this objective naturally brakes down into three separate directions, as described next.

II. Summary of work performed. After a preliminary phase during which a detailed literature survey on the larger topic of variance reduction in MCMC estimation was conducted, the first major task was the development of the necessary *theory*: A firm mathematical foundation was created for the application of control variates to Markov chains in general. In the second phase, the resulting theoretical insights led to the introduction of new generic *methodologies*: For the large class of so-called conjugate Gibbs samplers, we developed a clear, detailed and provably effective method for the direct use of control variates in the estimation process. Then, for the complementary class of Metropolis-Hastings samplers, we proposed several variants of the earlier basic methodology, which can be applied to a large number of the most common MCMC algorithms used in Bayesian inference studies.

The third part of the work was devoted to *applications*: We examined a large array of scientific and engineering problems of active research interest across the range of disciplines that require intensive use of statistical computing via MCMC, including genetics, health studies, environmental epidemiology, animal development and ecology, demography, statistical physics, computer science, signal processing, econometrics, astronomy and neuroscience. In each case we studied how our methodology needs to be tailored in order to be maximally effective, and in all cases the empirical results we obtained showed that this methodology is very effective in producing significantly more reliable statistical conclusions. Indeed, in some cases the difference was dramatic enough to suggest that certain tasks that up to now may have been considered impossible, may in fact now become possible.

Our findings are described in detail in a series of papers (see [1,2,3,4] below) as well as in the Ph.D. thesis of Zoi Tsourti, a graduate student at the Athens University of Economics and Business, jointly supervised by Kontoyiannis and Dellaportas. We have given more than 15 seminar and conference talks in leading institutions around the world, and in September of 2009 we organized a small, focused workshop on the topics of this project, which was attended by several of the world's leading experts.

III. Main results achieved. During the first phase of the project we concentrated on reversible MCMC algorithms, and the largest part of our work focused on the class of random-scan, conjugate Gibbs samplers. Our starting point was the adoption of a class of control variate functions proposed by Henderson in his 1997 Ph.D. thesis. Henderson's idea had been used effectively in the simulation of complex networks, but there were three serious obstacles in its potential application to statistical MCMC samplers. First, there was no general guideline for choosing among a vast array of possible such functions. Second, even if these functions had somehow been appropriately selected, there was no known way to compute or even approximate the coefficients in their optimal linear combination that should be used in the estimation process. And third, in the case of Metropolis-Hastings samplers, the direct method of constructing control variates introduced by Henderson could not be applied at all.

Our **first major breakthrough** was a theoretical result which allowed us to express the values of these optimal scalar coefficients (for the case of an arbitrary reversible chain) in a form that did not involve any implicitly defined quantities. This made it possible to introduce a new estimator for these coefficients and to rigorously establish its optimality, solving a problem which had been open for at least a decade. We emphasize that this is one of the rare instances in statistics where the theory does not come to simply justify interesting procedures after the fact; on the contrary, it is the theory itself that has prescriptive and operational value, suggesting a completely novel methodological procedure. The **second major result** we obtained also originated in a theoretical breakthrough. For the random-scan Gibbs sampler applied to a multivariate Gaussian distribution, we were able to obtain an explicit, exact solution for the associated *Poisson equation*. This is perhaps the only known example of an 'interesting' Markov chain naturally arising in applications, for which the Poisson equation has ever been solved. The existence of such a solution, together with the classical fact that the posterior distribution of interest is approximately Gaussian in most Bayesian inference problems, allowed us to overcome the other obstacle in the application of control variates in MCMC estimation: It provided an explicit guideline for choosing a very effective vector of control variate functions.

Combining these two breakthroughs led to our first methodological proposal: A simple algorithm for using control variates to dramatically reduce the MCMC estimation error, which can be applied to any 'conjugate' sampler used in any Bayesian inference problem. Numerous such examples based on real data are presented in [1,2,4].

In the last phase of this project we concentrated on the greatest challenge described in the original proposal, namely the extension of our results to the family of Metropolis-Hastings algorithms, a class of samplers which consists of many different MCMC methods with varying degrees of complexity, including the classical random walk Metropolis-Hastings method, the reversible jump Metropolis algorithm, the broad family of Metropolis-within-Gibbs methods, as well as numerous adaptive Metropolis schemes. Although the original class of control variates introduced by Henderson are not directly applicable in this case, we proposed several different methods in which our earlier methodology can be adapted. In the recent manuscript [3] we describe these ideas, and illustrate their applicability in practice by revisiting several problems of applied statistical inference. The results, in the settings examined so far, strongly indicate that these methods can often be extremely effective for variance reduction, and they are very encouraging for the continuation of this ongoing work.

V. Potential impact and use. By its very nature, methodological research into concrete problems in applied statistics is interdisciplinary; see, e.g., the list of potential applications mentioned above. In addition to working on the interface between applied statistical methodology and various other scientific disciplines, this project also involves a deeply theoretical component, which has provided *fundamentally new operational insights* on estimation problems using MCMC sampling. In terms of importance for science at large, in view of the enormous practical importance of the applications considered and the range of applicability of the MCMC algorithms under investigation, *the potential impact of even moderate theoretical or methodological advances can hardly be overestimated.*

[1] P. Dellaportas and I. Kontoyiannis. "Notes on using control variates for estimation with reversible MCMC samplers." *Technical report*, July 2009. Available online at the arXiv.

[2] P. Dellaportas and I. Kontoyiannis. "Control variates for estimation based on reversible MCMC samplers." Submitted to *J. Royal Stat. Society, Series B*, under revision.

[3] P. Dellaportas, I. Kontoyiannis and Z. Tsourti. "Control variates for Metropolis-Hastings samplers." *Technical report*, January 2011. Available online at `www.cvmcmc.eu`.

[4] P. Dellaportas and I. Kontoyiannis. "Applications of control variates in MCMC, beyond the basic methodology." *Preprint*, February 2011.