

Empirical and Theoretical Analysis of Bayesian Inference in Gaussian Process Models

Hannes Nickisch¹, Tobias Pfingsten¹, Carl Edward Rasmussen², Matthias Seeger³



The Gaussian process (GP) predictor is a popular kernel method allowing to express uncertainty about smooth functions [5]. In Bayesian inference, a GP prior is combined with the data to yield the posterior reflecting the remaining uncertainty about the function. Only for the case of regression with Gaussian noise, this procedure is analytically feasible. In all other cases, including probabilistic classification, one has to resort to approximations in order to make predictions. Even in regression, Bayesian model averaging is prohibitive. Approximate model selection can be done in two ways: The evidence framework [2] optimizes the marginal likelihood and the predictive approach [7] uses cross validation to approximate and optimize the predictive performance of a model.

Extending the work of [1] on the benefits of Expectation Propagation (EP) over the Laplace approximation (LA) in approximate classification, we analyze direct Kullback-Leibler divergence minimization (KL) and Variational Bounding (VB) which is an analytically convenient special case of KL [3]. Furthermore, the Factorial Variational (FV) algorithm, a mean-field approach as well as a heuristic approach termed label regression (LR) are included. In a unified treatment, we analytically characterize expected properties of the algorithms, we later confirm and broaden by extensive simulations including MCMC sampling on several benchmark datasets. We provide a clear statement about both theoretical and algorithmic advantages and shortcomings of the algorithms (Figure 1). In practice, an approximation method has to satisfy a wide range of requirements. If runtime is the major concern or one is interested in error rate only, LA or LR should be considered. Only EP and – although a lot slower – KL deliver accurate marginals as well as reliable class probabilities and allow for faithful model selection. For weakly coupled training data, FV can lead to quite reasonable approximations.

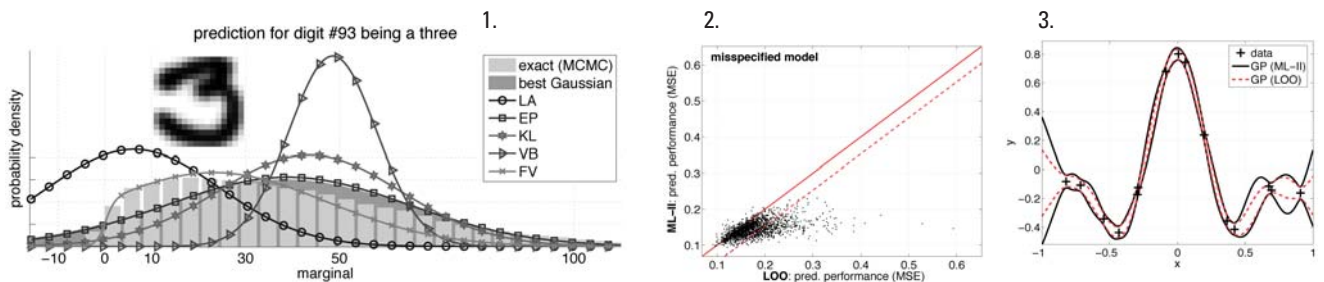
In GP regression, we compare predictive performance obtained by optimizing either the marginal likelihood or the leave-one-out predictive probability with Bayesian averaging done by Markov Chain Monte Carlo (MCMC) sampling [4]. In the absence of theoretical results, an empirical study based on a set of realistic

regression problems gives valuable insights into the properties of the methods. Compared to evidence optimization, the predictive framework, leads to predictions with significantly higher confidence and higher tendency to overfit training data (Figure 3). The predictive approach is consistently outperformed by the evidence framework, particularly on difficult problems (Figure 2). The best results are obtained using Bayesian averaging, however the computational burden of sampling is significant compared to optimization.

On the theoretical side, novel worst-case information consistency results for sequence prediction with GPs were obtained in [6]. As opposed to most other consistency results for nonparametric techniques, these results explicitly depend on the covariance function and the covariate distribution, so are in principle useful for model selection. They render new insights into the role of the spectral decay of the kernel operator, and have important implications for nonparametric minimum description length (MDL).

References

1. Kuss, M., C. E. Rasmussen: Assessing Approximate Inference for Binary Gaussian Process Classification. *Journal of Machine Learning Research* 6, 1679–1704 (2005).
2. MacKay, D. J. C.: Information-based Objective Functions for Active Data Selection. *Neural Computation* 4(4), 589–603 (1992).
3. Nickisch, H., C. E. Rasmussen: Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 9, 2035–2078 (2008).
4. Pfingsten, T., C. E. Rasmussen: Fully Bayesian Inference and Model Selection for Gaussian Process Regression. To be published.
5. Rasmussen, C. E., C. K. I. Williams: *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA (2006).
6. Seeger, M., S. Kakade, D. Foster: Information Consistency of Nonparametric Gaussian Process Methods. *IEEE Transactions on Information Theory* 54(5), 2376–2382 (2008).
7. Stone, M.: Cross-validators Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society B* 36(2), 111–147 (1974).



1. GP classification predictive marginal: Approximation results for handwritten digit recognition are compared to ground truth obtained by extensive MCMC sampling and the best Gaussian (same moments as ground truth). We show four Gaussian approximations and one non-Gaussian but factorial approximation as lines. The Laplace method (LA) underestimates the mean since marginal mode and joint mode do not coincide in high dimensions, Expectation Propagation (EP) yields close to exact results, direct Kullback-Leibler divergence minimization (KL) slightly underestimates the variance and overestimates the mean. Variational Bounding (VB) magnifies these effects. Finally, the Factorial Variational (FV) approximation yields quite appealing training marginals but has other shortcomings.
2. Mean Squared Error (MSE) on a difficult 10 dimensional dataset. Solid line: identity Figure 3: Illustrative example with lines indicating predictive mean \pm std. dev.

¹ Now at: OC&C Strategy Consultants, Düsseldorf, Germany; ² Also at: University of Cambridge, UK;

³ Now at: Universität des Saarlandes, Saarbrücken, Germany

Bayesian Experimental Design

Matthias Seeger¹, Hannes Nickisch, Florian Steinke, Koji Tsuda

Bayesian experimental design concerns the optimization of measurements in scientific experiments or industrial sensing, aiming for cost reductions or improved temporal resolution. Examples include systems biology problems, such as regulatory network reconstruction from protein or mRNA concentrations, optimization of neuroscientific experiments, or improved reconstruction in medical imaging. Roughly, the uncertainty about parameters of interest is quantified in a Bayesian model, and new measurements are optimized for maximum expected information gain. We recently developed novel Bayesian inference and experimental design frameworks for sparse models [5, 1, 2], and showed how they can be scaled up to large domains [4, 3]. The generality of these methods allows us to address problems in systems biology [5], compressed sensing [2], and magnetic resonance imaging [4]. Beyond extensions of this preliminary work, we plan to address problems of neural temporal data analysis (such as fMRI) and optimization of neuroscientific experiments. Our efficient Bayesian inference algorithms have applications beyond experimental design, for example in low level computer vision, digital photography, or in supporting decisions about causality. For the study in gene regulatory network recovery [5], we simulated data from realistic synthetic networks with nonlinear dynamics, comparing our novel method with the state-of-the-art. Our

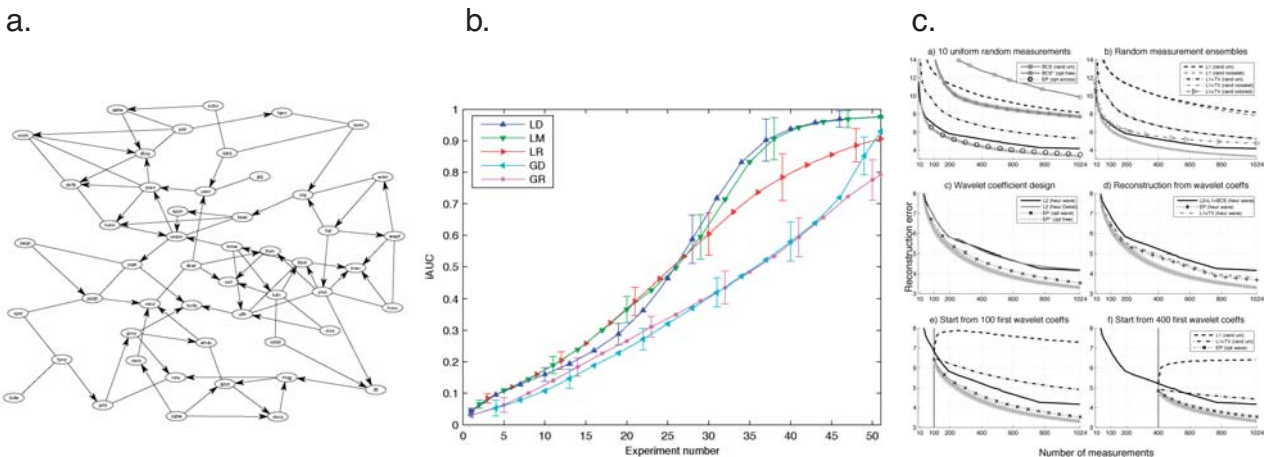
framework involves Bayesian inference on sparsity-favouring linear models, for which a novel variant of expectation propagation was used. For the study in [2], our measurement optimization algorithm was compared to a range of classical and modern compressed sensing alternatives, on a set of natural images frequently used in computer vision research. Our methods for the optimization of MRI sequences [4] are described separately [p. 55].

Our approach [5] outperformed the previous state-of-the-art in terms of recovery of interactions. Both the sparsity prior and the actively chosen experiments were shown to be of significant importance. We demonstrated in [2] that compressed sensing for natural images does not work well for random measurement filters, and explained why this is consistent with the underlying theory. Fixed measurement setups used in state-of-the-art image compression schemes significantly outperform random designs, yet can be improved upon by filters chosen by our Bayesian method (see Figure 1).



References

1. **Seeger, M.:** Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9: 759–813, 2008.
2. **Seeger, M., H. Nickisch:** Compressed sensing and Bayesian experimental design. In A. McCallum, S. Roweis, R. Silva, editors, *International Conference on Machine Learning 25*. Omni Press, (2008).
3. **Seeger, M., H. Nickisch:** Large scale variational inference and experimental design for sparse generalized linear models. *Technical Report TR-175*, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, (9/2008).
4. **Seeger, M., H. Nickisch, R. Pohmann, B. Schölkopf:** Bayesian experimental design of magnetic resonance imaging sequences. *Advances in Neural Information Processing Systems 21, Proceedings of the 2008 Conference*. (2009).
5. **Steinke, F., M. Seeger, K. Tsuda:** Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. *BMC Systems Biology*, 1(51), (2007).



1a. Example synthetic regulatory network. b. Results for network recovery: Our method (LD) outperforms a variant without sparsity prior (GD), and the approach without optimized measurements (LR). c. Results of compressed sensing study.

¹ Now at: Universität des Saarlandes, Saarbrücken, Germany

Bayesian Linear Gaussian State-space Based Models

Silvia Chiappa, David Barber¹, Jan Peters



Mixtures of and switching Linear Gaussian State-Space Models (LGSSMs) are probabilistic hidden variable models which are used in many real-world applications for time-series clustering and segmentation.

As an example, such models can be used to discover the underlying dynamical structure in Figure 1 (b), consisting of five different dynamical regimes M1-M5, from the unsegmented time-series in Figure 1 (a). In the most common scenario, the number of mixtures or segment-types (dynamical regimes) K is a priori unknown and needs to be estimated. Most clustering and segmentation models are designed to have a fixed structure, so that the determination of K is achieved by training and comparing several separate models with different structure. This approach has the disadvantage of introducing high computational overhead.

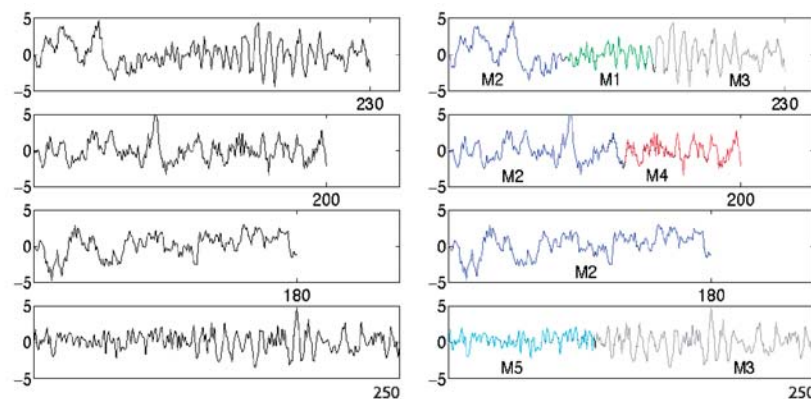
The goal of this project was to develop a Bayesian approach to mixtures of and switching LGSSMs which defines a prior distribution for the model parameters. The resulting models have a flexible structure such that the optimal K can be selected within the model during training, without the need to train and compare several separate models.

In our approach, an appropriate prior distribution on the parameters is used to enforce a sparse parametrization, such that the model automatically selects a small number of dynamical systems to explain the data. As the resulting model is computationally intractable, the major challenge was to develop an efficient approximation method in terms of accuracy, stability

and computational cost. We introduced a novel variational approximation in which a reformulation of the problem enables to use state-of-the-art inference algorithms (developed for the non-Bayesian approach), ensuring accuracy and stability. We demonstrated that our approach can successfully perform clustering and segmentation in a number of artificially generated datasets [1, 2]. We applied the mixture model to spike sorting and to cluster human-movement trajectories originating from different executions of the ball-in-a-cup game of dexterity for robot imitation learning [3]. In addition, we applied the switching model to segment long human-movement trajectories into basic units of motions [2].

References

1. **Chiappa, S., D. Barber:** Output Grouping using Dirichlet Mixtures of Linear Gaussian State-Space Models. In *Proceedings of the 5th International Symposium on Image and Signal Processing and Analysis (ISPA 2007)*, 446-451, IEEE Computer Society, Los Alamitos, CA, USA (2007).
2. **Chiappa, S.:** A Bayesian Approach to Switching Linear Gaussian State-Space Models for Unsupervised Time-Series Segmentation. In *7th Proceedings of the IEEE International Conference on Machine Learning and Applications*, 3-9, IEEE Computer Society, Los Alamitos, CA, USA (2008).
3. **Chiappa, S., J. Kober, J. Peters:** Using Bayesian Dynamical Systems for Motion Template Libraries. In *Advances in Neural Information Processing Systems 21: Proceedings of the 2008 Conference*, (2009).



1. (a) Four time-series and (b) their segmentation into five underlying dynamical regimes.

Causal Inference Rules

Dominik Janzing, Xiaohai Sun¹, Bernhard Schölkopf

Inferring causal relations between random variables using purely observational data is usually based on testing statistical independences and choosing causal graphs that satisfy the causal Markov condition and the faithfulness condition. The former states that every variable is conditionally independent of its non-descendants, given its parents. The latter selects graphs for which the joint distribution is generic, in the sense that it only satisfies the independences imposed by the Markov condition. The goal of this project is to address the following problems of independence-based causal inference: First, non-parametric conditional independence tests for continuous variables are challenging. In current algorithms, type one and type two errors can lead to major errors in edge orientation. Second, there is often a large number of (“Markov equivalent”) graphs satisfying Markov and faithfulness conditions.

To implement independence tests that do not require any assumptions other than smoothness of densities, we have further developed kernel independence tests and successfully applied them to learning causal Bayesian networks [1] and detecting non-linear Granger-causality [2]. Here, we have tested the performance on continuous and discrete variables as well as combinations of both. To address the second problem, several methods have been proposed by us and other groups [3, 4, 5, 6] that distinguish between Markov equivalent graphs by taking into account asymmetries in the shapes of conditional probability densities. It is likely that these approaches will turn out to be part of a whole family of new inference methods yet to be discovered. Our recent results indicate that new inference rules of the above kind can be justified by an algorithmic information theory approach, because conditional densities contain algorithmic information that provides additional hints about causal directions. For instance, the shortest description of the density $P(\text{cause}, \text{effect})$ is generically given by separate descriptions of $P(\text{cause})$ and $P(\text{effect}|\text{cause})$.

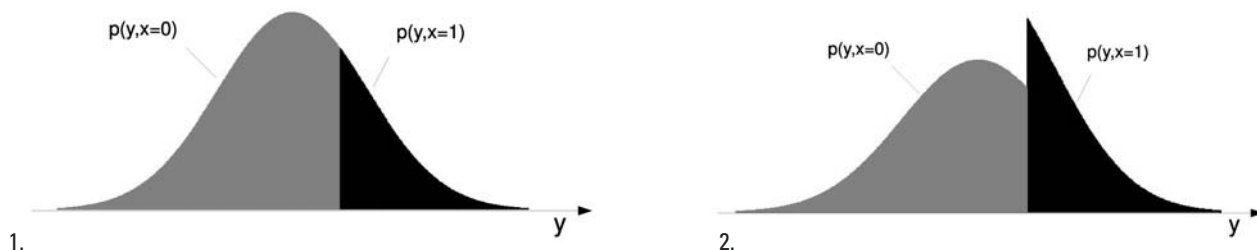
To use algorithmic information in a systematic way, we have postulated an *algorithmic* causal Markov condition [7] that links algorithmic conditional dependences between single observations to the underlying causal structure. By proving the equivalence

of three different versions of the Markov condition (the d-separation criterion, the local condition, and a sum rule for the joint algorithmic information) we have developed a consistent graphical model framework. Further implications of the theory for *statistical* causal inference remain to be studied.



References

1. Sun, X., D. Janzing, B. Schölkopf, K. Fukumizu: A Kernel-based Causal Learning Algorithm. In *Proceedings of the 24th International Conference on Machine Learning*, 855–862. (Ed.) Ghahramani, Z., Corvallis, OR, (2007).
2. Sun, X.: Assessing Nonlinear Granger Causality from Multivariate Time Series. In *Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2008*, 440-455. (Eds.) Daelemans, W., B. Goethals, K. Morik, Springer, Berlin, Germany (2008).
3. Kano, Y., S. Shimizu: Causal Inference using Non-normality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, 261–270. (Eds.) Higuchi, T., Y. Iba, M. Ishiguro, ISM Report on Research and Education, No.17, The Institute of Statistical Mathematics, Tokyo, Japan (2003).
4. Sun, X., D. Janzing, B. Schölkopf: Causal Inference by Choosing Graphs with Most Plausible Markov Kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics (AI&Math 2006)*, 1–11 (2006).
5. Sun, X., D. Janzing, B. Schölkopf: Causal Reasoning by Evaluating the Complexity of Conditional Densities with Kernel Methods. *Neurocomputing* 71, 1248–1256 (2008).
6. Hoyer, P., D. Janzing, J. Mooij, J. Peters, B. Schölkopf: Nonlinear Causal Discovery with Additive Noise Models. In *Advances in Neural Information Processing Systems 21: Proceedings of the 2008 Conference* (2009).
7. Janzing, D., B. Schölkopf: Causal Inference Using the Algorithmic Markov Condition. <http://arxiv.org/abs/0804.3678>, 1–46 (2008).



1. Joint density $p(x, y)$ of a real-valued random variable Y and a binary variable X . The marginal distribution $p(y)$ is Gaussian. The causal hypothesis $Y \rightarrow X$ is plausible: the conditional $p(x|y)$ corresponds to setting $x = 1$ for all y above a certain threshold. We reject the converse hypothesis $X \rightarrow Y$ because $p(y|x)$ and $p(x)$ share algorithmic information: given $p(y|x)$, only *specific* choices of $p(x)$ reproduce the Gaussian $p(y)$, whereas generic choices of $p(x)$ would yield “odd” densities of the type in Figure 2.

2. Joint density $p(x, y)$ obtained by changing $p(x)$ and keeping $p(y|x)$ as in Figure 1.

¹ Now at: Altran CIS, Frankfurt, Germany

Asymmetries of Time Series under Time Inversion

Jonas Peters, Dominik Janzing, Arthur Gretton, Bernhard Schölkopf



Consider the following problem: We are given m ordered values X_1, \dots, X_m from a time series, but we do not know if the sample has been reversed. Our task is to find out whether X_1, \dots, X_m or X_m, \dots, X_1 represent the true time direction.

The aim of this project is to learn more about the statistical asymmetries between cause and effect. As the past causes the future, but not vice versa (every cause precedes its effect), studying the direction of time series might provide insights helpful for causal inference. We can use the temporal ordering of cause and effect to solve the time direction problem in the following way: if we identify one cause and its effect in the data X_1, \dots, X_m , we can order the whole series. Say, we found that X_5 is causing X_2 , then the true time ordering must be X_m, \dots, X_1 . This idea can be formalized using a linear model for time series, namely an autoregressive moving average (ARMA) process. These processes are defined to be stationary and to satisfy

$$X_t = \sum_{i=1}^p \phi_i X_{t-i} + \sum_{j=1}^q \theta_j \epsilon_{t-j} + \epsilon_t \quad \forall t \in \mathbb{Z},$$

for some coefficients ϕ_i, θ_j and iid noise ϵ_t . Such a process is called causal if the noise does not depend on the last values X_{t-1}, \dots of the time series.

For a normally distributed ARMA process it is well-known that the reversed process satisfies the same (causal) ARMA equation. We proved, however, that this is the only case, where the reversed process is again a causal ARMA process; that means if the noise of a causal ARMA process follows any distribution other than the Gaussian, the reversed process is not a causal ARMA process anymore.

Now we assume the data to follow a causal ARMA process (with non-Gaussian noise), which we regard as a relative simple model. The statement above tells us that the reversed time series is not that simple anymore and satisfies a more complex model. This simplicity idea of causal reasoning can also be found in the LiNGAM approach, for example.

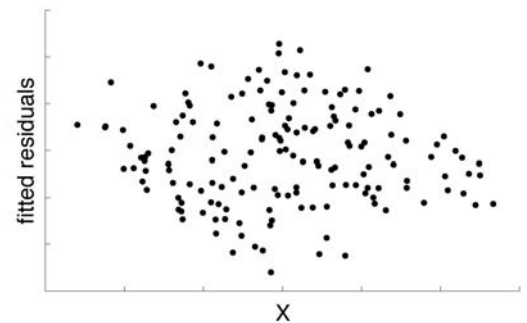
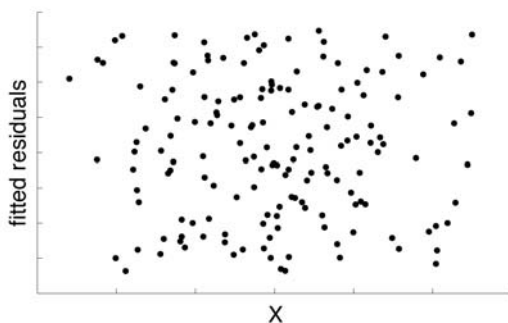
We use our theoretical result to propose a method for detecting the true time direction as follows:

1. Fit an ARMA process to both directions (X_1, \dots, X_m) and (X_m, \dots, X_1) and compute the fitted residuals.
2. If the residuals seem to be Gaussian do not make a decision. To check this, we used the so-called Jarque-Bera test.
3. Use an independence test (e.g. HSIC [1]) to test if the forward residuals depend on X_{t-1}, X_{t-2}, \dots or if the backward residuals depend on X_{t+1}, X_{t+2}, \dots . According to our result only one dependence should be found. If the independence is indeed rejected for only one direction, propose the other direction to be the correct one (see Figure). If both directions seem to lead to dependent noise, conclude that the model fit is not good enough and do not decide.

Our method works well on simulated and on real data if the assumption of an ARMA model is reasonable.

References

1. Gretton, A., O. Bousquet, A. Smola, B. Schölkopf: Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005*, 63–78 (2005).



Simulated AR process with uniformly distributed noise: The fitted residuals of the forward model (left) and of the backward model (right) are plotted against past time series values. The fit in the wrong direction leads to a strong dependence between residuals and time series, the residuals of the forward model are regarded as independent.

Learning Structural Causal Models from Data

Joris Mooij, Dominik Janzing, Jonas Peters, Bernhard Schölkopf

A *structural causal model* [1] is defined as follows. Given a directed acyclic graph (DAG), denote the parents of node i as $\pi(i)$. For each node i , the corresponding random variable x_i (“effect”) is a function $x_i = f_i(x_{\pi(i)}, \epsilon_i)$ of the random variables $x_{\pi(i)}$ (“causes”) associated with the parents $\pi(i)$ of i and an independent noise source ϵ_i . Most existing methods for learning structural causal models from observational data assume that all variables are discrete, or that all variables are continuous, but that all functions are linear and the noise is additive Gaussian. Apart from the fact that those assumptions are often not met in practice, it has been pointed out recently that they can actually exacerbate the problem of causal inference. Indeed, for linear models, *non-Gaussianity* in the data can actually aid in distinguishing causal directions [2]; similarly, we have shown recently that *nonlinearity* of the functional relationships can aid in identifying the causal structure [3].

We intend to develop methods for learning structural causal models from observational data that exploit the asymmetries between cause and effect generated by nonlinearity and non-Gaussianity. In addition, we plan to tackle the case of non-additive noise, which often occurs in practice, in particular for causal relations between discrete variables.

We have studied the case of *additive* noise, and proposed the following approach to learning: [3] for a given candidate DAG, one solves a regression problem for each node i to obtain the function that yields the best fit to the data. The candidate model is accepted if all residuals are independent. For the more general case of non-additive noise, more sophisticated methods have to be developed. In a nutshell, the idea is to unite the regression

and independence tests by maximizing an independence measure, while simultaneously penalizing the complexity of the regression functions.

We proved for additive noise that the causal structure is identifiable in the special case where there are only two variables and that either the function is nonlinear or the noise is non-Gaussian [3]. We have also done a preliminary empirical study of the causal inference method described above, which has shown that if all assumptions are met, the correct causal structure can often be identified.

We have proposed a novel method for learning causal models that exploit nonlinearities or non-Gaussianities in the data, with promising results. Overcoming the limitation of the additive noise assumption would open up a large range of possible applications in different fields.



References

1. Pearl, J., *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA (2000).
2. Shimizu, S., P. O. Hoyer, A. Hyvärinen, A. J. Kerminen: A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* 7, 2003–2030 (2006).
3. Hoyer, P. O., D. Janzing, J. Mooij, J. Peters, B. Schölkopf: Nonlinear Causal Discovery with Additive Noise Models. In *Advances in Neural Information Processing Systems 21: Proceedings of the 2008 Conference* (2009).

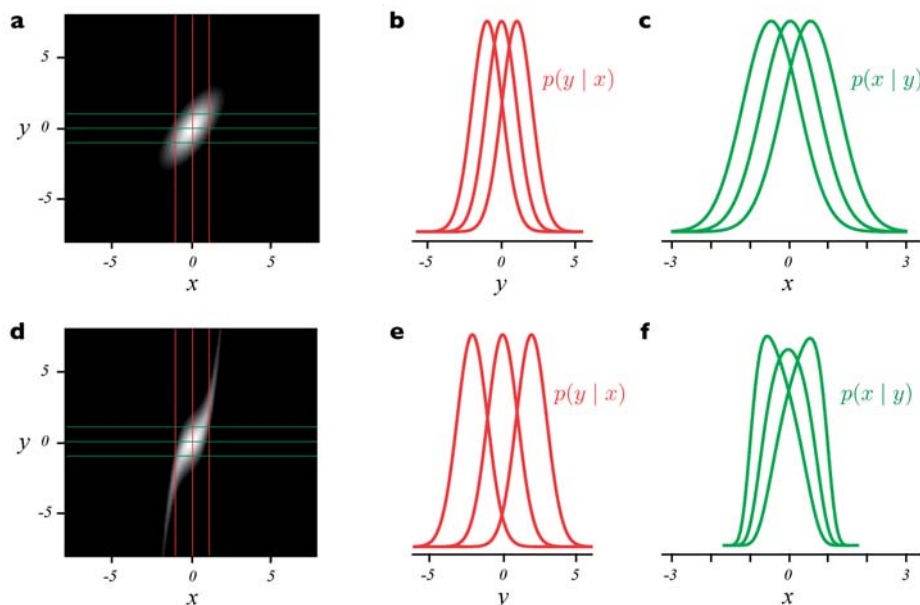


Illustration of the causal identifiability due to nonlinearities for the case of two variables. The model generating the data is $y = f(x) + \epsilon$, where x and ϵ are both Gaussian and independent. The linear case $f(x) = x$ is shown in (a–c) and the nonlinear case $f(x) = x + x^3$ is illustrated in (d–f). The joint densities $p(x, y)$ of the observed variables are shown in (a) and (d). In both cases, the conditional densities $p(y|x)$ have the same shape for all values of x , as shown in (b) and (e). In general, there is no reason to believe that this relationship would also hold for the conditionals $p(x|y)$, but, as is well known and illustrated in (c), for the linear-Gaussian model this is actually the case. However, notice how in the nonlinear case (f) the conditional densities $p(x|y)$ look different for different values of y , indicating that a reverse causal model of the form $x = g(y) + \eta$ (with y and η independent) would not fit the joint density.