Network reconstruction with realistic models

Marco Grzegorczyk[1], Andrej Aderhold[2], Dirk Husmeier[2]

[1] Johann Bernoulli Institute, Groningen University, The Netherlands
[2] School of Mathematics and Statistics, Glasgow University, UK

E-mail for correspondence: m.a.grzegorczyk@rug.nl

Abstract: We extend a recently proposed gradient-matching method for inferring interactions in complex systems described by differential equations in various respects: improved gradient inference, evaluation of the influence of the prior on kinetic parameters, comparative evaluation of two model selection paradigms: marginal likelihood versus DIC (divergence information criterion), comparative evaluation of different numerical procedures for computing the marginal likelihood, extension of the methodology from protein phosphorylation to transcriptional regulation, based on a realistic simulation of the underlying molecular processes with Markov jump processes.

Keywords: Gene Networks; Semi-Mechanistic Models; Bayesian Model Selection

1 INTRODUCTION

A challenging problem for computational statistics is to infer the structure of regulatory networks from postgenomic data. Two approaches can be distinguished. The first paradigm aims to apply generic models like Bayesian networks. The second paradigm is based on mechanistic models and the detailed mathematical description of the underlying interactions with differential equations (DEs). The advantage of this paradigm is a more faithful representation of the interactions. The disadvantage are the substantially higher computational costs of inference. A novel approach, presented by Oates et al. (2014) and termed "chemical model averaging" (CheMa), aims for a compromise between the two paradigms by gradient matching. Given the concentration time series of some quantities ("species") whose interactions are to be inferred, the temporal derivatives of the concentrations are estimated from the data. These derivatives are then matched against those predicted from the DEs by standard statistical techniques.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

The resulting "semi-mechanistic" model is effectively a non-linear regression model, whose computational complexity of inference sits between the two paradigms. We take the work of Oates et al. (2014) further in five respects. We improve the gradient matching by adopting techniques from nonparametric Bayesian statistics with Gaussian processes. We assess the influence of the parameter prior. We carry out a comparative evaluation study to assess two model selection paradigms – the marginal likelihood versus the divergence information criterion. We adapt the method to a new type of application, namely to transcriptional regulation from gene expression data. We evaluate the method in a realistic simulation study based on a stochastic process description of the underlying molecular processes.

2 METHODS

A biopathway can be modelled as a system of ordinary DEs:

$$\frac{dx_i}{dt}|_{t=t_j} \;=\; c_i - v_{0,i}x_i(t_j) + f_i(\mathbf{x}_i(t_j), \boldsymbol{\theta}) \tag{1}$$

where i is one of N species, $x_i(t_j)$ is the concentration of i at time point t_j $(j = 1, \ldots, T)$, c_i is a baseline production rate, $v_{0,i}$ is a decay rate, $f_i(.)$ is a regulation function, $\boldsymbol{\theta}$ are parameters, and $\mathbf{x}_i(t_j)$ is a vector of concentrations of species that regulate species i. We follow Oates et al. (2014) and estimate the time derivatives $\frac{dx_i}{dt}|_{t=t_j}$ from the observed data D, and treat the problem as non-linear regression with the likelihood:

$$p(D|\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{j=1}^{T} \mathcal{N}(\xi_i(t_j)|f_i(\mathbf{x}_i(t_j), \boldsymbol{\theta}) - v_{0,i}x_i(t_j), \sigma_i^2) \tag{2}$$

where $\xi_i(t_j) = \frac{dx_i}{dt}|_{t=t_j}$, and $\mathcal{N}(.|\mu, \sigma^2)$ is the PDF of a normal distribution. Oates et al. (2014) obtained the temporal derivatives $\xi(t_j)$ with a finite difference quotient (**"numerical gradient"**). We here propose to apply a Gaussian process to smooth interpolation (**"analytical gradient"**). The superiority of this approach was recently established by Aderhold et al. (2014). In the CheMA approach Eq. (1) is implemented with $c_i = 0$, and $f_i(.)$ describes Michaelis-Menten kinetics:

$$\xi_i(t_j) = \frac{dx_i}{dt}|_{t=t_j} = -v_{0,i}x_i(t_j) + \sum_{j\in\pi_i} v_{j,i}\frac{I_{j,i}\cdot x_j(t_j) + (1 - I_{j,i})\cdot k_{j,i}}{x_j(t_j) + k_{j,i}} \tag{3}$$

where the sum is over all species $j \in \pi_i$ that are regulators of i. The indicator function $I_{j,i}$ indicates whether species j is an activator or inhibitor. The term $-v_{0,i}x_i(t)$ takes the degradation of $x_i(t)$ into account, while the parameters $v_{j,i}$ and $k_{j,i}$ are the "*maximum reaction rate*" and "*Michaelis-Menten*" parameters. Oates et al. (2004) impose truncated Normal distributions on $k_{j,i}$, $k_{j,i} \sim \mathcal{N}_{\{k_{j,i}\geq 0\}}(1, \nu)$, where $\nu > 0$, use Jeffrey's prior

for σ_i^2, and a truncated **g-prior** on the vector $\mathbf{V}_i$ of maximum reaction rate parameters: $\mathbf{V}_i \sim \mathcal{N}_{\{\mathbf{V}_i \geq 0\}}(\mathbf{1}, T\sigma_i^2(\mathbf{D}_i^{\mathrm{T}}\mathbf{D}_i)^{-1})$, where $\mathbf{D}_i$ is the design-matrix for species i. We show that a truncated **ridge regression prior**: $\mathbf{V}_i \sim \mathcal{N}_{\{\mathbf{V}_i \geq 0\}}(\mathbf{1}, \delta_i^2\sigma_i^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix, and δ_i^2 has an inverse Gamma prior, $\delta_i^2 \sim IG(a_\delta, b_\delta)$, yields a better network reconstruction accuracy. We refer to these methods as **CheMa** (with the g-prior on $\mathbf{V}_i$) and **iCheMa** (improved CheMa with the ridge regression prior on $\mathbf{V}_i$).

To infer the regulator sets π_i of the interaction processes described by Eq. (1), we compare the divergence information criterion (DIC) and the marginal likelihood (MLL). **DIC** is defined as

$$DIC(\pi_i) = 2\log p(D|\overline{\boldsymbol{\theta}}, \pi_i) - 4\int \log p(D|\boldsymbol{\theta}, \pi_i)p(\boldsymbol{\theta}|\pi_i, D)d\boldsymbol{\theta}$$

where $\overline{\boldsymbol{\theta}} = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\pi_i, D)d\boldsymbol{\theta}$ is the posterior mean of the parameters. The integrals are approximated by sums over parameters sampled from the posterior distribution $p(\boldsymbol{\theta}|G, D)$ with MCMC. The marginal likelihood is

$$p(D|\pi_i) \;=\; \int p(D|\boldsymbol{\theta}, \pi_i)p(\boldsymbol{\theta}|\pi_i)d\boldsymbol{\theta} \tag{4}$$

We compare two methods for approximating Eq. (4): Chib's method and thermodynamic integration (TI). **Chib's method** is based on

$$p(D|\pi_i) \;=\; \frac{p(D|\boldsymbol{\theta}^\star, G)p(\boldsymbol{\theta}^\star|\pi_i)}{p(\boldsymbol{\theta}^\star|\pi_i, D)} \tag{5}$$

where the posterior near a selected parameters $\boldsymbol{\theta}^\star$, $p(\boldsymbol{\theta}^\star|G, D)$, is approximated with MCMC. **TI** is based on the power posteriors:

$$p(\boldsymbol{\theta}|\pi_i, D, \tau) \;=\; \frac{p(D|\boldsymbol{\theta}, \pi_i)^\tau p(\boldsymbol{\theta}|\pi_i)}{\int p(D|\boldsymbol{\theta}', \pi_i)^\tau p(\boldsymbol{\theta}'|\pi_i)d\boldsymbol{\theta}'} \tag{6}$$

from which the marginal likelihood is computed via

$$p(D|\pi_i) \;=\; \int_0^1 E_{\boldsymbol{\theta},\tau}\big[\log p(D|\boldsymbol{\theta}, \pi_i)\big]d\tau \tag{7}$$

Here $E_{\boldsymbol{\theta},\tau}[.]$ is an expectation w.r.t. the power posterior in Eq. (6). These expectations are computed for various temperatures $0 \leq \tau \leq 1$ with population MCMC, and the integral in Eq. (7) is then approximated with the trapezium sum. We choose 10 temperatures $\tau_i = (\frac{i}{9})^m$ $(0 \leq i \leq 9)$, and we vary the exponent $m \in \{4, 8\}$ to obtain **TI-4** and **TI-8**.

3 DATA

We generate $T = 240$ data points for four species $x_1, \ldots, x_4$ from iid $\mathcal{N}(0, 1)$ distributions. Subsequently, to obtain non-negative concentrations, the observations of each individual species are shifted such that the lowest value

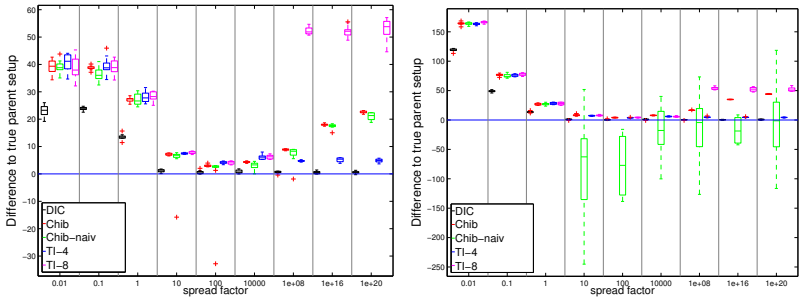

FIGURE 1. **Performance of iCheMa on synthetic data** The box plots show the log marginal likelihood differences between the true regulator set $\pi_y = \{x_2\}$ and an alternative over-complex set that includes one irrelevant regulator. The panels correspond to different prior distributions of the parameters. **Left panel**: the prior variance of $k_{j,i}$ was kept fixed at $\nu = 0.5$, and the hyperparameter δ_i^2 of the truncated ridge regression prior on $\mathbf{V}_i$ was set to *spread factor*. **Right panel**: both hyperparameters δ_i^2 and ν were set to *spread factor*. The box plots show the distributions of the average MLL differences across 9 parameter pairs $(v_{0,y}, v_{2,y})$, computed for 10 independent data instantiations. Positive differences indicate that the true model is favoured. Each panel shows: (i) the DIC difference (**DIC**), the MLL differences, approximated with (ii) a naive implementation of Chib's method (**Chib-naiv**), (iii) an improved implementation of Chib's method (**Chib**), (iv) TI with $m = 4$ (**TI-4**), and (v) TI with $m = 8$ (**TI-8**).

is equal to 0, before we re-scale the observations of each species to mean 1. With x_1 taking the role of the degradation process and x_2 being an activating regulator ($I_{2,y} = 0$) of a gradient ξ_y, which we here assume to be directly observable, we generate target observations with Eq. (3): $\xi_y(t_j) = -v_{0,y}x_1(t_j) + v_{2,y}\frac{x_2(t_j)}{x_2(t_j)+k_{2,y}} + \epsilon_{t_j}$, where $\epsilon_{t_j} \sim \mathcal{N}(0, \sigma^2)$ is additive iid Gaussian noise. We keep $k_{2,y} = 1$ fixed, and vary the rates $(v_{0,y}, v_{2,y}) = \{(1,1), (0.5,1), (1.5,1), (2,1), (0.2,1), (2,0.2), (3,0.1), (0.2,2), (0.1,2)\}$. Our goal is to infer the true regulator set $\pi_y = \{x_2\}$ out of all subsets of $\{x_2, x_3, x_4\}$. The degradation, modelled via x_1, is included in all models.
We also use the benchmark data from Aderhold et al. (2014), which contain realistically simulated gene expression time series for genes in the circadian clock of *Arabidopsis thaliana*. We focus on the mRNA data points and those time series generated for the wildtype circadian gene network, shown in the left panel of Figure 2. The molecular interactions were modelled as individual discrete events with a Markov jump process and practically simulated with the Biopepa software, see Aderhold et al. (2014) for details.
Finally, we apply the improved CheMa model to Arabidopsis gene expression data which were recently measured under the EU-FP7-funded Timing Metabolism (TiMet) research project. For space restrictions the latter results are not shown in this paper.

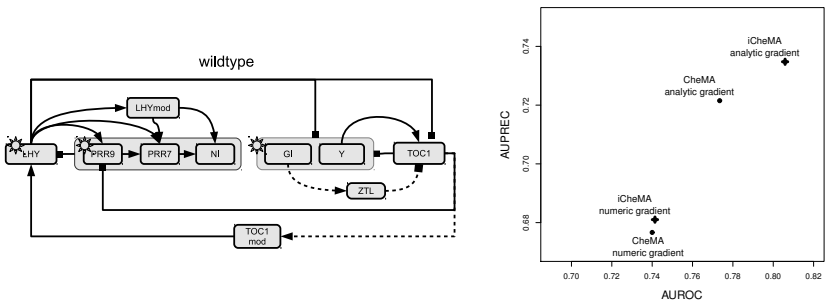

FIGURE 2. **Network reconstruction accuracy for the realistic wildtype Arabdidopsis gene network, based on marginal interaction probabilities ("model averaging").** *Left panel:* Hypothetical circadian clock network in *A.thaliana* from Pokhilko et al. (2010). *Right panel:* Mean AUROC and AUPREC scores to quantify the effects of the **prior** ('CheMa vs. iCheMa') and the **gradient type** ('numerical vs. analytical') on the network reconstruction accuracy.

4 RESULTS

SYNTHETIC DATA: Figure 1 shows the DIC and log marginal likelihood (MLL) differences between the true and an over-complex regulator set for different parameter priors. Each parameter prior was a Gaussian centred on $\mu = 1$, with different variances. For low variances, both DIC and MLL clearly favour the true network, because the prior 'pulls' the spurious interaction parameter from its true value of zero towards a wrong value of $\mu = 1$. As the prior becomes more diffuse, both the DIC and MLL differences become less pronounced, but still select the true model up to spread factors of about 100. As the prior becomes more diffuse, with the spread factor exceeding 100, DIC fails to select the correct model. MLL, on the other hand, starts to increasingly favour the true model as the spread factor further increases beyond 1000. This is a consequence of Lindley's paradox, whereby MLL increasingly penalizes the over-complex model for increasingly vague priors.

The left panel of Figure 1 shows that the different ways of computing the MLL give very similar results up to a prior spread factor of about 1e+08. For spread factors exceeding this value, the results differ. The MLL computed with Chib's method monotonically increases, as expected from Lindley's paradox. MLL computed with TI reaches a plateau, with different values obtained for different temperature schemes ($m = 4$ and $m = 8$). This is a numerical discretization error that results from the form of the integrand in Eq. (7), which has most of its area concentrated on values near $\tau = 0$. The right panel shows that a naive implementation of Chib's method can run into numerical instabilities, as diffuse prior distributions can yield

suboptimal attractor states in parameter configuration space. We fixed this instability by selecting exclusively pivot parameter vectors $\boldsymbol{\theta}^\star$ that are representative for the sampled parameter values. (We never observed unstable results when selecting the parameter vector $\boldsymbol{\theta}^\star$ with the highest posterior probability from only the sample phase of the MCMC simulation, rather than the total MCMC trajectory with the burn-in phase included.) With this numerical stabilization, Chib's method, whose numerical complexity compared to TI is lower by a factor of about 10 (as we use $K = 10$ temperatures τ_i for TI), is the favourite method for the realistic network data.
REALISTIC NETWORK DATA: The right panel of Figure 2 shows average AUROC (area under the ROC curve) and AUPREC (area under the precision recall curve) scores, obtained for five realistically simulated time series of the wildtype circadian clock network (shown in the left panel). The proposed analytical gradient (based on a Gaussian process) yields a significantly improved network reconstruction accuracy for both models: (i) CheMa with the truncated g-prior on $\mathbf{V}_i$, as proposed by Oates et al. (2014) and (ii) iCheMa with the truncated ridge regression prior, introduced here. It can also be seen that iCheMa has a slightly better network prediction accuracy than the original CheMa method given the numerical gradient, and a substantially better accuracy for the analytic gradient.

5 FURTHER RESULTS

At IWSM 2015 more results will be presented. For space restrictions some results could not be included in this paper. Most importantly, we will (i) show that the iCheMa model is superior to established network reconstruction methods, such as Hierarchical Bayesian regression, Sparse regression with L_1 penalty (Lasso), Sparse regression with L_1 and L_2 penalty (ElasticNet), Sparse regression with change-points (Tesla), Sparse Bayesian regression, Graphical Gaussian models, Bayesian spline autoregression, State-space models, Gaussian processes, and Bayesian networks, and we will (ii) apply the novel iCheMa method to reverse-engineer the circadian clock network in *A. thaliana* from TiMet gene expression time series.

References

Aderhold, A., Husmeier, D., and Grzegorczyk, M. (2014). Statistical inference of regulatory networks for circadian regulation. *Statistical applications in genetics and molecular biology*, **13(3)**, 227 – 273.

Oates et al. (2014) Causal network inference using biochemical kinetics. *Bioinformatics*, **30(17)**, 468 – 474.

Pokhilko et al. (2010). Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, **6(1)**, online article.