Predictions in Hierarchical Bayesian Models

Georgios Arampatzis

Professorship for Computational Science, ETH-Zurich, CH-8092, Switzerland

September 23, 2018

Let Y_i, Θ_i for $i = 1, \ldots, N$ and Ψ be random variables with $Y_i \in \mathbb{R}^{N_{Y_i}}$, $\Theta_i \in \mathbb{R}^{N_\Theta}$ and $\Psi \in \mathbb{R}^{N_\Psi}$. The dependency of the random variables is described by the Directed Acyclic Graph (DAG) given in Figure 1. The joint probability distribution is given by

$$p(\boldsymbol{d}, \boldsymbol{\vartheta}, \psi) = \prod_{i=1}^{N} p(d_i|\vartheta_i)\, p(\vartheta_i|\psi) p(\psi) = p(\boldsymbol{d}|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\psi)\, p(\psi) = p(\boldsymbol{d}|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}, \psi)\ , \tag{1}$$

with $\boldsymbol{d} = (d_1, \ldots, d_N)$ and $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_N)$. In Bayesian inverse problems this graph describes the output out a model, f, that depends on parameters ϑ, under the hypothesis of random errors. The noise is usually modelled and the function $p(y_i|\vartheta_i)$ is fully determined. A usual assumption is the following,

$$d_i = f(\vartheta_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Sigma_i)\ , \tag{2}$$

with $\Sigma_i := \Sigma(\vartheta_i)$ and thus $p(d_i|\vartheta_i) = \mathcal{N}(d_i \,|\, f(\vartheta_i), \Sigma(\vartheta_i))$. The probability distributions $p(\vartheta_i|\psi)$ and $p(\psi)$ are modelled and are problem dependent.

The Bayesian inversion problem seeks answers for the following questions: After observing some data for the random variables $Y_1, \ldots, Y_N$, i.e., we know that $\boldsymbol{Y} = \boldsymbol{d} = \{d_1, \ldots, d_N\}$,

1. What is the probability of the parameters ϑ_i and ψ conditioned on the observations $\boldsymbol{d}$, $p(\boldsymbol{\vartheta}, \psi|\boldsymbol{d})$? What is the probability of ϑ^{new} conditioned on the data $\boldsymbol{d}$?

2. What is the probability distribution of a new random variable y^{new} conditioned to the data $\boldsymbol{d}$ and prior to observing any data for y^{new}, $p(y^{\text{new}}|\boldsymbol{d})$?

3. What is the probability distribution of a new random variable y_i^{new}, that depends on ϑ_i, conditioned on the data $\boldsymbol{d}$, $p(y_i^{\text{new}}|\boldsymbol{d})$?

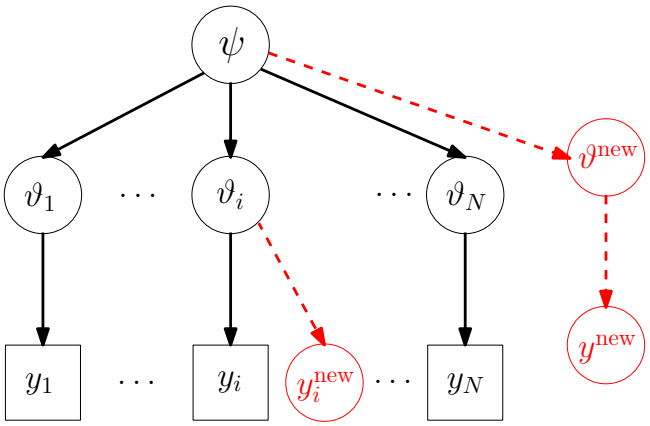

Figure 1: Directed Acyclic Graph (DAG) showing the dependency of the random variables y_i, ϑ_i and ψ. The circles show unobserved and the squares observed random variables. The variables ϑ_i and ϑ^{new} are being inferred after observing the variables y_i. Then, the probability of y_i^{new} and y^{new} conditioned on the observations $\boldsymbol{d}$ can be quantified.

1 Conditional probability of $\boldsymbol{\vartheta}$ and ψ

The joint distribution approach. We ask to draw samples from the distribution $p(\boldsymbol{\vartheta}, \psi|\boldsymbol{d})$. Using Bayes theorem,

$$p(\boldsymbol{\vartheta}, \psi|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\boldsymbol{\vartheta}, \psi)\, p(\boldsymbol{\vartheta}, \psi)}{p(\boldsymbol{d})} = \frac{p(\boldsymbol{d}|\boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta}|\psi)\, p(\psi)}{p(\boldsymbol{d})}\ , \tag{3}$$

where by (1) the likelihood function is simplified as $p(\boldsymbol{d}|\boldsymbol{\vartheta}) = \prod_{i=1}^{N} p(d_i|\vartheta_i)$. The problem with the sampling if this distribution is the dimension of the state space of the target probability which is equal to $NN_\vartheta + N_\psi$. Unless this number is small the sampling of (3) is highly inefficient. This approach was considered in [1] in low dimensional problems. The conditional probability of ϑ_i is given by marginalizing (3),

$$p(\vartheta_i|\boldsymbol{d}) = \int p(\boldsymbol{\vartheta}, \psi|\boldsymbol{d})\, \mathrm{d}\psi\, \mathrm{d}\boldsymbol{\vartheta}_{\setminus i}, \tag{4}$$

where $\boldsymbol{\vartheta}_{\setminus i} = (\vartheta_1, \ldots, \vartheta_{i-1}, \vartheta_{i+1}, \ldots, \vartheta_{N_\vartheta})$ and the conditional distribution of ϑ^{new},

$$\begin{aligned} p(\vartheta^{\mathrm{new}}|\boldsymbol{d}) &= \int p(\vartheta^{\mathrm{new}}, \psi|\boldsymbol{d})\, \mathrm{d}\psi \\ &= \int p(\vartheta^{\mathrm{new}}|\psi, \boldsymbol{d})\, p(\psi|\boldsymbol{d})\, \mathrm{d}\psi = \int p(\vartheta^{\mathrm{new}}|\psi)\, p(\psi|\boldsymbol{d})\, \mathrm{d}\psi \\ &\approx \frac{1}{N_s} \sum_{k=1}^{N_s} p(\vartheta^{\mathrm{new}}|\psi^{(k)}), \quad \psi^{(k)} \sim p(\psi^{(k)}|\boldsymbol{d})\ , \end{aligned} \tag{5}$$

and $p(\psi|\boldsymbol{d})$ is given by marginalizing (3) over $\boldsymbol{\vartheta}$.

The two step approach. In this approach the conditional distribution of ψ is sampled first,

$$p(\psi|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\psi)p(\psi)}{p(\boldsymbol{d})}\ , \tag{6}$$

where

$$\begin{aligned} p(\boldsymbol{d}|\psi) &= \prod_{i=1}^{N} p(d_i|\psi) \\ &= \prod_{i=1}^{N} \int p(d_i|\vartheta_i)\, p(\vartheta_i|\psi)\, \mathrm{d}\vartheta_i \\ &\approx \prod_{i=1}^{N} \frac{1}{N_s} \sum_{k=1}^{N_s} p(d_i|\vartheta_i^{(k)}), \quad \vartheta_i^{(k)} \sim p(\vartheta_i^{(k)}|\psi)\ . \end{aligned} \tag{7}$$

Notice that one evaluation of $p(\boldsymbol{d}|\psi)$ involves NN_s evaluations of the likelihood $p(y|\vartheta)$ and thus NN_s evaluations of the model. In order to alleviate the huge computational cost an importance sampling scheme was proposed in [Wu2016] with $p(\vartheta_i|d_i, \mathcal{M}_i)$ as instrumental density in each of the integrals in (7). Here, $p(\vartheta_i|d_i, \mathcal{M}_i)$ is the posterior distribution of ϑ_i after observing the data d_i and by considering d_i and ϑ_i independent of ψ,

$$p(\vartheta_i|d_i, \mathcal{M}_i) = \frac{p(d_i|\vartheta_i, \mathcal{M}_i)\, p(\vartheta_i|\mathcal{M}_i)}{p(d_i|\mathcal{M}_i)}\ , \tag{8}$$

where the likelihood function $p(d_i|\vartheta_i, \mathcal{M}_i)$ is chosen equal to the likelihood function $p(d_i|\vartheta_i)$. Finally, the likelihood (7) is written as

$$p(\boldsymbol{d}|\psi) = \prod_{i=1}^{N} p(d_i|\psi)\ , \tag{9}$$

with

$$\begin{aligned} p(d_i|\psi) &= \int p(d_i|\vartheta_i)\, p(\vartheta_i|\psi)\, \mathrm{d}\vartheta_i \\ &= \int \frac{p(d_i|\vartheta_i)\, p(\vartheta_i|\psi)}{p(\vartheta_i|d_i, \mathcal{M}_i)}\, p(\vartheta_i|d_i, \mathcal{M}_i)\, \mathrm{d}\vartheta_i \\ &= \int \frac{p(d_i|\vartheta_i)\, p(\vartheta_i|\psi)\, p(d_i|\mathcal{M}_i)}{p(d_i|\vartheta_i, \mathcal{M}_i)\, p(\vartheta_i|\mathcal{M}_i)}\, p(\vartheta_i|d_i, \mathcal{M}_i)\, \mathrm{d}\vartheta_i \\ &= \int \frac{p(\vartheta_i|\psi)\, p(d_i|\mathcal{M}_i)}{p(\vartheta_i|\mathcal{M}_i)}\, p(\vartheta_i|d_i, \mathcal{M}_i)\, \mathrm{d}\vartheta_i\ , \end{aligned} \tag{10}$$

and approximated by

$$p(\boldsymbol{d}|\psi) \approx \prod_{i=1}^{N} \frac{p(d_i|\mathcal{M}_i)}{N_s} \sum_{k=1}^{N_s} \frac{p(\vartheta_i^{(k)}|\psi)}{p(\vartheta_i^{(k)}|\mathcal{M}_i)} , \quad \vartheta_i^{(k)} \sim p(\vartheta_i^{(k)}|d_i, \mathcal{M}_i) . \tag{11}$$

With this scheme the model function is evaluated NN_s times and the $\vartheta_i^{(k)}$ samples are stored. Then, the evaluation of $p(\boldsymbol{d}|\psi)$ involves no evaluations of the likelihood function and thus the computationally expensive model function.

The posterior samples for a new parameter ϑ^{new} can be obtained by,

$$\begin{aligned} p(\vartheta^{\text{new}}|\boldsymbol{d}) &= \int p(\vartheta^{\text{new}}, \psi|\boldsymbol{d}) \, \mathrm{d}\psi \\ &= \int p(\vartheta^{\text{new}}|\boldsymbol{d}, \psi) \, p(\psi|\boldsymbol{d}) \, \mathrm{d}\psi \\ &= \int p(\vartheta^{\text{new}}|\psi) \, p(\psi|\boldsymbol{d}) \, \mathrm{d}\psi \\ &\approx \frac{1}{N_s^{\psi}} \sum_{k=1}^{N_s^{\psi}} p(\vartheta^{\text{new}}|\psi^{(k)}), \quad \psi^{(k)} \sim p(\psi|\boldsymbol{d}) . \end{aligned} \tag{12}$$

Notice that (12) is easy to be sampled once $p(\psi|\boldsymbol{d})$ is sampled, since $p(\vartheta^{\text{new}}|\psi^{(k)})$ is a known function.

The posterior samples for the parameter ϑ_i can be obtained by,

$$\begin{aligned} p(\vartheta_i|\boldsymbol{d}) &= \int p(\vartheta_i, \psi|\boldsymbol{d}) \, \mathrm{d}\psi \\ &= \int p(\vartheta_i|\boldsymbol{d}, \psi) \, p(\psi|\boldsymbol{d}) \, \mathrm{d}\psi \\ &= \int p(\vartheta_i|d_i, \psi) \, p(\psi|\boldsymbol{d}) \, \mathrm{d}\psi \\ &= p(d_i|\vartheta_i) \int \frac{p(\vartheta_i|\psi)}{p(d_i|\psi)} \, p(\psi|\boldsymbol{d}) \, \mathrm{d}\psi \\ &\approx \frac{p(d_i|\vartheta_i)}{N_s^{\psi}} \sum_{k=1}^{N_s^{\psi}} \frac{p(\vartheta_i|\psi^{(k)})}{p(d_i|\psi^{(k)})}, \quad \psi^{(k)} \sim p(\psi|\boldsymbol{d}) . \end{aligned} \tag{13}$$

Notice that that the term $p(d_i|\psi^{(k)})$ can be precomputed using (10) since it is independent of ϑ_i,

$$\begin{aligned} p(d_i|\psi^{(k)}) &\approx p(d_i|\mathcal{M}_i) \, \frac{1}{N_s^{\vartheta}} \sum_{\ell=1}^{N_s^{\vartheta}} \frac{p(\vartheta_i^{(\ell)}|\psi^{(k)})}{p(\vartheta_i^{(\ell)}|\mathcal{M}_i)} \\ &:= p(d_i|\mathcal{M}_i) \, b_{i,k} . \end{aligned} \tag{14}$$

Finally, the log-posterior is given by,

$$\log p(\vartheta_i|\boldsymbol{d}) \approx \log p(d_i|\vartheta_i) - \log p(d_i|\mathcal{M}_i) + \log \frac{1}{N_s^{\psi}} \sum_{k=1}^{N_s^{\psi}} \frac{p(\vartheta_i|\psi^{(k)})}{b_{i,k}} . \tag{15}$$

2 Propagation of Uncertainty in y

The conditional distribution of the output of the model on the observations $\boldsymbol{d}$ is given by,

$$\begin{aligned} p(y^{\text{new}}|\boldsymbol{d}) &= \int p(y^{\text{new}}|\vartheta^{\text{new}}, \boldsymbol{d}) \, \mathrm{d}\vartheta^{\text{new}} \\ &= \int p(y^{\text{new}}|\vartheta^{\text{new}}) \, p(\vartheta^{\text{new}}|\boldsymbol{d}) \, \mathrm{d}\vartheta^{\text{new}} \\ &\approx \frac{1}{N_s} \sum_{k=1}^{N_s} p(y^{\text{new}}|\vartheta^{\text{new}(k)}), \quad \vartheta^{\text{new}(k)} \sim p(\vartheta^{\text{new}(k)}|\boldsymbol{d}) , \end{aligned} \tag{16}$$

and $p(\vartheta^{\text{new}(k)}|\boldsymbol{d})$ can be sampled by (12). Under the assumption that $p(y|\vartheta) = \mathcal{N}(y|f(\vartheta), \Sigma(\vartheta))$ the distribution $p(y^{\text{new}}|\boldsymbol{d})$ is a Gaussian mixture given by,

$$p(y^{\text{new}}|\boldsymbol{d}) = \sum_{k=1}^{N_s} \frac{1}{N_s} \mathcal{N}\big(y^{\text{new}}|f(\vartheta^{(k)}), \Sigma(\vartheta^{(k)})\big) \; . \tag{17}$$

The posterior distribution of a new observation on the i-th branch of the DAG in Figure 1, conditioned on the observations $\boldsymbol{d}$, is given by

$$\begin{aligned} p(y_i^{\text{new}}|\boldsymbol{d}) &= \int p(y_i^{\text{new}}, \vartheta_i|\boldsymbol{d}) \, \mathrm{d}\vartheta_i \\ &= \int p(y_i^{\text{new}}|\vartheta_i) \, p(\vartheta_i|\boldsymbol{d}) \, \mathrm{d}\vartheta_i \\ &\approx \frac{1}{N_s} \sum_{k=1}^{N_s} p(y_i^{\text{new}}|\vartheta_i^{(k)}), \quad \vartheta_i^{(k)} \sim p(\vartheta_i^{(k)}|\boldsymbol{d}) \; , \end{aligned} \tag{18}$$

where ϑ_i can be sampled by (15). In the special case that $p(y|\vartheta) = \mathcal{N}(y|f(\vartheta), \Sigma(\vartheta))$ the distribution $p(y^{\text{new}}|\boldsymbol{d})$ is a Gaussian mixture similar to (17).

3 Optimization in Bayesian Networks

In Mixed Effects Models (MEM) the likelihood of the data conditioned on the hyper-parametes (6) is optimized on ψ,

$$\psi^\star = \arg\max_{\psi} \; p(\boldsymbol{d}|\psi) \; . \tag{19}$$

Then, the probability distributions of the parameters ϑ_i conditioned on the data are given by (15) for $p(\psi|\boldsymbol{d}) = \delta(\psi - \psi^\star)$,

$$p(\vartheta_i|\boldsymbol{d}) = \frac{p(d_i|\vartheta_i) \, p(\vartheta_i|\psi^\star)}{p(d_i|\psi^\star)} \; . \tag{20}$$

Finally, the posterior distributions for the predictions of y^{new} and y_i^{new} are given by (16) and (18), respectively.

3.1 Algorithms for Optimization in Bayesian Networks

Expectation Maximization. It can be shown that the following iterative procedure maximizes the likelihood function (19) starting from $\psi^{(0)}$:

- Expectation step: Evaluate the function defined by,

$$\begin{aligned} Q(\psi|\psi^{(k)}) &= \mathbb{E}_{\boldsymbol{\vartheta}|\boldsymbol{d},\psi^{(k)}} \left[\log p(\boldsymbol{d}, \boldsymbol{\vartheta}|\psi)\right] \\ &= \int \log\Big(p(\boldsymbol{d}, \boldsymbol{\vartheta}|\psi)\Big) \, p(\boldsymbol{\vartheta}|\boldsymbol{d}, \psi^{(k)}) \, \mathrm{d}\boldsymbol{\vartheta} \\ &= \int \log\Big(p(\boldsymbol{d}|\boldsymbol{\vartheta}) \, p(\boldsymbol{\vartheta}|\psi)\Big) \, p(\boldsymbol{\vartheta}|\boldsymbol{d}, \psi^{(k)}) \, \mathrm{d}\boldsymbol{\vartheta} \\ &= \prod_{i=1}^{N} \int \log\Big(p(d_i|\vartheta_i) \, p(\vartheta_i|\psi)\Big) \, p(\vartheta_i|d_i, \psi^{(k)}) \, \mathrm{d}\vartheta_i \\ &\approx \prod_{i=1}^{N} \frac{1}{N_s} \sum_{\ell=1}^{N_s} \log p(d_i|\vartheta_i^{(\ell)}) + \log p(\vartheta_i^{(\ell)}|\psi), \quad \vartheta_i^{(\ell)} \sim p(\vartheta_i^{(\ell)}|d_i, \psi^{(k)}) \; , \end{aligned} \tag{21}$$

  where $p(\vartheta_i^{(\ell)}|d_i, \psi^{(k)})$ can be sampled as in (20).

- Maximization step: Maximize the function in (21),

$$\psi^{(k+1)} = \arg\max_{\psi} Q(\psi|\psi^{(k)}) \; . \tag{22}$$

Importance Sampling. The likelihood function in (19) can be approximated by the importance sampling scheme described in Section 1 and summarized in equation (11).

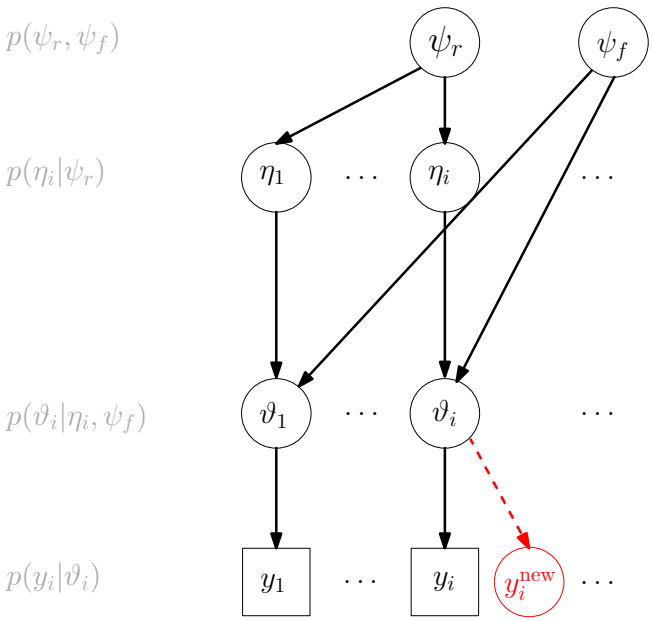

Figure 2: Directed Acyclic Graph (DAG) showing the dependency of the random variables in a two layer hierarchical model.

4 A two layer hierarchical model

Working in the same way as in (9) and (10)

$$p(\boldsymbol{d}|\psi_r, \psi_f) = \prod_{i=1}^{N} \int p(d_i|\vartheta_i)\, p(\vartheta_i|\psi_r, \psi_f)\ \mathrm{d}\vartheta_i\ , \tag{23}$$

where

$$\begin{aligned} p(\vartheta_i|\psi_r, \psi_f) &= \int p(\vartheta_i|\eta_i, \psi_f)\, p(\eta_i|\psi_r)\, \mathrm{d}\eta_i \\ &\approx \frac{1}{N_s} \sum_{k=1}^{N_s} p(\vartheta_i|\eta_i^{(k)}, \psi_f), \quad \eta_i^{(k)} \sim p(\eta_i|\psi_r)\ . \end{aligned} \tag{24}$$

Let $\psi = (\psi_r, \psi_f)$.
Similarly with eq. (15), the posterior distribution of parameters ϑ_i is given by,

$$\begin{aligned} p(\vartheta_i|\boldsymbol{d}) &= \int p(\vartheta_i|\psi, \boldsymbol{d}) p(\psi|\boldsymbol{d})\, \mathrm{d}\psi \\ &= p(d_i|\vartheta_i) \int \frac{p(\vartheta_i|\psi)}{p(d_i|\psi)} p(\psi|\boldsymbol{d})\, \mathrm{d}\psi \\ &\approx p(d_i|\vartheta_i)\, \frac{1}{N_s} \sum_{k=1}^{N_s} \frac{p(\vartheta_i|\psi^{(k)})}{p(d_i|\psi^{(k)})}, \quad \psi^{(k)} \sim p(\psi|\boldsymbol{d}) \\ &\approx p(d_i|\vartheta_i)\, \frac{1}{N_s} \sum_{k=1}^{N_s} \frac{1}{p(d_i|\psi^{(k)})}\, \frac{1}{N_s} \sum_{\ell=1}^{N_s} p(\vartheta_i|\eta_i^{(\ell)}\psi_f^{(k)}), \quad \eta_i^{(\ell)} \sim p(\eta_i|\psi_r^{(k)}), \quad \psi^{(k)} \sim p(\psi|\boldsymbol{d})\ . \end{aligned} \tag{25}$$

The joint posterior distribution of parameter (η_i, ϑ_i) is given by,

$$\begin{aligned} p(\eta_i, \vartheta_i|\boldsymbol{d}) &= \int p(\eta_i, \vartheta_i|\psi, \boldsymbol{d})\, p(\psi|\boldsymbol{d})\, \mathrm{d}\psi \\ &= \int p(\eta_i, \vartheta_i|\psi, d_i)\, p(\psi|\boldsymbol{d})\, \mathrm{d}\psi \\ &= \int \frac{p(d_i|\eta_i, \vartheta_i, \psi)\, p(\eta_i, \vartheta_i|\psi)}{p(d_i|\psi)}\, p(\psi|\boldsymbol{d})\, \mathrm{d}\psi \\ &= \int \frac{p(d_i|\vartheta_i)\, p(\eta_i, \vartheta_i|\psi)}{p(d_i|\psi)}\, p(\psi|\boldsymbol{d})\, \mathrm{d}\psi\ . \end{aligned} \tag{26}$$

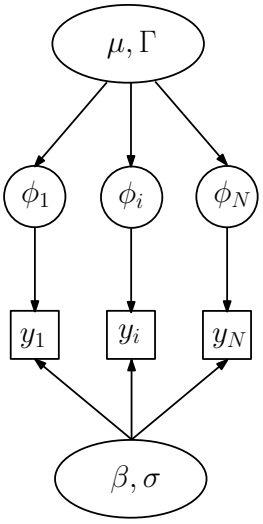

Figure 3: Directed Acyclic Graph (DAG) showing the dependency of the random variables in a two layer hierarchical model.

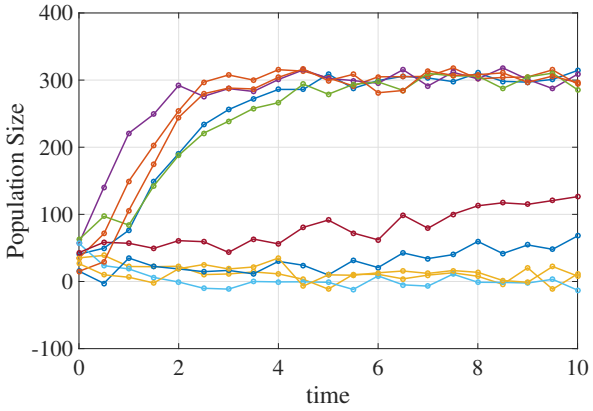

Figure 4: Independent synthetic data sets from equation (29) using the parameters in (31) .

By noticing that

$$p(\eta_i, \vartheta_i|\psi) = p(\vartheta_i|\eta_i, \psi_f)\, p(\eta_i|\psi_r)\ , \tag{27}$$

the posterior distribution (26) can be written as,

$$p(\eta_i, \vartheta_i|\boldsymbol{d}) = p(d_i|\vartheta_i) \int \frac{p(\vartheta_i|\eta_i, \psi_f)\, p(\eta_i|\psi_r)}{p(d_i|\psi)}\, p(\psi|\boldsymbol{d})\, \mathrm{d}\psi\ . \tag{28}$$

5 Mixed Effects Model

6 Numerical example

We consider the following function that models the growth of population

$$f(t;\vartheta) = K\, P_0\, \frac{e^{\lambda\, t}}{K + P_0(e^{\lambda\, t} - 1)}\ , \tag{29}$$

where $\vartheta = (K, P_0, \lambda)$ and K is the capacity, P_0 is the initial population and λ the growth rate. We create synthetic data at times $t_k = 0.5k$, for $k = 1, \ldots, 20$,

$$d_k = f(t_k;\vartheta) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_d^2)\ , \tag{30}$$

using the following nominal values for the parameters,

$$K = 300, \quad P_0 \sim \mathcal{N}(40, 10^2), \quad \lambda \sim \mathcal{N}(0.5, 1), \quad \sigma_d = 10\ , \tag{31}$$

where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and standard deviation σ. See Figure 4 for an example of 10 data sets. From now on we consider the extended parameter vector $\vartheta = (K, P_0, \lambda, \sigma_d)$ that contains the standard deviation of the noise.

References

[1] J. B. Nagel and B. Sudret. Hamiltonian Monte Carlo and borrowing strength in hierarchical inverse problems. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 0(0):B4015008, 2015.

[2] J. Stirnemann, A. Samson, and J.-C. Thalabard. Individual predictions based on nonlinear mixed modeling: application to prenatal twin growth. *Statistics in Medicine*, 31(18):1986–1999, 2012.