Computational tools III: Variational approximations

Sorry, new notation !

For a joint distribution $p(\mathbf{x}, \mathbf{y})$ of hidden variables $\mathbf{x}$ and observed data $\mathbf{y}$ the posterior

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})}$$

describes our knowledge about $\mathbf{x}$ when we observe $\mathbf{y}$.

- The computation of the marginal probability of the data $p(\mathbf{y}) = \int d\mathbf{x}\, p(\mathbf{x}, \mathbf{y})$ requires high dimensional sums or integrals and is often intractable.

- For the same reasons we often can't compute marginals $p_i(x_i|\mathbf{y})$, or expectations using these densities which are e.g. required in the EM algorithm.

The Variational Approximation

Approximate $p(\mathbf{x})$ by $q(\mathbf{x}) \in \mathcal{F}$ where $\mathcal{F}$ tractable family of distributions such that the Kullback-Leibler divergence

$$KL(q,p) = \int d\mathbf{x} q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \geq 0$$

is minimized.

- Setting $p(\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{y})}{Z}$ with $Z = p(\mathbf{y})$, we get an **upper bound** for any q

  $$-\ln Z = F(q) - KL(q,p) \leq F(q) \doteq \int d\mathbf{x}\, q(\mathbf{x}) \ln q(\mathbf{x}) - \langle \ln p(\mathbf{x},\mathbf{y}) \rangle_q$$

  with the **variational free energy** $F(q)$

- Dependency on parameters for optimal q:

$$\frac{dF(q^*(\theta),\theta)}{d\theta} = \frac{\partial F(q^*,\theta)}{\partial \theta}$$

The Mean Field Method

An important case is given by the family of factorising densities

$$q(\mathbf{x}) = \prod_{i=1}^{M} q_i(x_i)$$

In this case, we speak of a **mean field approximation**. Optimise q_i such that the free energy

$$F(q) = \int d\mathbf{x}\, q(\mathbf{x}) \ln q(\mathbf{x}) - \langle \ln p(\mathbf{x}, \mathbf{y}) \rangle_q$$

is minimial. The solution is: $q_i^*(x) = \frac{1}{Z_i} \exp \langle \ln p(\mathbf{x}, \mathbf{y}) \rangle_{\setminus i}$ with $\langle \ldots \rangle_{\setminus i}$ the average over all variables except x_i.

Proof: For any q_i, we have

$F(q) = -\int dx\, q_i(x) \langle \ln p(\mathbf{x}, \mathbf{y}) \rangle_{\setminus i} + \sum_j \int dx\, q_j(x)\ \ln q_j(x)$

$= KL(q_i, q_i^*) - \ln Z_i^* + \sum_{j, j \neq i} \sum_x q_j(x)\ \ln q_j(x)$. Minimal for $q_i = q_i^*$. Requires *selfconsistent solution* (e.g. sequential update).

MF Example

$$p(\mathbf{x}) = \frac{1}{Z} \prod_i \psi_i(x_i) \; \text{exp}\left[\frac{1}{2}\mathbf{x}^T \mathbf{J} \mathbf{x}\right]$$

with $J_{ii} = 0$. For this case, we have

$$q_i(x) = \frac{1}{Z_i} \psi_i(x) \, \text{exp}\left[x \underbrace{\sum_j J_{ij} \langle x_j \rangle_q}_{\gamma_i} \right]$$

Introduce

$$Z_i(\gamma) = \int dx \; \psi_i(x) \, \text{exp}\,[x\gamma]$$

$$m_i(\gamma) = \frac{d \ln Z_i}{d\gamma}$$

we get the relation (exact for Gaussian models)

$$\langle x_i \rangle_q = m_i \left(\sum_j J_{ij} \langle x_j \rangle_q \right)$$

Variational EM Algorithm

Optimise model parameters by Maximum Likelihood using free energy bound

$$-\ln p(\mathbf{y}|\boldsymbol{\theta}) \leq \int q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{y},\mathbf{x}|\boldsymbol{\theta})} \equiv F(q,\boldsymbol{\theta})$$

Iterate:

1. Mimimise $F(q,\boldsymbol{\theta}_t)$ with respect to the distribution $q \in \mathcal{F} \to q_t$. Note, that the unconstrained variation gives $q_t(\mathbf{x}) = p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ (exact EM algorithm)!

2. Minimise $F(q_t,\theta)$ with respect to $\boldsymbol{\theta}$.

   This iterations will not increase (and possibly decrease) **an upper bound** on $-\ln p(\mathbf{y}|\boldsymbol{\theta})$!

Variational Bayes algorithm

This aims at performing an approximation to a full Bayesian posterior i.e. $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$. We use the bound

$$-\ln p(\mathbf{y}|m) \leq F(q) = \int d\mathbf{x}\, d\boldsymbol{\theta}\; q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{q(\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}|m)}$$

Look for minima in the space of factorising distributions $q(\mathbf{x}, \boldsymbol{\theta}) = q(\mathbf{x})q(\boldsymbol{\theta})$.

Alternate between

1. **VB - E Step**: Minimise $F(q(\mathbf{x}), q_t(\boldsymbol{\theta}))$ w.r.t. $q(\mathbf{x})$

$$q_{l+1}(\mathbf{x}) \propto \exp\left[\int q_l(\boldsymbol{\theta})) \ln p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m)\; d\boldsymbol{\theta}\right]$$

2. **VB - M Step**: Minimise $F(q_{l+1}(\mathbf{x}), q(\boldsymbol{\theta}))$ w.r.t. $q(\boldsymbol{\theta})$

$$q_{l+1}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \exp\left[\int q_l(\mathbf{x})) \ln p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m)\; d\mathbf{x}\right]$$

Dynamical Bayes Models with hidden factors

(M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, D. L. Wild)

- Hidden causes or unmeasured genes may simplify network structure & lead to better interpretability.

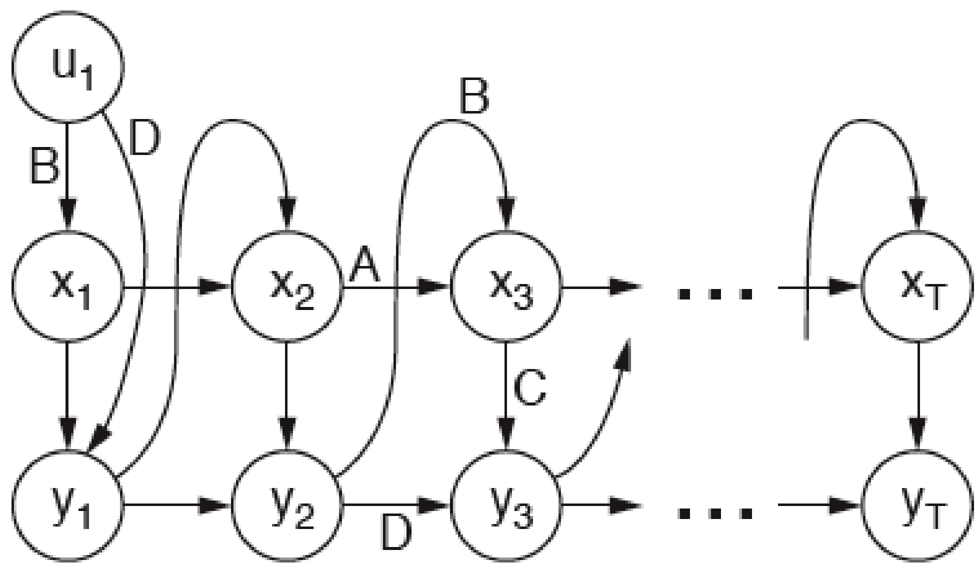


Fig. 1. The feedback graphical model with outputs feeding into inputs. Gene expression levels at time t are represented by y_t, whilst the hidden factors are represented by x_t.

- Gaussian state space models

$$\begin{aligned}\mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{y}_{t-1} + \mathbf{w}_t \qquad \mathbf{w}_t \sim N(0, \mathbf{I}) \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{y}_{t-1} + \mathbf{v}_t \qquad \mathbf{w}_t \sim N(0, \mathbf{R})\end{aligned}$$

- Bayesian approach: Use (conjugate) Gaussian prior distributions over matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ and a Gamma prior over the elements of the diagonal matrix $\mathbf{R}$.

- Goal: Fit the model by maximising $p(\mathbf{y}|m)$ where m denotes the model, i.e. the dimensionality of the hidden states.

- Make predictions about *interactions* using the posterior distribution $p(\theta|\mathbf{y}, m)$ where $\theta = \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$.

  Problem: This is intractable! Approximate inference is necessary.

We have

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m) = \prod_t p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{t-1}) \times \prod_t p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$$

with

$$p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{t-1}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}_t - \mathbf{C}\mathbf{x}_t - \mathbf{D}\mathbf{y}_{t-1})^\top \mathbf{R}^{-1}(\mathbf{y}_t - \mathbf{C}\mathbf{x}_t - \mathbf{D}\mathbf{y}_{t-1})\right]$$

and

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \propto \exp\left[-\frac{1}{2}(\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \mathbf{B}\mathbf{y}_{t-1})^\top (\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \mathbf{B}\mathbf{y}_{t-1})\right]$$

- Hidden variables possibly represent ''combination of complex molecular events linking two genes''

- This leads to effective interactions (activation or inhibition) between measured genes is given by $I_{ij} = (\mathbf{CB} + \mathbf{D})_{ij}$.

- Significant evidence of interactions if $|I_{ij}|$ far away from 0 relative to standard deviation.

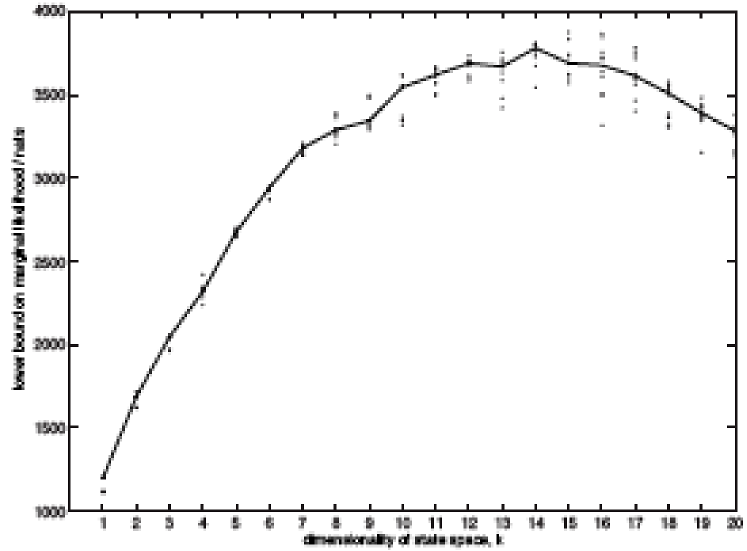


Fig. 2. Variation of $\mathcal{F}$ with hidden state dimension k for 10 random initializations of VBEM. The line represents the median $\mathcal{F}$ value.

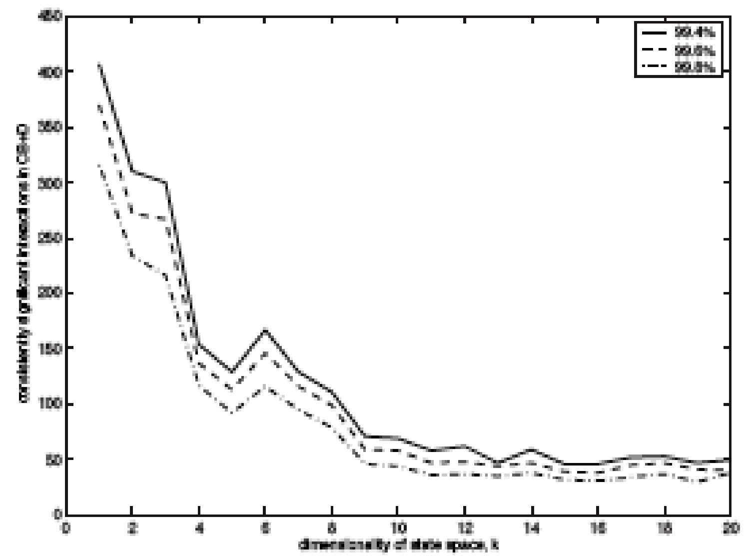


Fig. 3. The number of significant interactions that are repeated in all 10 runs of VB-EM at each value of k. There are 3 plots, each corresponding to a different significance level.

Linear Response Correction

We can get an explicit improvement on MF by estimating the neglected correlations using the following 'trick': Introduce

$$p(\mathbf{x}|\mathbf{y},\mathbf{h}) = \frac{1}{Z(\mathbf{h})} \underbrace{p(\mathbf{x},\mathbf{y})e^{\mathbf{h}^T\mathbf{x}}}_{p(\mathbf{x},\mathbf{y}|\mathbf{h})}$$

and $F(\mathbf{h}) = -\ln Z(\mathbf{h}) = -\ln \int d\mathbf{x}\, p(\mathbf{x},\mathbf{y})e^{\mathbf{h}^T\mathbf{x}}$.

Then

$$-\frac{\partial F}{\partial h_i} = \langle x_i \rangle \qquad -\frac{\partial \langle x_i \rangle}{\partial h_j} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

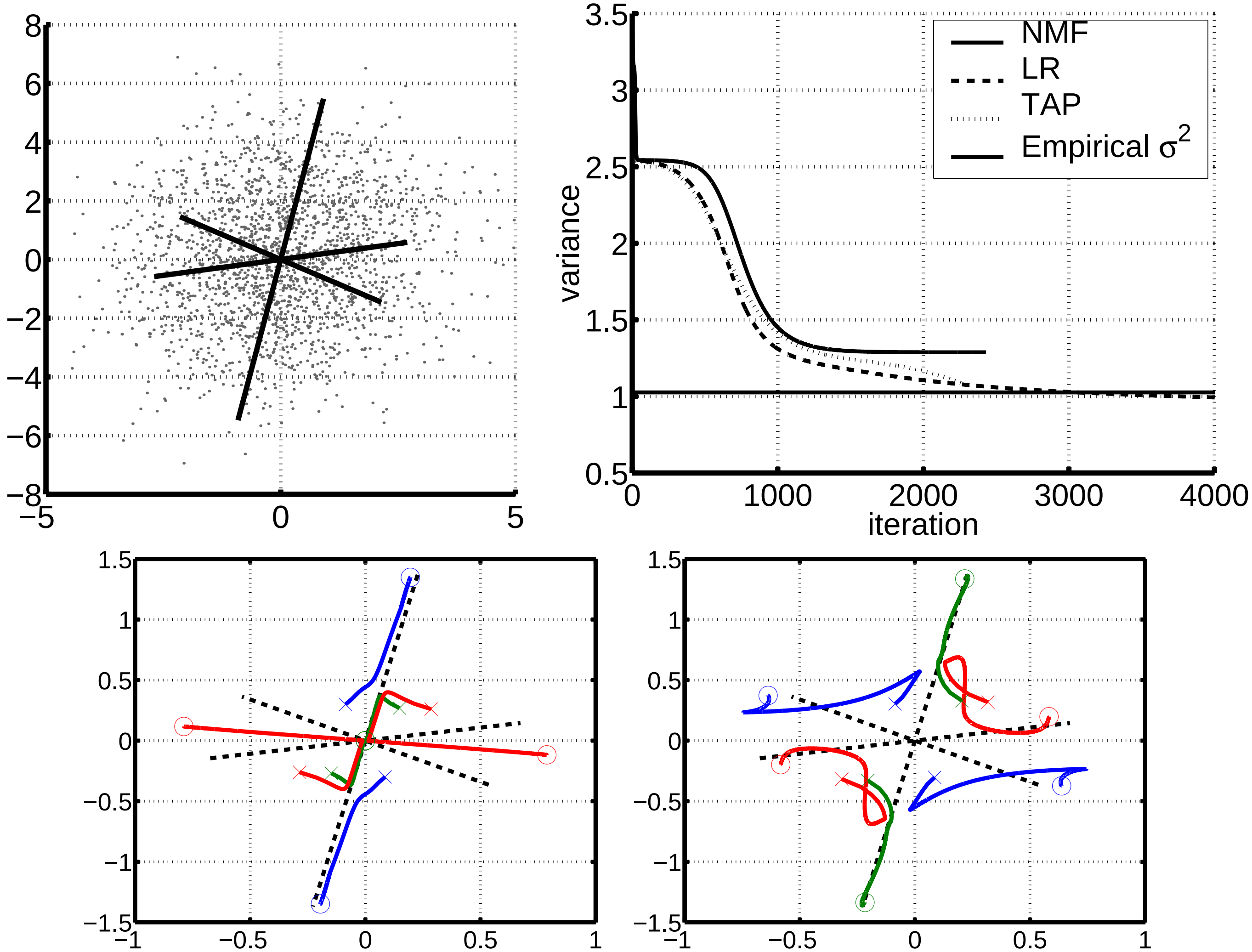
Evaluate in Mean Field approximation:

$$C_{ij} \doteq \frac{\partial \langle x_i \rangle_{q(\mathbf{h})}}{\partial h_j} \approx \frac{\partial m_i(\gamma_i(\mathbf{h}) + h_i)}{\partial h_j}$$

leads to the approximate covariance (exact for Gaussian models)

$$\mathbf{C} = (\mathbf{\Lambda} - \mathbf{J})^{-1} \qquad \text{where} \qquad \mathbf{\Lambda} = \mathrm{diag}\left\{1/\left(\langle x_i^2 \rangle_q - \langle x_i \rangle_q^2\right)\right\}$$

ICA (artificial Data) 2 Sensors, 3 Sources



$\rho(S)$ bimodal. **left**: MF **right:** MF + linear response. (Højen-Sørensen, Winther & Hansen)

LR: Gaussian Models

$$\ln p(\mathbf{x}) = \frac{1}{2} \sum_{ij, i \neq j} x_i J_{ij} x_j + \sum_i (h_i x_i - b_i x_i^2/2)$$

Variational distribution (with $J_{ii} = -b_i$):

$$q_i(x_i) \propto \exp[-\frac{b_i}{2}(x_i - [\mathbf{J}^{-1}\mathbf{h}]_i)^2]$$

LR approximation

$$\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle = -(\mathbf{J}^{-1})_{ij}$$

comes out exact!

Gauss-Variational method (C. Archambeau & M. Opper)

Let $\mathbf{y}$ be observations and $\mathbf{x}$ latent parameters. Approximate posterior

$$p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}) = \frac{p(\mathbf{y},\mathbf{x},\boldsymbol{\theta})}{p(\mathbf{y},\boldsymbol{\theta})},$$

by a **tractable density** $q(\mathbf{x})$ minimising the **variational free energy**

$$F(q,\boldsymbol{\theta}) = -H[q] - \langle \log p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}) \rangle_q$$

Gaussian variational densities

$$q(\mathbf{x}) = (2\pi)^{-N/2}|\mathbf{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

The variational free energy becomes

$$F(q,\boldsymbol{\theta}) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{N}{2} - \langle \log p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta})\rangle_q$$

Setting $\nabla\mathcal{F}(q,\boldsymbol{\theta}) = 0$, we obtain

$$\begin{aligned} 0 &= \nabla_{\boldsymbol{\mu}}\langle \log p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta})\rangle_q = \left\langle \frac{\partial \log p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta})}{\partial \mathbf{x}} \right\rangle_q \\ \mathbf{\Sigma}^{-1} &= -2\nabla_{\mathbf{\Sigma}}\langle \log p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta})\rangle_q = -\left\langle \frac{\partial^2 \log p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta})}{\partial \mathbf{x}^T\partial \mathbf{x}} \right\rangle_q \end{aligned}$$

Useful Results for Gaussian expectations

To compute the minimum, we need

$$\frac{\partial \ln|\mathbf{\Sigma}|}{\partial \mathbf{\Sigma}} = -2\frac{\partial \ln \int d\mathbf{x} \exp\left[-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}\mathbf{x}\right]}{\partial \mathbf{\Sigma}} = \langle \mathbf{x}\mathbf{x}^T \rangle = \mathbf{\Sigma}^{-1}$$

Introducing the characteristic function

$$G(\mathbf{k}) = E_q\left[e^{i\mathbf{k}^T\mathbf{x}}\right] = \exp\left[-\frac{1}{2}\mathbf{k}^T\mathbf{\Sigma}\mathbf{k} + i\mathbf{k}^T\mathbf{m}\right]$$

of the measure q

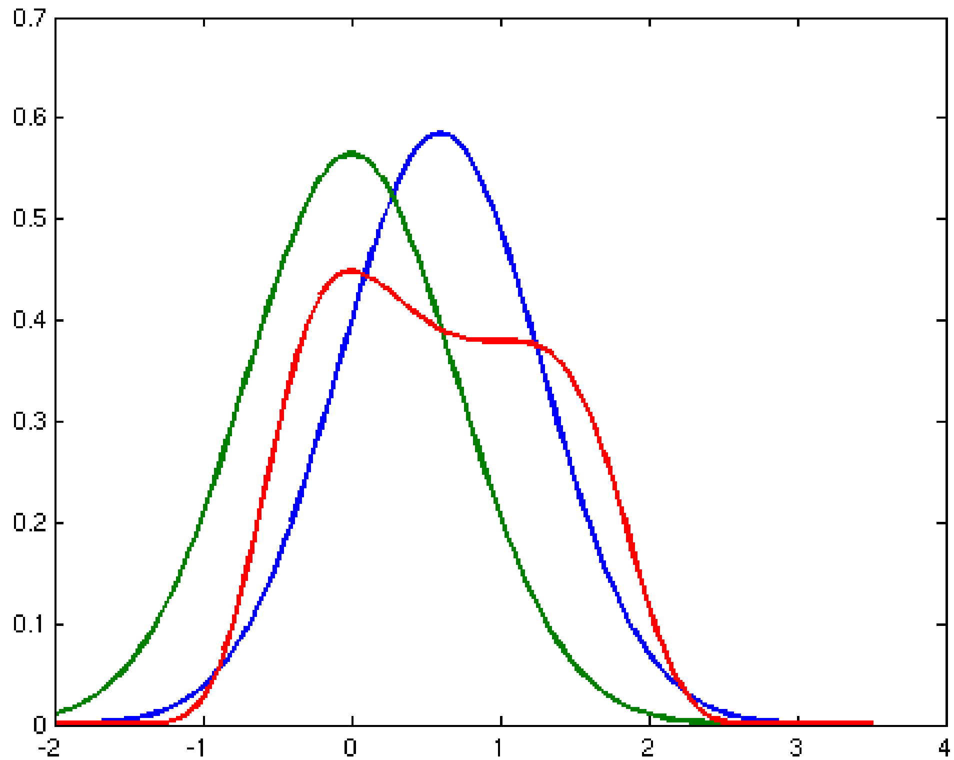
$$\begin{aligned}\int d\mathbf{x}\; q(\mathbf{x})\; F(\mathbf{x}) = \int d\mathbf{y}\; E_q\left[\delta(\mathbf{x}-\mathbf{y})\right] F(\mathbf{y}) &= \frac{1}{(2\pi)^n} \int d\mathbf{y}\; d\mathbf{k}\; G(\mathbf{k}) e^{-i\mathbf{k}^T\mathbf{y}} F(\mathbf{y}) \\ &= \frac{1}{(2\pi)^n} \int d\mathbf{y}\; d\mathbf{k}\; \exp\left[-\frac{1}{2}\mathbf{k}^T\mathbf{\Sigma}\mathbf{k} + i\mathbf{k}^T(\mathbf{m}-\mathbf{y})\right] F(\mathbf{y})\end{aligned}$$

Thus

$$\frac{\partial E_q[F(\mathbf{x})]}{\partial \mathbf{m}} = E_q\left[\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}\right]$$

and

$$\begin{aligned}\frac{\partial E_q[F(\mathbf{x})]}{\partial \boldsymbol{\Sigma}} &= -\frac{1}{2}\int d\mathbf{y}\ d\mathbf{k}\ \exp\left[-\frac{1}{2}\mathbf{k}^T\boldsymbol{\Sigma}\mathbf{k} + i\mathbf{k}^T(\mathbf{m}-\mathbf{y})\right]\mathbf{k}\mathbf{k}^T F(\mathbf{y}) \\ &= \frac{1}{2}E_q\left[\frac{\partial^2 F(\mathbf{x})}{\partial \mathbf{x}^T \partial \mathbf{x}}\right] = \frac{1}{2}\frac{\partial^2 E_q\left[F(\mathbf{x})\right]}{\partial \mathbf{m}^T \partial \mathbf{m}}\end{aligned}$$


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-2
-1
0
1
2
3
4

GPs with factorising likelihood

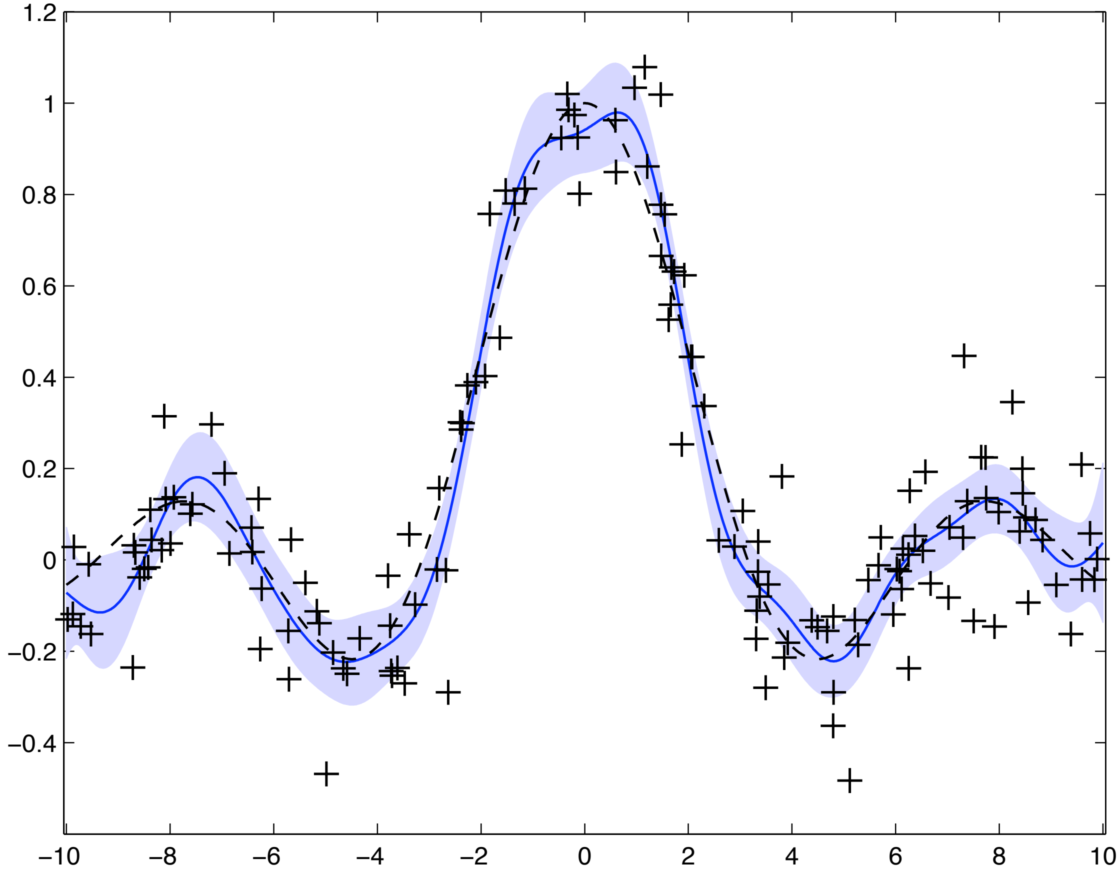
$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z_0} \exp\left(-\sum_n V_n(y_n, x_n) - \frac{1}{2}\mathbf{x}^T \mathbf{K}^{-1}\mathbf{x}\right),$$

Covariance

$$\mathbf{\Sigma}^{-1} = \mathbf{K}^{-1} + \text{diag}\left\langle \frac{\partial^2 V_n}{\partial x_n^2} \right\rangle_q$$

is parametrised by N elements!

sinc function with Cauchy noise (GP with Gaussian likelihood)



sinc function with Cauchy noise (Var - GP with Cauchy likelihood)