Gaussian Process Regression Networks

Andrew Gordon Wilson

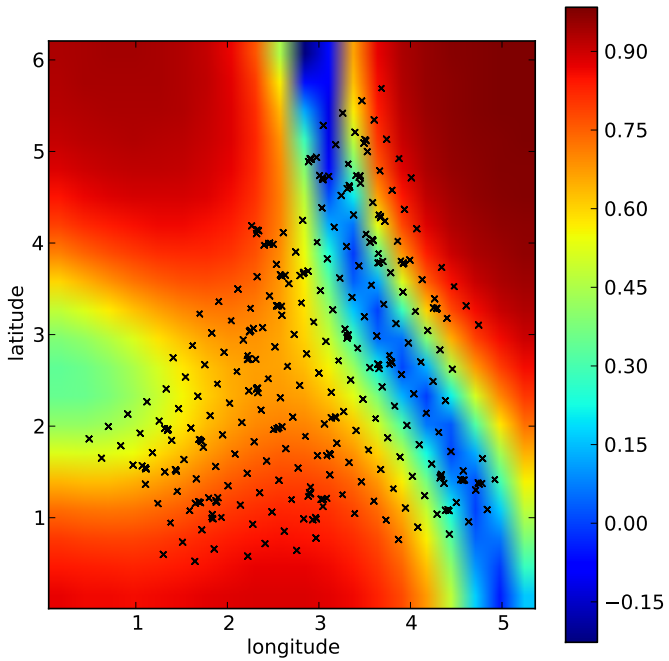
agw38@cam.ac.uk
mlg.eng.cam.ac.uk/andrew
University of Cambridge

Joint work with David A. Knowles and Zoubin Ghahramani

June 27, 2012
ICML, Edinburgh

Multiple responses with input dependent covariances

- Two response variables:
  $y_1(x)$: concentration of cadmium at a spatial location x.
  $y_2(x)$: concentration of zinc at a spatial location x.
- The values of these responses, at a given spatial location x_*, are correlated.
- We can account for these correlations (rather than assuming $y_1(x)$ and $y_2(x)$ are independent) to enhance predictions.
- We can further enhance predictions by accounting for how these correlations vary with geographical location x. Accounting for input dependent correlations is a distinctive feature of the Gaussian process regression network.



Motivation for modelling dependent covariances

Promise

- Many problems in fact have input dependent uncertainties and correlations.
- Accounting for dependent covariances (uncertainties and correlations) can greatly improve statistical inferences.

Uncharted Territory

- For convenience, response variables are typically seen as independent, or as having fixed covariances (e.g. multi-task literature).
- The few existing models of dependent covariances are typically not expressive (e.g. Brownian motion covariance structure) or scalable (e.g. < 5 response variables).

Goal

- We want to develop expressive and scalable models (> 1000 response variables) for dependent uncertainties and correlations.

Outline

- Gaussian process review
- Gaussian process regression networks
- Applications

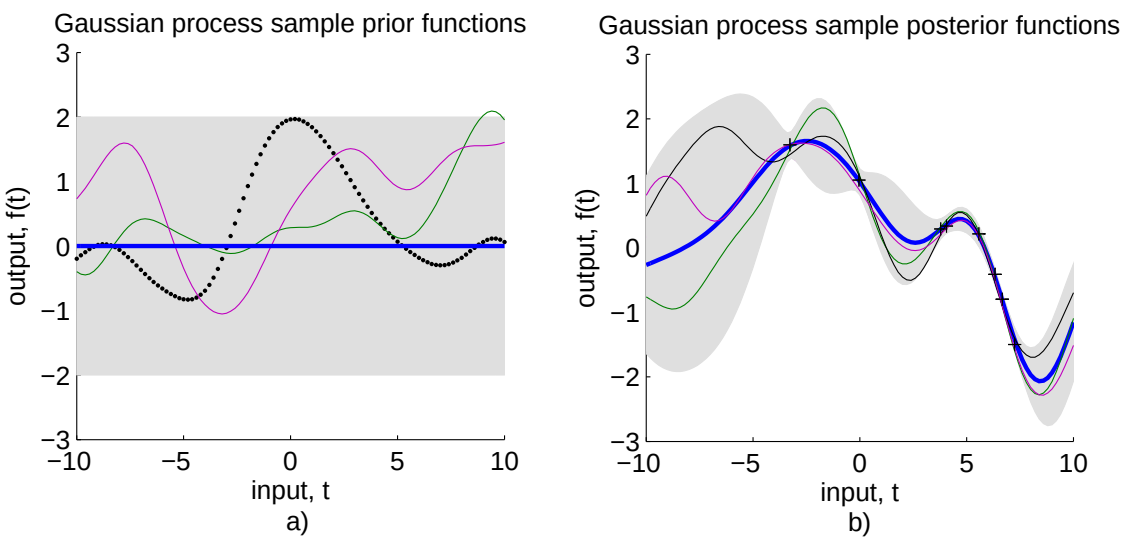
Definition

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Nonparametric Regression Model

- Prior: $f(x) \sim \mathcal{GP}(m(x), k(x,x'))$, meaning $(f(x_1), \ldots, f(x_N)) \sim \mathcal{N}(\boldsymbol{\mu}, K)$, with $\boldsymbol{\mu}_i = m(x_i)$ and $K_{ij} = \mathrm{cov}(f(x_i), f(x_j)) = k(x_i, x_j)$.

$$\overbrace{p(f(x)|\mathcal{D})}^{\text{GP posterior}} \propto \overbrace{p(\mathcal{D}|f(x))}^{\text{Likelihood}} \overbrace{p(f(x))}^{\text{GP prior}}$$



"How can Gaussian processes possibly replace neural networks? Did we throw the baby out with the bathwater?"

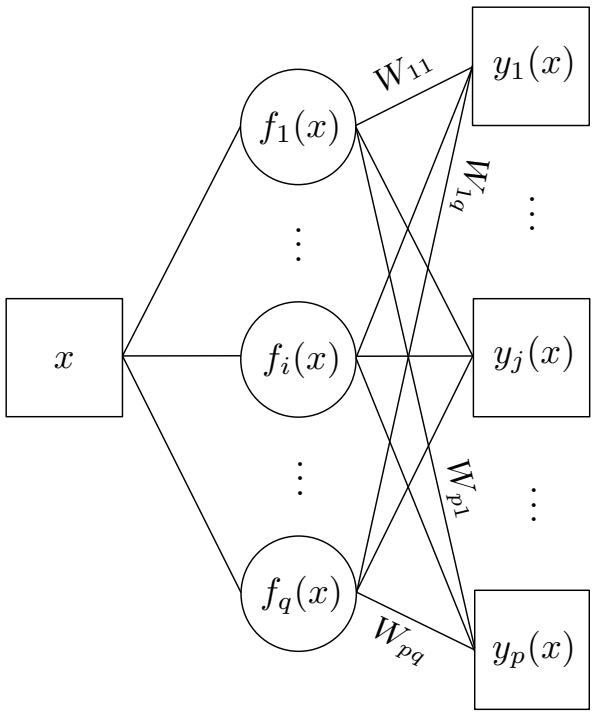
David MacKay, 1998.

Semiparametric Latent Factor Model

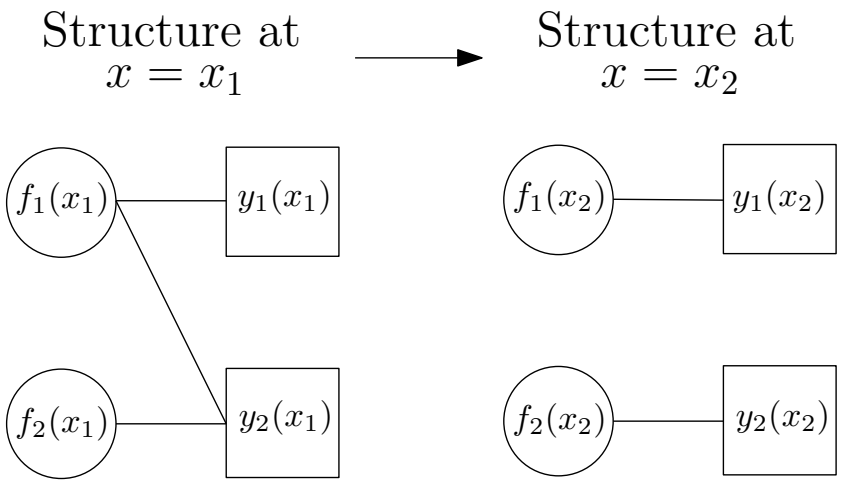
The semiparametric latent factor model (SLFM) (Teh, Seeger, and Jordan, 2005) is a popular multi-output (multi-task) GP model for fixed *signal* correlations between outputs (response variables):

$$\overbrace{\boldsymbol{y}(x)}^{p\times 1} = \overbrace{W}^{p\times q}\overbrace{\boldsymbol{f}(x)}^{q\times 1} + \sigma_y \overbrace{\boldsymbol{z}(x)}^{\mathcal{N}(0,I_p)}$$

- x: input variable (e.g. geographical location).
- $\boldsymbol{y}(x)$: $p \times 1$ vector of output variables (responses) evaluated at x.
- W: $p \times q$ matrix of mixing weights.
- $\boldsymbol{f}(x)$: $q \times 1$ vector of Gaussian process functions.
- σ_y: hyperparameter controlling noise variance.
- $\boldsymbol{z}(x)$: i.i.d Gaussian white noise with $p \times p$ identity covariance I_p.



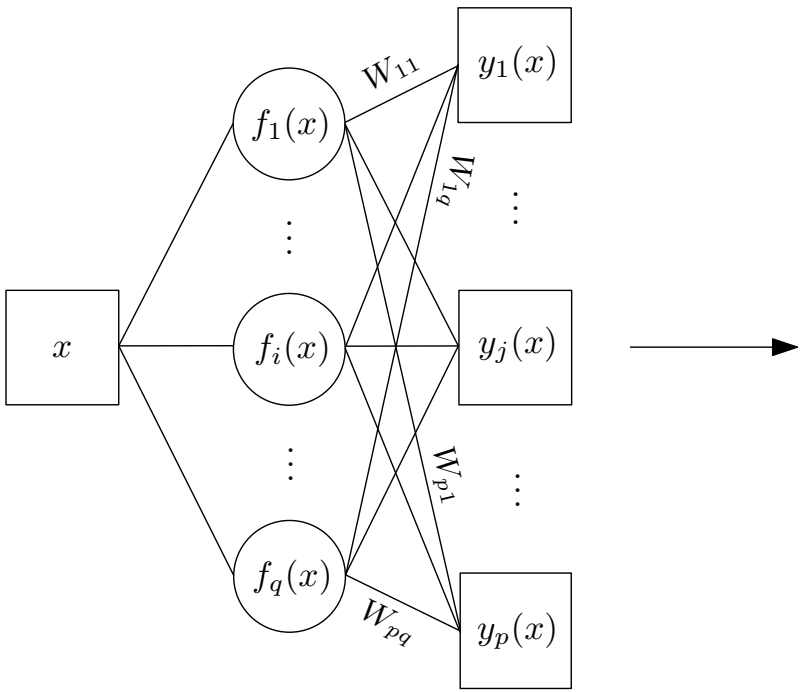
Deriving the GPRN



At $x = x_1$ the two outputs (responses) y_1 and y_2 are correlated since they share the basis function f_1. At $x = x_2$ the outputs are independent.

From the SLFM to the GPRN

SLFM



GPRN

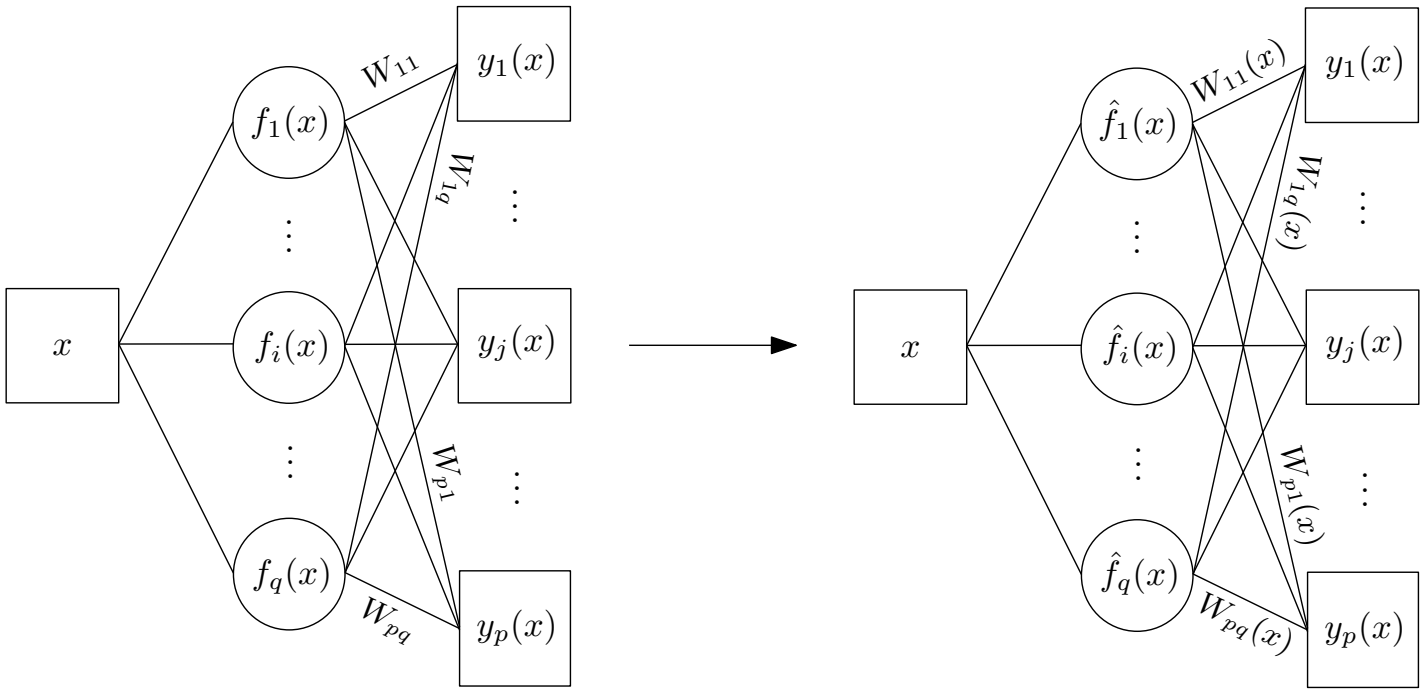
?

$$\overbrace{\boldsymbol{y}(x)}^{p\times 1} = \overbrace{W}^{p\times q}\overbrace{\boldsymbol{f}(x)}^{q\times 1} + \sigma_y \overbrace{\boldsymbol{z}(x)}^{\mathcal{N}(0,I_p)}$$

From the SLFM to the GPRN

SLFM

GPRN



$$\overbrace{\boldsymbol{y}(x)}^{p\times 1} = \overbrace{W}^{p\times q}\overbrace{\boldsymbol{f}(x)}^{q\times 1} + \sigma_y \overbrace{\boldsymbol{z}(x)}^{\mathcal{N}(0,I_p)}$$

$$\overbrace{\boldsymbol{y}(x)}^{p\times 1} = \overbrace{W(x)}^{p\times q}\underbrace{[\overbrace{\boldsymbol{f}(x)}^{q\times 1} + \sigma_f \overbrace{\epsilon(x)}^{\mathcal{N}(0,I_q)}]}_{\hat{\boldsymbol{f}}(x)} + \sigma_y \overbrace{\boldsymbol{z}(x)}^{\mathcal{N}(0,I_p)}$$

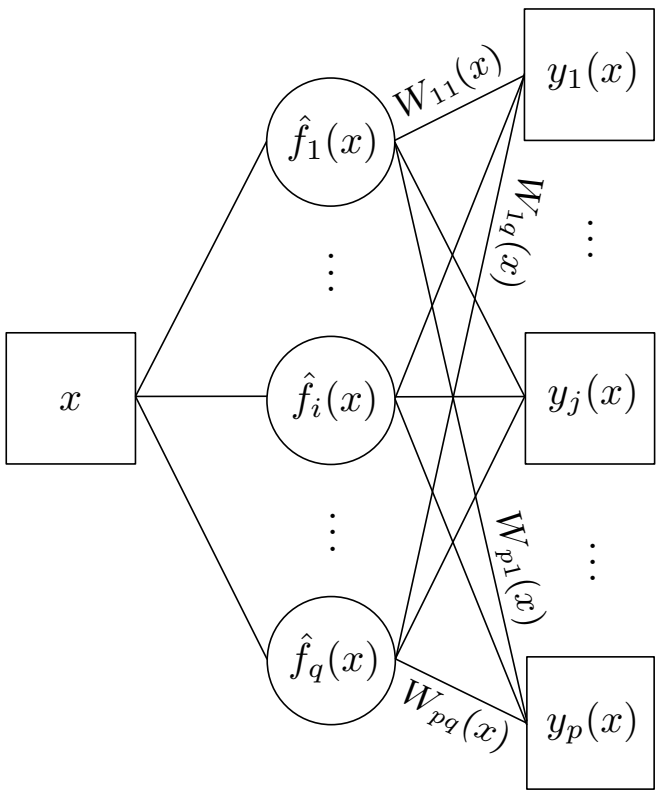
Gaussian process regression networks

$$\overbrace{\boldsymbol{y}(x)}^{p\times 1} = \overbrace{W(x)}^{p\times q}\underbrace{[\overbrace{\boldsymbol{f}(x)}^{q\times 1} + \sigma_f \overbrace{\epsilon(x)}^{\mathcal{N}(0,I_q)}]}_{\hat{\boldsymbol{f}}(x)} + \sigma_y \overbrace{\boldsymbol{z}(x)}^{\mathcal{N}(0,I_p)}$$

or, equivalently,

$$\boldsymbol{y}(x) = \overbrace{W(x)\boldsymbol{f}(x)}^{\text{signal}} + \overbrace{\sigma_f W(x)\boldsymbol{\epsilon}(x) + \sigma_y \boldsymbol{z}(x)}^{\text{noise}}.$$

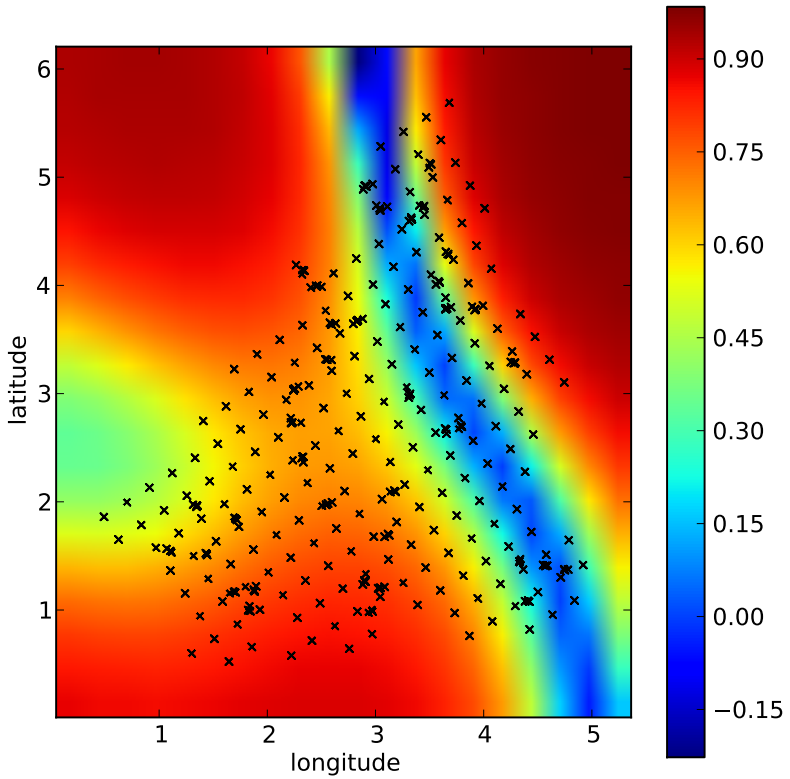
- $\boldsymbol{y}(x)$: $p \times 1$ vector of output variables (responses) evaluated at x.
- $W(x)$: $p \times q$ matrix of weight *functions*. $W(x)_{ij} \sim \mathcal{GP}(0, k_w)$.
- $\boldsymbol{f}(x)$: $q \times 1$ vector of Gaussian process node functions. $\boldsymbol{f}(x)_i \sim \mathcal{GP}(0, k_f)$.
- σ_f, σ_y: hyperparameters controlling noise variance.
- $\boldsymbol{\epsilon}(x), \boldsymbol{z}(x)$: Gaussian white noise.

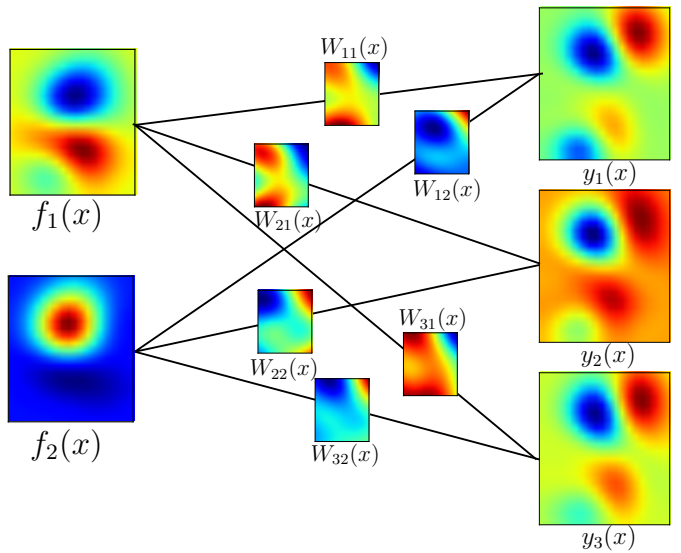


GPRN Inference

- We sample from the posterior over Gaussian processes in the weight and node functions using elliptical slice sampling (ESS) (Murray, Adams, and MacKay, 2010). ESS is especially good for sampling from posteriors with correlated Gaussian priors.
- We also approximate this posterior using a message passing implementation of variational Bayes (VB).
- The computational complexity is cubic in the number of data points and linear in the number of response variables, per iteration of ESS or VB.
- Details are in the paper.

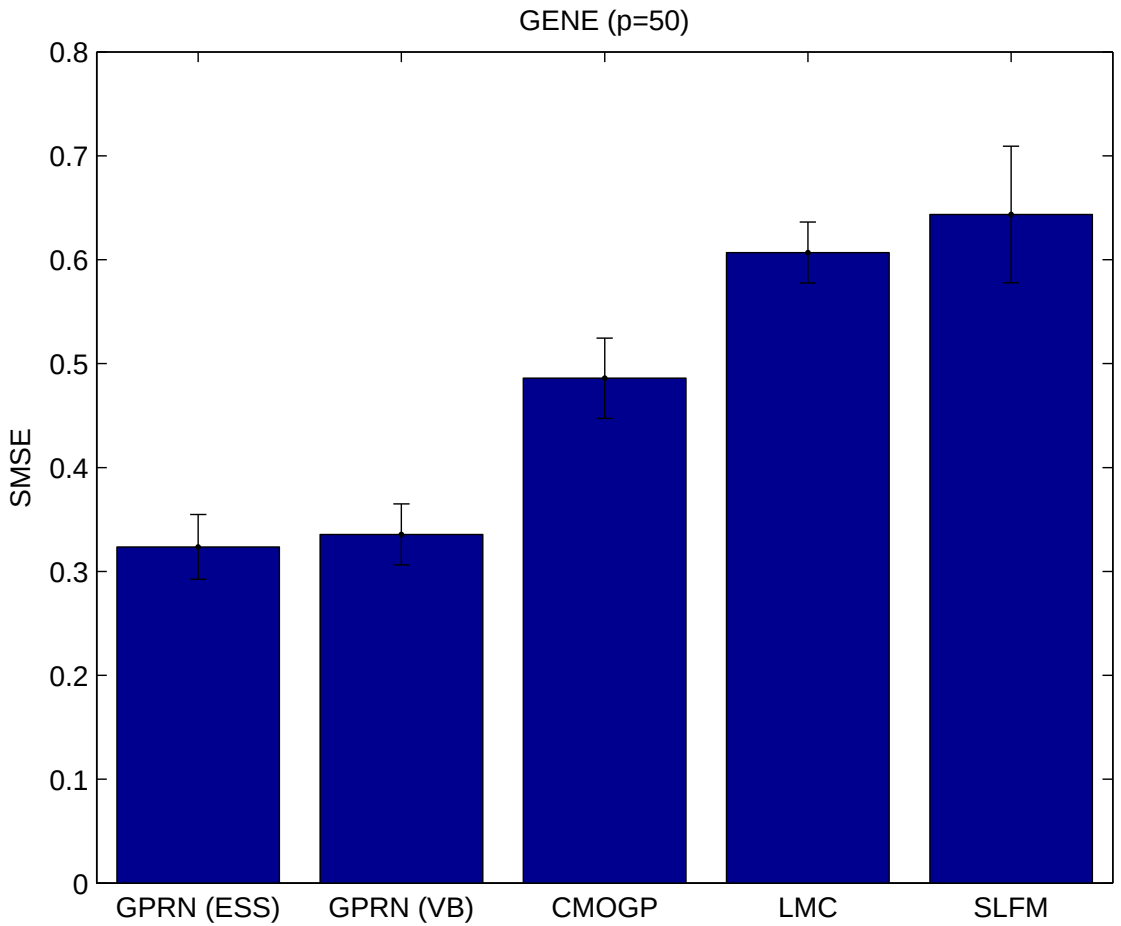
GPRN Results, Jura Heavy Metal Dataset



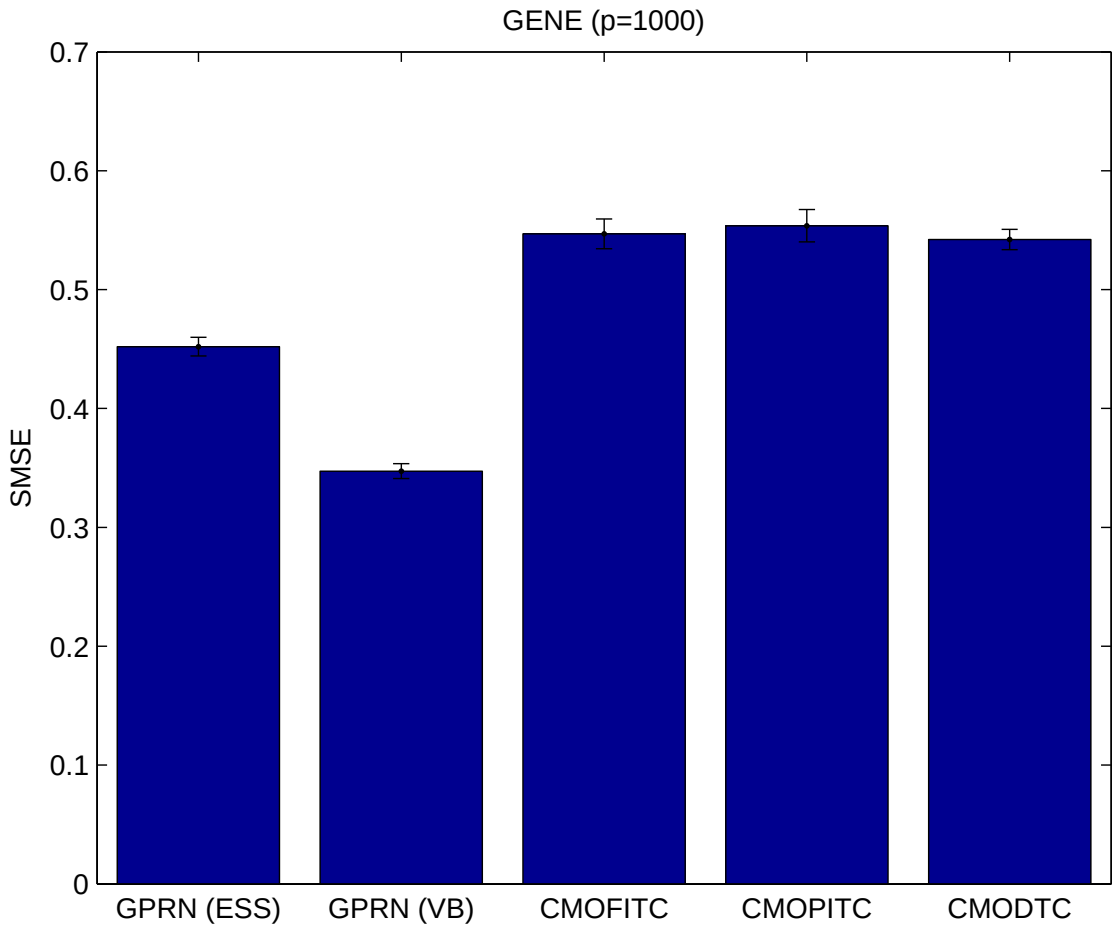


$$\boldsymbol{y}(x) = \overbrace{W(x)\boldsymbol{f}(x)}^{\text{signal}} + \overbrace{\sigma_f W(x)\boldsymbol{\epsilon}(x) + \sigma_y \boldsymbol{z}(x)}^{\text{noise}}.$$

GPRN Results, Gene Expression 50D



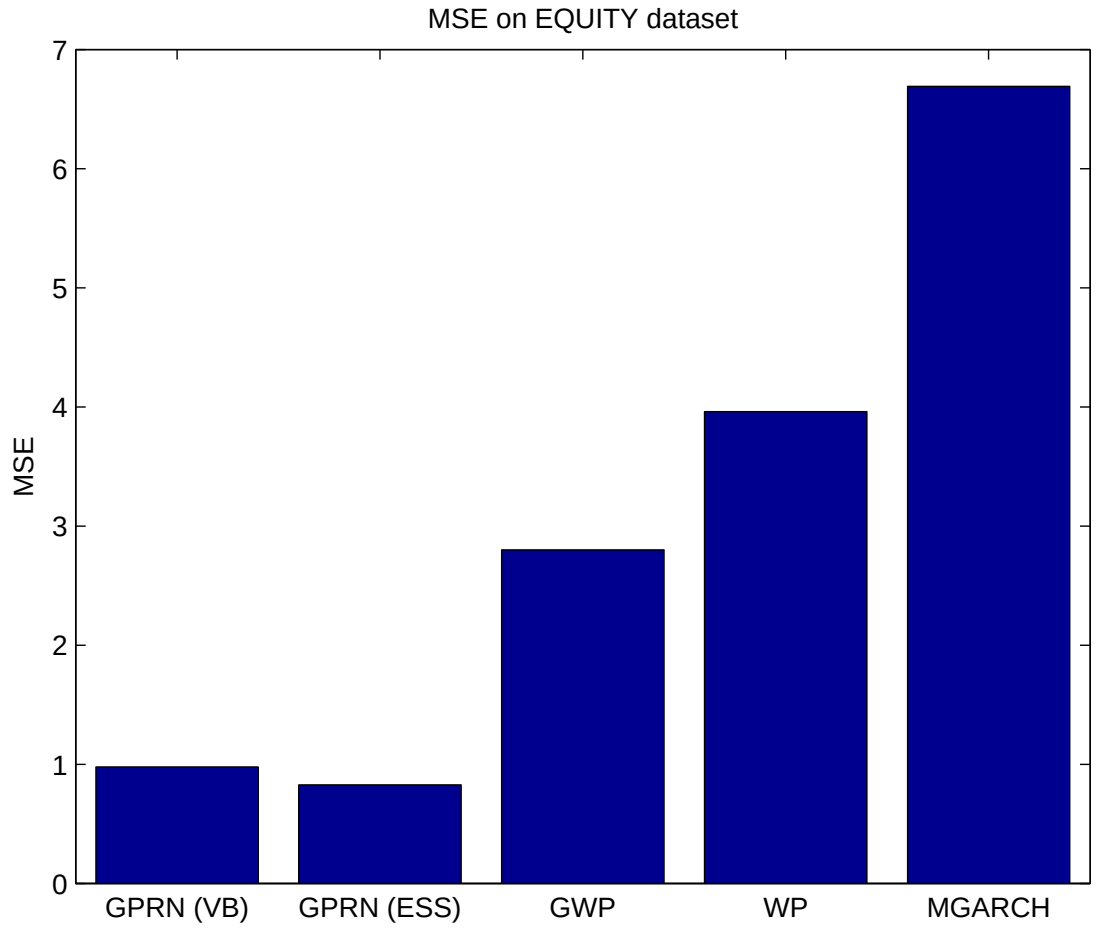
GPRN Results, Gene Expression 1000D



Training Times on GENE

Training time	GENE (50D) (s)	GENE (1000D) (s)
GPRN (VB)	12	330
GPRN (ESS)	40	9000
LMC, CMOGP, SLFM	minutes	days

Multivariate Volatility Results



Summary

- A *Gaussian process regression network* is used for multi-task regression and multivariate volatility, and can account for input dependent signal and noise covariances.
- Can scale to thousands of dimensions.
- Outperforms multi-task Gaussian process models and multivariate volatility models.

Generalised Wishart Processes

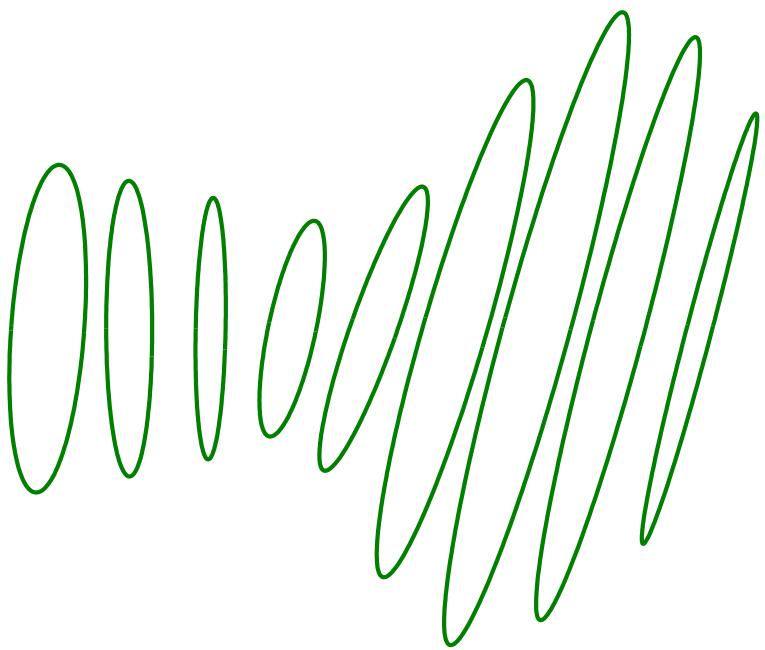
Recall that the GPRN model can be written as

$$\boldsymbol{y}(x) = \overbrace{W(x)\boldsymbol{f}(x)}^{\text{signal}} + \overbrace{\sigma_f W(x)\boldsymbol{\epsilon}(x) + \sigma_y \boldsymbol{z}(x)}^{\text{noise}}.$$

The induced noise process,

$$\Sigma(x)_{\text{noise}} = \sigma_f^2 W(x) W(x)^{\mathrm{T}} + \sigma_y^2 I\,,$$

is an example of a *Generalised Wishart Process* (Wilson and Ghahramani, 2010). At every x, $\Sigma(x)$ is marginally Wishart, and the dynamics of $\Sigma(x)$ are governed by the GP covariance kernel used for the weight functions in $W(x)$.



GPRN Inference New

$$\boldsymbol{y}(x) = \overbrace{W(x)\boldsymbol{f}(x)}^{\text{signal}} + \overbrace{\sigma_f W(x)\boldsymbol{\epsilon}(x) + \sigma_y \boldsymbol{z}(x)}^{\text{noise}}.$$

Prior is induced through GP priors in nodes and weights

$$p(\boldsymbol{u}|\sigma_f, \boldsymbol{\gamma}) = \mathcal{N}(0, C_B)$$

Likelihood

$$p(\mathcal{D}|\boldsymbol{u}, \sigma_f, \sigma_y) = \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{y}(x_i); W(x_i)\hat{\boldsymbol{f}}(x_i), \sigma_y^2 I_p)$$

Posterior

$$p(\boldsymbol{u}|\mathcal{D}, \sigma_f, \sigma_y, \boldsymbol{\gamma}) \propto p(\mathcal{D}|\boldsymbol{u}, \sigma_f, \sigma_y) p(\boldsymbol{u}|\sigma_f, \boldsymbol{\gamma})$$

We sample from the posterior using elliptical slice sampling (Murray, Adams, and MacKay, 2010) or approximate it using a message passing implementation of variational Bayes.