1 Data

Name	Symbol	Dimension
Replicates		R
Samples		N
Timepoints		T
Genes		D
Confounders		Q
Expression	$\mathbf{Y}$	$NRT \times D$
Latent Variables of Confounders	$\mathbf{X}$	$NRT \times Q$
Confounders	$\mathbf{C}$	$NRT \times D$

Table 1: Data explanation

2 Assumption on Confounder influence

The confounders are assumed to additively contribute to gene expression:

$$\mathbf{Y} = \mathbf{Y}_{\text{true}} + \mathbf{C} + \sigma^2 \mathbf{I} \ , \tag{1}$$

where in the linear case the confounders are

$$\mathbf{C} = \mathbf{X}\mathbf{W} \tag{2}$$

3 Confounder Simulation

3.1 Linear

$$\mathbf{X} = \text{randn}(NRT, Q) \tag{3}$$
$$\mathbf{W} = \text{randn}(Q, D) \tag{4}$$
$$\mathbf{C} = \mathbf{X}\mathbf{W} \tag{5}$$

4 Confounder Learning

GPLVM: $p(\mathbf{Y}|X, t, t', \theta) = \mathcal{N}(\mathbf{Y}|\mathbf{0}, \mathbf{K}(X, t, t', \theta))$ In the following we will discuss different choices of $\mathbf{K}(X, t, t', \theta)$

4.1 Linear Confounders

Learn confounders with linear covariance:

$$K = \mathbf{X}\mathbf{A}\mathbf{X}^\top + \sigma^2 \mathbf{I}, \quad \text{where} \tag{6}$$

$\mathbf{A}$ has dimensional weights on diagonal and $\mathbf{X}$ are linear learned confounders. See Figure 2 for results.

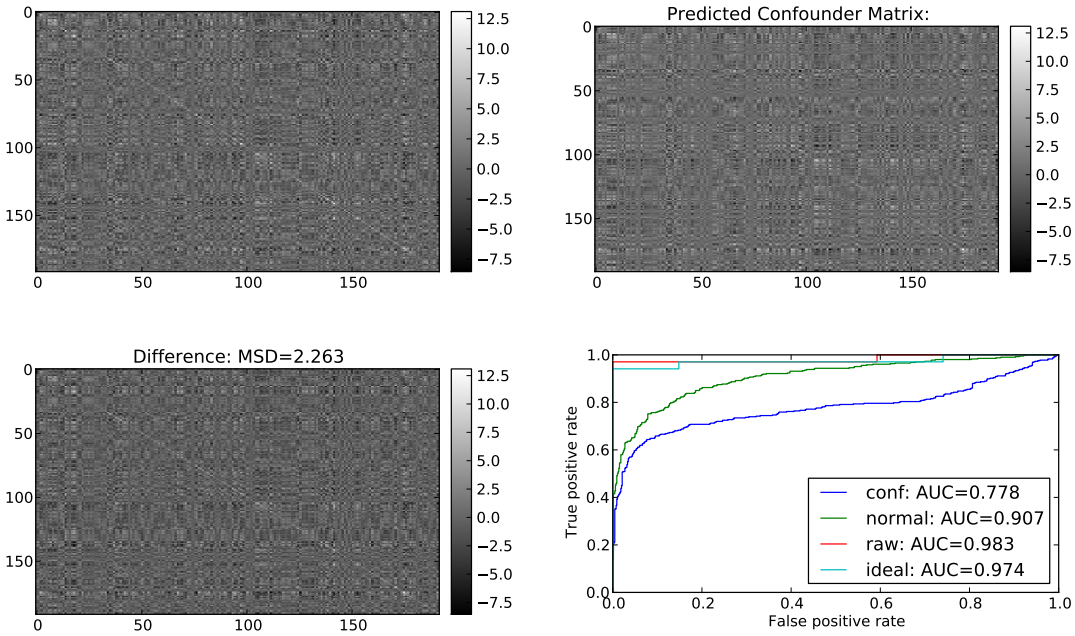

Figure 1: 4.1

4.2 Condition specific linear Confounders

Learn confounders with linear covariance and condition matrix:

$$K = \mathbf{X}\mathbf{A}\mathbf{X}^\top + \mathbf{K}_c + \sigma^2\mathbf{I}, \quad \text{where} \tag{7}$$

$\mathbf{A}$ has dimensional weights on diagonal, $\mathbf{X}$ are linear learned confounders and $\mathbf{K}_c$ depicts the condition structure of the data:

$$\begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots & & \\ 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & & & \vdots & & \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix} \tag{8}$$

See Figure 2 for results.

5 GPTwoSample Model Ideas:

5.1 Model 1: Confounders included in Confounders (Currently used)

Learn confounders $\mathbf{X}$ through GPLVM. Include Condounders as Covariance Matrix

$$\mathbf{K}_\mathbf{X} = \mathbf{X}\mathbf{X}^\top \tag{9}$$

into model, as follows:

$$p(\mathbf{Y}|\mathbf{t}, \theta, \mathbf{X}) = \prod_d^D \mathcal{N}(\mathbf{y}_d|\mathbf{0}, \mathbf{K}_\theta(\mathbf{t}) + \mathbf{K}_\mathbf{X} + \sigma^2\mathbf{I}) \quad . \tag{10}$$

5.2 Model 2: Added Confounders

Learn confounders $\mathbf{X}$ and predict confounder matrix by GPLVM. Subtract confounder matrix from observed gene expression and run normal GPTwoSample on residuals.

$$\mathbf{Y}_{\text{non-confounded}} = \mathbf{Y} - \text{GPLVM.predict}(\mathbf{X}) \tag{11}$$

$$p(\mathbf{Y}_{\text{non-confounded}}) = \prod_{d}^{D} \mathcal{N}(\mathbf{y}_d | \mathbf{0}, \mathbf{K}_\theta(\mathbf{t}) + \sigma^2 \mathbf{I}) \quad . \tag{12}$$

5.3 Model 3: One Confounder Matrix per Condition

Learn confounders $\mathbf{X}_1$ and $\mathbf{X}_2$ on condition $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively. Then either predict or incorporate confounders as covariance into GPTwoSample.

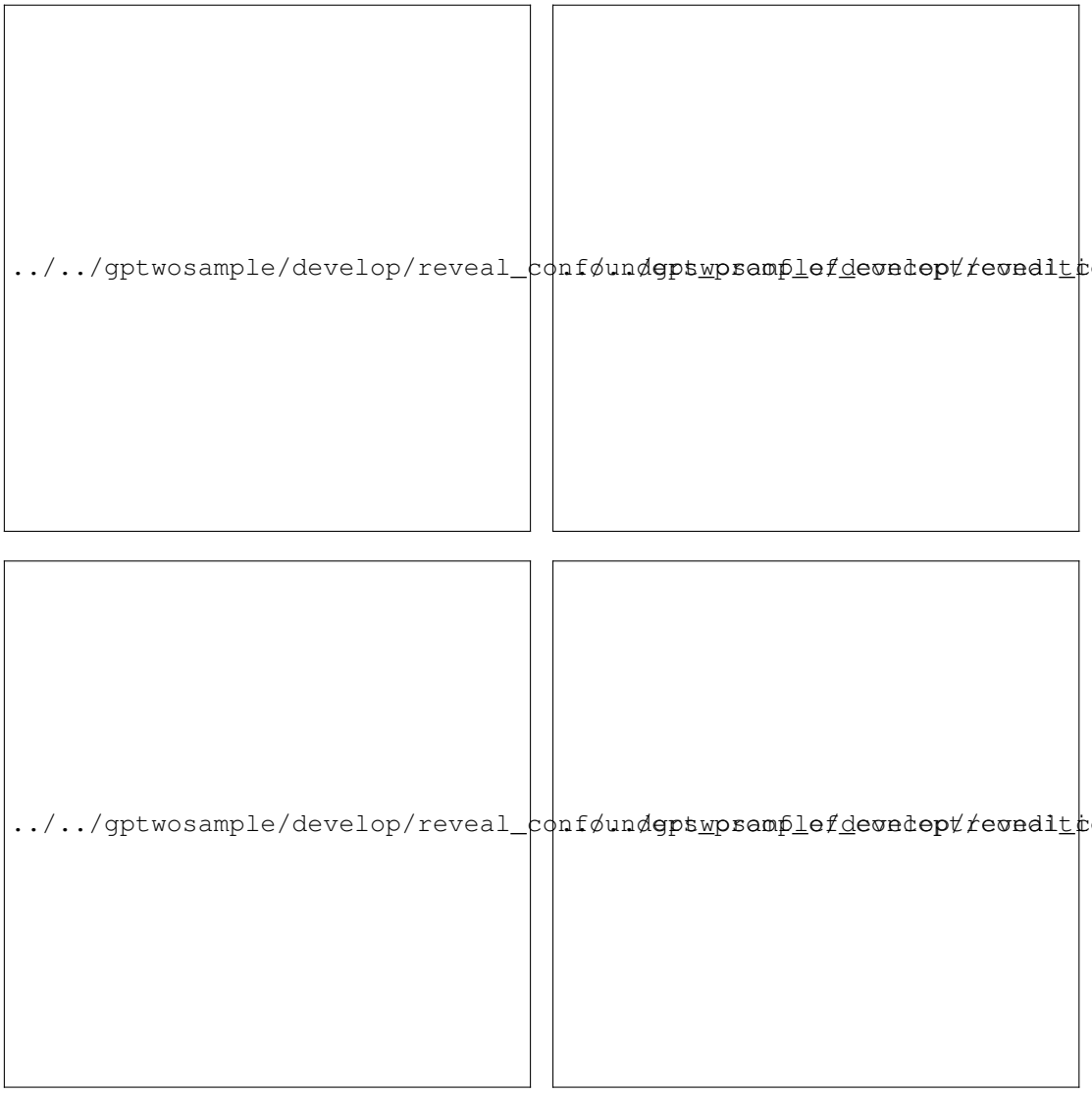
../../gptwosample/develop/reveal_confounder_simulation/reveal_confounder_simulation.ipynb
../../gptwosample/develop/reveal_confounder_different_sample_size/reveal_confounder_different_sample_size.ipynb

Figure 2: 4.2