Statistical Model

	I	Br	Cl	MA	CS	FA
ion:	1	2	3	4	5	6

Solvent -> (Polarity, minimum volume of enclosing eclipse)

Solubility of solution x: (solubility = - Energy)

$$V_x = \sum_{i=1}^{6} \alpha_i Z_i^x + \beta_x + \zeta + f(Z_7^x, Z_8^x)$$

- $Z_i^x = 1$: ion i present in solution x, for i = 1...6
- Z_7^x, Z_8^x : Polarity, maximum volume of enclosing eclipse
- $f(x_1, x_2)$: a 2-d Gaussian process with prior mean $\mu_0(\cdot)$ and covariance $\Sigma_0(\cdot, \cdot)$
- Place normal prior distributions on α_i, β_x, ζ
- Estimate hyper-parameters with MLE based on historical data

Statistical Model - Cont

Since a normal prior is placed on each of the parameter,

$$\begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_{135} \end{pmatrix} \sim N(\mu^0, \Sigma^0)$$

- After each observation, we can update the mean and covariance.
- We use μ^n, Σ^n to denote the mean and covariance of the multivariate normal after n observations.

Using EI to decide where to sample next

Let $\hat{y}^*$ be the largest solubility among the n observed samples so far. The expected improvement (EI) of a solution x is:

$$EI(x) = \mathbb{E}\Big[\max\{V_x - \hat{y}^*, 0\}\Big|\mu^n, \Sigma^n\Big]$$

Next solution to evaluate:

$$x^* = \text{argmax}_{x\in\{1,\ldots,135\}} EI(x)$$