TSA Tutorial 3
12.5.2020

Manuel Brenner
manuel.brenner@zi-mannheim.de

Organizational Point

Moodle was down from 12-14, but should be up again...but if your hand-in is slightly delayed that is okay

I’m aiming for a deadline next week so correction of some of the sheets can be delayed a little bit

Likelihood of a model

- The likelihood function (often simply called the likelihood) measures the quality of fit of a statistical model to a sample of data for given **values** of the unknown parameters.

  → Maximum Likelihoood finds the best parameters of the model based on the sample

Gaussian likelihood in a prediction framework

- Likelihood can be thought of as composed of a prediction term+ an entropy terms

$$\mathrm{L}(\{\alpha_i\},\sigma) = \prod_{t=p+1}^{T} (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2}[x_t-(\alpha_0+\sum_{i=1}^{p}\alpha_i x_{t-i})]^2/\sigma^2} = (2\pi)^{-(T-p)/2} \,|\, \sigma^2\mathbf{I} \,|^{-1/2}\; e^{-(1/2)\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\sigma^{-2}} .$$

$$\log \mathrm{L}(\{\alpha_i\},\sigma) = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \frac{1}{2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\sigma^{-2},$$

Likelihood in a prediction framework

- inherent noisiness of the process gives you a measure of how good your predictions can be ---> find a tradeoff between precision of prediction and stochasticity of the data

  → if you expect the world to be really uncertain, you don't put a lot of weight on your predictions

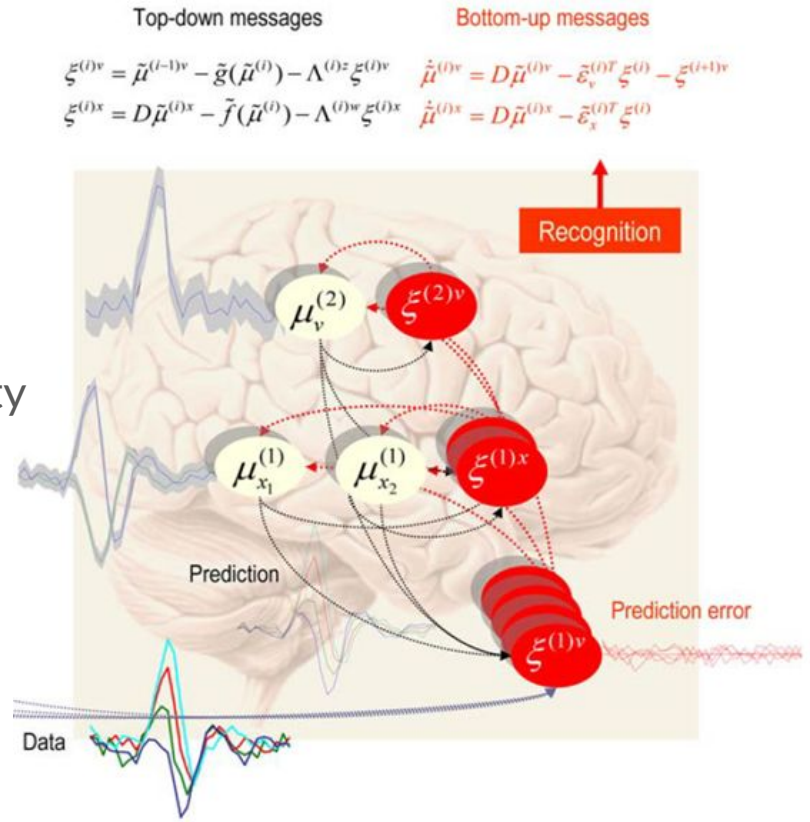
ELBO for Gaussian Process+Observation Models

$$
\begin{aligned}
E_{\mathbf{Z}}\{\log p(\mathbf{X},\mathbf{Z}\mid\boldsymbol{\theta})\} &= E_{\mathbf{Z}}\{\log[p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\theta})p(\mathbf{Z}\mid\boldsymbol{\theta})]\} \\
&= E_{\mathbf{Z}}\left\{\log\left[p(\mathbf{z}_1\mid\boldsymbol{\theta})p(\mathbf{x}_1\mid\mathbf{z}_1,\boldsymbol{\theta})\prod_{t=2}^{T}p(\mathbf{z}_t\mid\mathbf{z}_{t-1},\boldsymbol{\theta})p(\mathbf{x}_t\mid\mathbf{z}_t,\boldsymbol{\theta})\right]\right\} \\
&= E_{\mathbf{Z}}\left\{-\frac{1}{2}\left[T\log|\boldsymbol{\Sigma}|+(\mathbf{z}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}_1-\boldsymbol{\mu}_0)+\sum_{t=2}^{T}(\mathbf{z}_t-\mathbf{A}\mathbf{z}_{t-1})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}_t-\mathbf{A}\mathbf{z}_{t-1})\right.\right. \\
&\qquad \left.\left.+T\log|\boldsymbol{\Gamma}|+\sum_{t=1}^{T}(\mathbf{x}_t-\mathbf{B}\mathbf{z}_t)^T\boldsymbol{\Gamma}^{-1}(\mathbf{x}_t-\mathbf{B}\mathbf{z}_t)\right]+const.\right\}
\end{aligned}
$$

Connection to Neuroscience

Brain makes predictions

→ model inversion by prediction errors, weighted by expecations over uncertainty in environment



Model selection in machine learning

We need to limit the capacity of the model as not to over-explain the data. In ML practice usually with some regularization method.

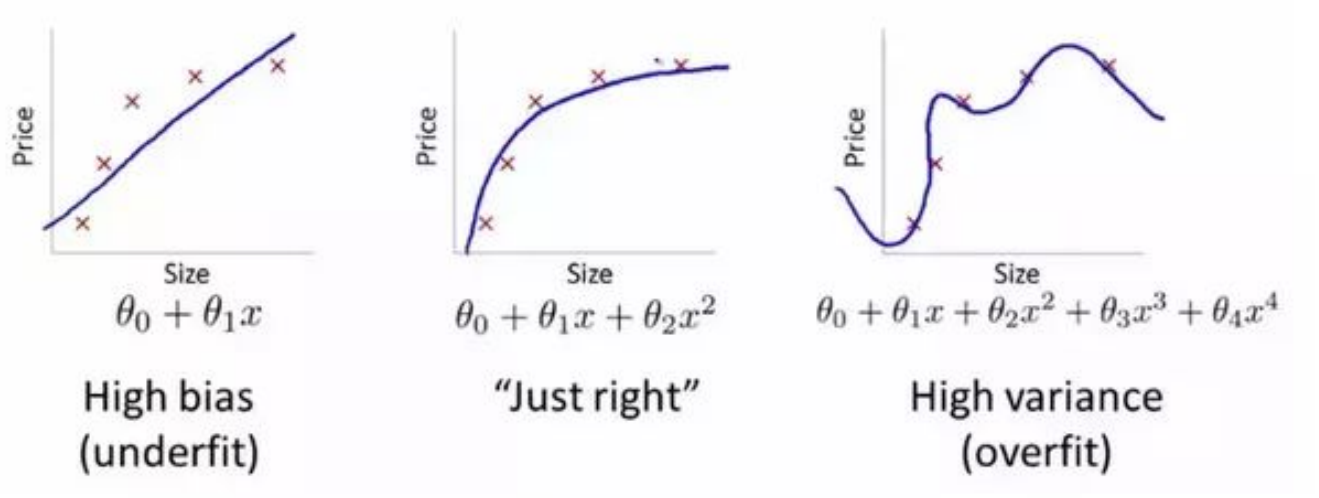
In many cases, dropout or L1/ L2 regularization with a large enough data set is enough to do the trick (see Computational Statistics Lecture)

Many modern ML practices are focused on starting with large models and reducing the size consecutively

Bias-Variance Tradeoff

- Often the number of parameters we can use is not a priori clear, and we can add more or less parameters
- We want models that both capture the data well but generalizes well to unseen data
- → We need to balance the ratio between bias and variance by finding the optimal trade-off between overfitting and underfitting

Bias-Variance Tradeoff

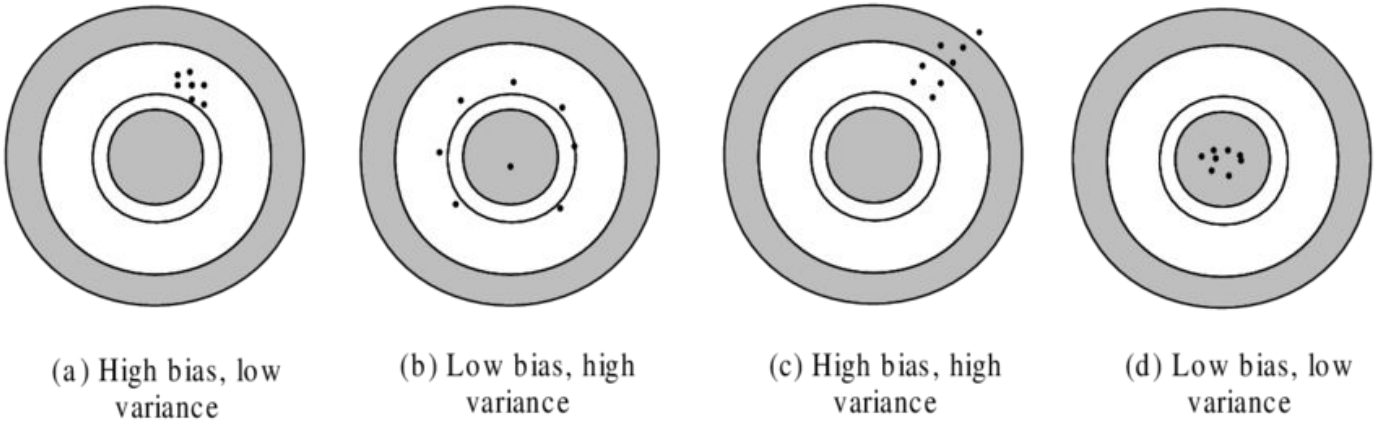


Bias Variance Tradeoff

$$Err(x) = \left(E[\hat{f}\,(x)] - f(x)\right)^2 + E\left[\left(\hat{f}\,(x) - E[\hat{f}\,(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Bias-Variance Tradeoff



(a) High bias, low variance

(b) Low bias, high variance

(c) High bias, high variance

(d) Low bias, low variance

Bias-Variance Tradeoff

In high bias models, you are underfitting the signal

In high variance models, you are overfitting the noise terms

→ this is why we always ask ourselves if the additional model parameters explain significantly more variance or if the residuals after fitting the models are simply white noise

We don't want to "overfit" by adding degrees of freedom

Noise and entropy

The entropy gives a measure of inherent uncertainty in a distribution

See information theoretic interpretation of entropy: how well can I predict the outcome of sampling from the distribution?

In the case of Gaussian distributions, this is reflected in the covariance:

narrow distributions have small entropy, wide distributions have large entropy

Back to ELBO

$$
\begin{aligned}
\log p(\boldsymbol{X}, \boldsymbol{Z}) = & -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_0\right| - \frac{1}{2}\left(\boldsymbol{z}_1 - \boldsymbol{\mu}_0\right)^{\intercal}\boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{z}_1 - \boldsymbol{\mu}_0\right) \\
& -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Gamma}\right| - \frac{1}{2}\left(\boldsymbol{x}_1 - \boldsymbol{\mu}_{x_1}\right)^{\intercal}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{x}_1 - \boldsymbol{\mu}_{x_1}\right) \\
& + \sum_{t=2}^{T}\Bigg(\left(-\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\left(\boldsymbol{z}_t - \boldsymbol{\mu}_{z_t}\right)^{\intercal}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{z}_t - \boldsymbol{\mu}_{z_t}\right)\right) \\
& + \left(-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Gamma}\right| - \frac{1}{2}\left(\boldsymbol{x}_t - \boldsymbol{\mu}_{x_t}\right)^{\intercal}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{x}_t - \boldsymbol{\mu}_{x_t}\right)\right)\Bigg).
\end{aligned}
$$