Temporal Time Series Analysis (Part II)

Joaquín Rapela

Gatsby Computational Neuroscience Unit
University College London

April 27, 2025

Contents

Contents

Forecasting

Forecasting is the problem of predicting the value of $x_{n+h}, h > 0$, of a stationary time series, in term of the previous m values$\{x_n, \ldots, x_{n-(m-1)}\}$. The mean of such predictor is

$$\text{mean}(\text{pred}(x_{n+h}|x_n, \ldots, x_{n-(m-1)})) = \mu + \mathbf{a_m}^\intercal \begin{bmatrix} x_n - \mu \\ \cdots \\ x_{n-(m-1)} - \mu \end{bmatrix}$$

and its variance is

$$\text{var}(\text{pred}(x_{n+h}|x_n, \ldots, x_{n-(m-1)})) = \gamma(0) - \mathbf{a_m}^\intercal \gamma_m(h)$$

Forecasting

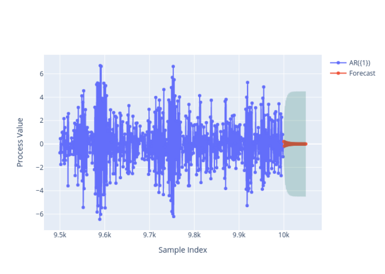
with

$$
\begin{aligned}
\Gamma_m \mathbf{a_m} &= \gamma_m(h) \\
\Gamma_m = [\gamma(i-j)]_{i,j=1}^m &= \begin{bmatrix}
\gamma(0) & \gamma(1) & \gamma(2) & \gamma(3) & \dots & \gamma(m-1) \\
\gamma(1) & \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(m-2) \\
\gamma(2) & \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(m-3) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\gamma(m-1) & \gamma(m-2) & \gamma(m-3) & \gamma(m-4) & \dots & \gamma(0)
\end{bmatrix} \\
\mathbf{a_m} &= [a_1, \dots, a_m]^\intercal \\
\gamma_m(h) &= [\gamma(h), \gamma(h+1), \dots, \gamma(h+m-1)]^\intercal
\end{aligned}
$$

AR(1) forecasting example

Example (Forecasting with an AR(1) model)

Simulate N=1,000 samples from an AR(1) stochastic process with $\phi = -0.9$ and $\sigma_w = 1.0$. Use the last 500 samples to forecast 50 samples (i.e., $n = 1,000, m = 500, h = 1, \ldots, 50$).



Marginals and conditionals of Gaussians are Gaussians

Theorem 1 (Marginals and conditionals of Gaussians are Gaussians)

Given $\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}$ *such that*

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x} \middle| \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right) \\ &= \mathcal{N}\left(\mathbf{x} \middle| \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}^{-1}\right) \end{aligned}$$

Then

$$\begin{aligned} p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}\left(\mathbf{x}_a \middle| \boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b), \Lambda_{aa}^{-1}\right) && (1) \\ &= \mathcal{N}\left(\mathbf{x}_a \middle| \boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right) && (2) \\ p(\mathbf{x}_b) &= \mathcal{N}\left(\mathbf{x}_b \middle| \boldsymbol{\mu}_b, \Sigma_{bb}\right) && (3) \end{aligned}$$

Proof in the Appendix.

Relevance of the conditional density of Gaussians

The expression of the conditional density of jointly Gaussian random variables is used in the derivation of

1. Bayesian linear regression (Bishop, 2016),
2. Gaussian process regression (Williams and Rasmussen, 2006),
3. Gaussian process factor analysis (Yu et al., 2009),
4. linear dynamical systems (Durbin and Koopman, 2012).

Derivation of the forecasting equations using the expression of the conditional of Gaussians

Take $\mathbf{x}_a = [x_{m+h}]$ and $\mathbf{x}_b = [x_n, \ldots, x_{n-m+1}]^\intercal$ in Eq. 2.

Derivation of estimator of missing values using the expression of the conditional of Gaussians

Exercise 1

You are given an AR(1) time series with missing values $[x_{n+1}, \ldots, x_{n+h}]$*. Use the expression of the conditional of Gaussians to find the optimal estimator, in the mean square error sense, of the missing values using observations* $[x_n, \ldots, x_{n-(m-1)}]$ *and* $[x_{n+h+1}, \ldots, x_{n+h+m}]$*.*

Hint: take $\mathbf{x}_a = [x_{n+1}, \ldots, x_{n+h}]$ and
$\mathbf{x}_b = [x_{n+h+1}, \ldots, x_{n+h+m}, x_n, \ldots, x_{n-(m-1)}]$ Eq. 2.

Contents

Yule-Walker equations

Claim 1 (Yule-Walker equations for AR(p) model)

If $\{x_t\}$ is and AR(p) random process

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t \quad \textit{with } w_t \sim N(0, \sigma^2)$$

then

$$\gamma(h) = \phi_1 \gamma(h-1) + \cdots + \phi_p \gamma(h-p) \quad h = 1, \ldots, p \tag{4}$$

$$\gamma(0) = \phi_1 \gamma(h-1) + \cdots + \phi_p \gamma(h-p) + \sigma^2 \tag{5}$$

Proof.

See board. □

Yule-Walker equations

In matrix form the Yule-Walker equations 4 and 5 can be written as:

$$\Gamma_p \phi = \gamma_p \tag{6}$$

$$\gamma(0) = \phi^\intercal \gamma_p + \sigma^2 \tag{7}$$

with

$$\Gamma_p = [\gamma(i-j)]_{i,j=1}^{p} = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \gamma(3) & \dots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(p-2) \\ \gamma(2) & \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(p-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \gamma(p-3) & \gamma(p-4) & \dots & \gamma(0) \end{bmatrix}$$

$$\phi = [\phi(1), \dots, \phi(p)]^\intercal$$

$$\gamma_p = [\gamma(1), \dots, \gamma(p)]^\intercal$$

Yule-Walker estimators

Replacing γ by its estimate $\hat{\gamma}$ in Eqs. 6 and 7, we obtain the Yule-Walker estimators

$$
\begin{aligned}
\hat{\phi} &= \hat{\Gamma}_p^{-1}\hat{\gamma}_p \\
\hat{\sigma}^2 &= \hat{\gamma}(0) - \hat{\phi}^\intercal\hat{\gamma}_p \\
&\text{with} \\
\hat{\Gamma}_p &= [\hat{\gamma}(i-j)]_{i,j=1}^p \\
\hat{\phi} &= [\hat{\phi}(1), \ldots, \hat{\phi}(p)]^\intercal \\
\hat{\gamma}_p &= [\hat{\gamma}(1), \ldots, \hat{\gamma}(p)]^\intercal
\end{aligned}
$$

Large-sample distribution of Yule-Walker estimators

Theorem 2 (Large-sample distribution of Yule-Walker estimators)

For a large sample from an AR(p) random process

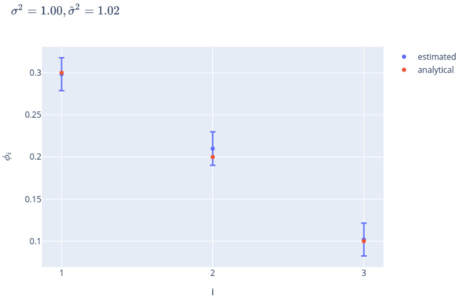
$$\hat{\phi} \sim N\left(\phi, n^{-1}\sigma^2\Gamma_p^{-1}\right).$$

Proof.

See Brockwell and Davis (1991, Section 8.10) □

Estimate coefficients of AR(3) model using the Yule-Walker estimators

Sample a time series of length $N = 1000$ from an AR(3) model. Estimate the coefficients of this model, and the variance of the noise, using the Yule-Walker estimators. Also, calculate the large sample estimates of the coefficients' variance. Plot the true and estimated coefficients. Add a 95% confidence bounds to the estimated coefficients.



Contents

Likelihood function

Definition 3 (Likelihood function)

Consider a random process $\{x_t\}$ with a probability density function parameterised by parameters θ, $f(\{x_t\}|\theta)$. Given a sample $\{x_i\}_{i=1}^{N}$, the **likelihood function** of θ, $\mathcal{L}(\theta)$, assigns to θ the value $f(\{x_i\}_{i=1}^{N}|\theta)$.

Likelihood function

Claim 2 (Likelihood function for an AR(1) random process)

The log likelihood function for the parameters $\theta = \{\phi, \sigma^2\}$ of an AR(1) random process, given observations $\{x_1, \dots, x_N\}$ is

$$\log \mathcal{L}(\phi, \sigma^2) = -\frac{N-1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=2}^{N}(x_n - \phi x_{n-1})^2 - \frac{1}{2}\log(2\pi\sigma^2) - \frac{x_1^2}{2\sigma^2}$$

Likelihood function for an AR(1) random process.

See board □

Maximum likelihood parameter estimates

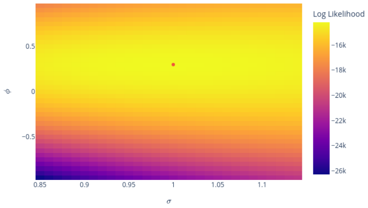
Definition 4 (Maximum likelihood parameters estimates)

Given a data sample $\{x_t\}$, the **maximum likelihood parameters estimates** are $\hat{\theta}_{ML} = \arg\max_\theta \mathcal{L}(\theta)$.

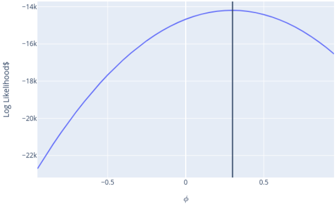
Maximum likelihood estimates of parameters of AR(1) process

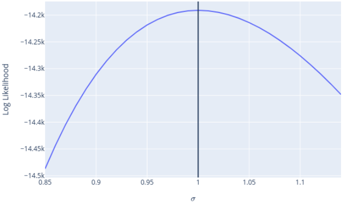
Example 5

Simulate a time series of length $N = 10,000$ from an AR(1) random process with $\phi = 0.3$ and $\sigma = 1$. Calculate the log-likelihood function on the simulated time series in the grid of parameters $0.85 \leq \sigma \leq 1.10$ (spacing $\delta_\sigma = 0.01$) and $-0.95 \leq \phi \leq 0.95$ (spacing $\delta_\phi = 0.05$). Verify that the calculated log likelihood is maximised at the simulated parameter values.



Maximum likelihood estimates of parameters of AR(1) process





Summary

References

time series analysis

- Brockwell and Davis (2002)
- Shumway and Stoffer (2016)
- Priestley (1981)

machine learning

- Bishop (2016)
- Murphy (2022)

Contents

Proof: the conditional of a Gaussian is a Gaussian (Theorem 1, Eq. 1)

Claim 3 (Quadratic form of Gaussian log pdf)

$p(\mathbf{x})$ is a Gaussian pdf with mean $\boldsymbol{\mu}$ and precision matrix Λ if and only if $\int p(\mathbf{x})d\mathbf{x} = 1$ and

$$\log p(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) + K \tag{8}$$

where K is a constant that does not depend on $\mathbf{x}$.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 1, Eq. 1)

Proof of Claim 3.

$\rightarrow$)

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}\Lambda^{-\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\Lambda(\mathbf{x}-\boldsymbol{\mu})\right\}\\
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\Lambda(\mathbf{x}-\boldsymbol{\mu})-\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\\
&= -\frac{1}{2}(\mathbf{x}^\intercal\Lambda\mathbf{x}-2\mathbf{x}^\intercal\Lambda\boldsymbol{\mu})-\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}-\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\\
&= -\frac{1}{2}(\mathbf{x}^\intercal\Lambda\mathbf{x}-2\mathbf{x}^\intercal\Lambda\boldsymbol{\mu})+K
\end{aligned}
$$

with $K=-\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}-\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})$.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 1, Eq. 1)

Proof of Claim 3.

$\leftarrow$)

$$
\begin{aligned}
\log p(\mathbf{x}) =& -\frac{1}{2}(\mathbf{x}^\intercal\Lambda\mathbf{x}-2\mathbf{x}^\intercal\Lambda\boldsymbol{\mu})+K\\
\log p(\mathbf{x}) =& -\frac{1}{2}(\mathbf{x}^\intercal\Lambda\mathbf{x}-2\mathbf{x}^\intercal\Lambda\boldsymbol{\mu})-\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}-\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\\
&+K+\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}+\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\\
=& -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\Lambda(\mathbf{x}-\boldsymbol{\mu})-\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\\
&+K+\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}+\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\\
=& \log N(\mathbf{x}|\boldsymbol{\mu},\Lambda)+K+\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}+\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\\
p(\mathbf{x}) =& N(\mathbf{x}|\boldsymbol{\mu},\Lambda)\exp\left(K+\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}+\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right) \qquad (9)
\end{aligned}
$$

Proof: the conditional of a Gaussian is a Gaussian (Theorem 1, Eq. 1)

Proof of Claim 3.

$\leftarrow$) cont

$$
\begin{aligned}
1 &= \int p(\mathbf{x})d\mathbf{x} \\
&= \int N(\mathbf{x}|\boldsymbol{\mu},\Lambda)\exp\left(K+\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}+\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right)d\mathbf{x} \\
&= \exp\left(K+\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}+\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right)\int N(\mathbf{x}|\boldsymbol{\mu},\Lambda)d\mathbf{x} \\
&= \exp\left(K+\frac{1}{2}\boldsymbol{\mu}^\intercal\Lambda\boldsymbol{\mu}+\log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right)
\end{aligned}
$$

From Eq. 9 then $p(\mathbf{x}) = N(\mathbf{x}|\boldsymbol{\mu},\Lambda)$.

$\square$

Proof: the conditional of a Gaussian is a Gaussian (Theorem 1, Eq. 1)

Proof of Theorem 1, Eq. 1.

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

$$\log p(\mathbf{x}_a|\mathbf{x}_b) = \log p(\mathbf{x}) - \log p(\mathbf{x}_b) = \log p(\mathbf{x}) + K$$

Therefore, the terms of $\log p(\mathbf{x}_a|\mathbf{x}_b)$ that depend on $\mathbf{x}_a$ are those of $\log p(\mathbf{x})$.
Steps for the proof:

1. isolate the terms of $\log p(\mathbf{x})$ that depend on $\mathbf{x}_a$,
2. notice that these term has the quadratic form of Claim 3, therefore $p(\mathbf{x}_a|\mathbf{x}_b)$ is Gaussian,
3. identify $\boldsymbol{\mu}$ and Λ in this quadratic form.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 1, Eq. 1)

Proof of Theorem 1, Eq. 1.

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}|\Lambda|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\Lambda(\mathbf{x}-\boldsymbol{\mu})\right)\\
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\Lambda(\mathbf{x}-\boldsymbol{\mu}) + K_1\\
&= -\frac{1}{2}[(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal, (\mathbf{x}_b-\boldsymbol{\mu}_b)^\intercal] \left[\begin{array}{cc} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{array}\right] \left[\begin{array}{c} \mathbf{x}_a-\boldsymbol{\mu}_a \\ \mathbf{x}_b-\boldsymbol{\mu}_b \end{array}\right] + K_1\\
&= -\frac{1}{2}\left\{(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal\Lambda_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) + 2(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\right.\\
&\quad \left. + (\mathbf{x}_b-\boldsymbol{\mu}_b)^\intercal\Lambda_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)\right\} + K_1\\
&= -\frac{1}{2}\left\{\mathbf{x}_a^\intercal\Lambda_{aa}\mathbf{x}_a - 2\mathbf{x}_a^\intercal(\Lambda_{aa}\boldsymbol{\mu}_a - \Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b))\right\} + K_2\\
&= -\frac{1}{2}\left\{\mathbf{x}_a^\intercal\Lambda_{aa}\mathbf{x}_a - 2\mathbf{x}_a^\intercal\Lambda_{aa}(\boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b))\right\} + K_2
\end{aligned}
$$

Comparing the last equation with Eq. 8 we see that $\Lambda = \Lambda_{aa}$, $\boldsymbol{\mu} = \boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)$ and conclude that $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b), \Lambda_{aa})$ □

Proof: the conditional of a Gaussian is a Gaussian (Theorem 1, Eq. 2)

Claim 4 (Inverse of a partitioned matrix)

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix} \tag{10}$$

where

$$M = (A - BD^{-1}C)^{-1}$$

Proof.

Exercise. Hint: verify that the multiplication of the inverse of the matrix in the right hand side of Eq. 10 with the matrix in the left hand side of the same equation is the identity matrix.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 1, Eq. 2)

Proof of Theorem 1, Eq. 2.

Using the definition

$$\left(\begin{array}{cc} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{array}\right)^{-1} = \left(\begin{array}{cc} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{array}\right)$$

and using Eq. 10, we obtain

$$\begin{aligned} \Lambda_{aa} &= (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \\ \Lambda_{ab} &= -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \end{aligned}$$

Replacing the above equations in Eq. 1 we obtain Eq. 2. □

Bishop, C. M. (2016). *Pattern recognition and machine learning*. Springer-Verlag New York.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: Theory and methods*. Springer-Verlag, 2nd edition.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. OUP Oxford.

Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. MIT press. https://probml.github.io/pml-book/book1.html.

Priestley, M. (1981). Spectral analysis and time series.

Shumway, R. H. and Stoffer, D. S. (2016). *Time series analysis and its applications*. Springer, 4 edition.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of neurophysiology*, 102(1):614–635.