The Gaussian distribution
Linear models for regression
References

Inference

Joaquín Rapela

Gatsby Computational Neuroscience Unit
University College London

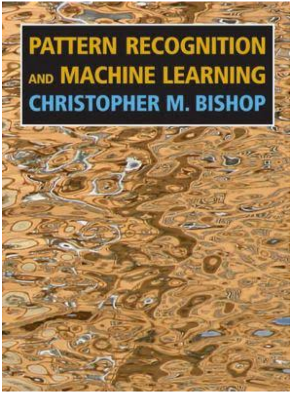
July 14, 2023

Contents

1. The Gaussian distribution
2. Linear models for regression

Main reference

I will mainly follow chapters two *Probability distributions* and three *Linear models for regression* from Bishop (2016).



Contents

1. The Gaussian distribution
2. Linear models for regression

The Gaussian distribution

- One-dimensional

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

- D-dimensional

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}\Sigma^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

The Gaussian is the maximum entropy distribution (Cover and Thomas, 1991)

Definition 1 (Differential entropy)

The differential entropy $h(X)$ of a continuous random variable X with a density $f(x)$ is defined as

$$h(X) = -\int_S f(X) \log f(x)\ dx$$

where S is the support set of the random variable.

Theorem 1 (The Gaussian is the maximum entropy distribution)

Let the random vector $X \in \mathbb{R}^n$ have zero mean and covariance K. Then $h(X) \leq \frac{1}{2} \log(2\pi e)^n |K|$, with equality if $X \sim \mathcal{N}(0, K)$.

The central limit theorem (Papoulis and Pillai, 2002)

Theorem 2 (The central limit theorem)

Given n independent and identically distributed random vectors $\mathbf{X}_i$*, with mean vector* $\boldsymbol{\mu} = E\{\mathbf{X}_i\}$ *and covariance matrix* $\boldsymbol{\Sigma}$*. Then*

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \rightarrow \mathcal{N}(0, \Sigma)$$

with convergence in distribution.

Very useful properties of the Gaussian distribution (Bishop, 2016)

Theorem 3 (Marginals and conditionals of Gaussians are Gaussians)

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

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

Then

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

Very useful properties of the Gaussian distribution (Bishop, 2016)

Theorem 4 (Marginals and conditionals of the linear Gaussian model)

Given the linear Gaussian model

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Lambda^{-1})$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|A\boldsymbol{\mu} + \mathbf{b}, L^{-1})$$

Then

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|A\boldsymbol{\mu} + \mathbf{b}, L^{-1} + A\Lambda^{-1}\Sigma^{\intercal})$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\Sigma\{A^{\intercal}L(\mathbf{y} - \mathbf{b}) + \Sigma\boldsymbol{\mu}\}, \Sigma)$$

where

$$\Sigma = (\Lambda + A^{\intercal}LA)^{-1}$$

Very useful properties of the Gaussian distribution (Bishop, 2016)

The conditional, $p(\mathbf{x}|\mathbf{y})$, of the linear Gaussian model is the fundamental result 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).

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Claim 1 (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

$$\log p(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) + K \tag{4}$$

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

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Proof of Claim 1.

$\rightarrow)$

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

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 3, Eq. 1)

Proof of Claim 1.

$\leftarrow$)

$$
\begin{aligned}
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) + K \\
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} - \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&\quad + K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal \Lambda (\mathbf{x} - \boldsymbol{\mu}) - \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&\quad + K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&= \log N(\mathbf{x}|\boldsymbol{\mu}, \Lambda) + K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
p(\mathbf{x}) &= N(\mathbf{x}|\boldsymbol{\mu}, \Lambda) \exp\left(K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right)
\end{aligned} \tag{5}
$$

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

The Gaussian distribution
Linear models for regression
References

Proof of Claim 1.

$\leftarrow$) 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. 5 then $p(\mathbf{x}) = N(\mathbf{x}|\boldsymbol{\mu}, \Lambda)$.

$\square$

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Proof of Theorem 3, 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 1, 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 3, Eq. 1)

Proof of Theorem 3, Eq. 1.

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

Comparing the last equation with Eq. 4 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 $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b), \Lambda_{aa})$ $\square$

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 2)

Claim 2 (Inverse of a partitioned matrix)

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix} \quad (6)$$

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. 6 with the matrix in the left hand side of the same equation is the identiy matrix.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 2)

Proof of Theorem 3, Eq. 2.

Using the definition

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

and using Eq. 6, we obtain

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

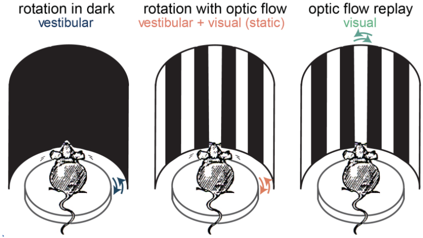
Replacing the above equations in Eq. 1 we obtain Eq. 2. $\square$

Contents

1 The Gaussian distribution

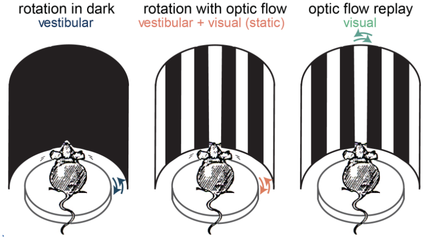
2 Linear models for regression

Linear regression example

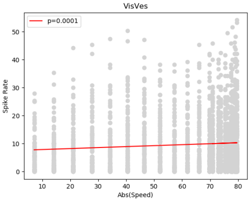


Keshavarzi et al., 2021

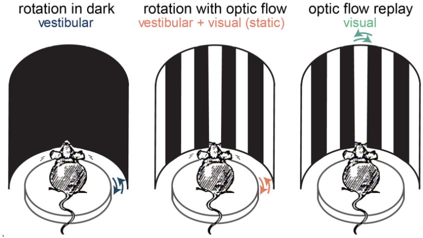
Linear regression example



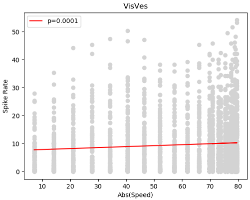
Keshavarzi et al., 2021



Linear regression example



Keshavarzi et al., 2021



Is there a linear relation between the speed of rotation and the firing rate of visual cells?

Linear regression model

simple linear regression model

$$y(x_i, \mathbf{w}) = w_0 + w_1 x_i = [1, x_i] \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = [\phi_0(x_i), \phi_1(x_i)] \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$
$$= \phi(x_i)^\mathsf{T} \mathbf{w}$$

polynomial regression model

$$y(x_i, \mathbf{w}) = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 = [1, x_i, x_i^2, x_i^3] \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$$
$$= [\phi_0(x_i), \phi_1(x_i), \phi_2(x_i), \phi_3(x_i)] \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} = \phi(x_i)^\mathsf{T} \mathbf{w}$$

basis functions regression model

$$y(x_i, \mathbf{w}) = \phi(x_i)^\mathsf{T} \mathbf{w} = \sum_{j=1}^{M} w_j \phi_j(x_i)$$

Linear regression model

$$\mathbf{y}(\mathbf{x},\mathbf{w}) = \begin{bmatrix} y(x_1,\mathbf{w}) \\ y(x_2,\mathbf{w}) \\ \dots \\ y(x_N,\mathbf{w}) \end{bmatrix} = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_M(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_M(x_2) \\ \vdots & \vdots & \dots & \vdots \\ \phi_1(x_N) & \phi_2(x_N) & \dots & \phi_M(x_N) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix}$$

$$= \mathbf{\Phi}\mathbf{w}$$

where $\mathbf{y}(\mathbf{x},\mathbf{w}) \in \mathbb{R}^N, \mathbf{\Phi} \in \mathbb{R}^{N\times M}, \mathbf{w} \in \mathbb{R}^M$.

Basis functions for regression

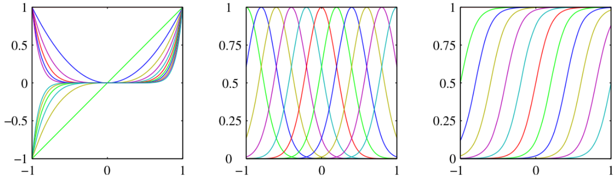


Figure 3.1 Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

Bishop (2016)

polynomial $\phi_i(x) = x^i$

Gaussian $\phi_i(x) = \exp(-\frac{(x-\mu_i)^2}{2\sigma^2})$

sigmoidal $\phi_i(x) = \frac{1}{1+\exp(-\frac{x-\mu_i}{\sigma^2})}$

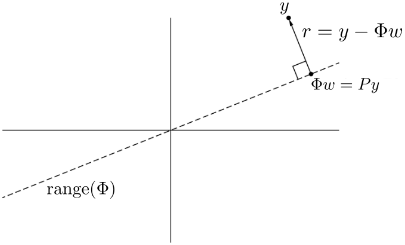
Least-squares estimation of model parameters (Trefethen and Bau III, 1997)

Definition 2 (Least-squares problem)

Given $\mathbf{\Phi} \in \mathbb{R}^{N\times M}$, $N \geq M$, $\mathbf{y} \in \mathbb{R}^N$, find $\mathbf{w} \in \mathbb{R}^M$ such that $||\mathbf{y} - \mathbf{\Phi w}||_2$ is minimized.

Theorem 5 (Least-squares solution)

Let $\mathbf{\Phi} \in \mathbb{R}^{N\times M}(N \geq M)$ and $\mathbf{y} \in \mathbb{R}^N$ be given. A vector $\mathbf{w} \in \mathbb{R}^M$ minimizes $||\mathbf{r}||_2 = ||\mathbf{y} - \mathbf{\Phi w}||_2$, thereby solving the least-squares problem, if and only if $\mathbf{r} \perp range(\mathbf{\Phi})$, that is, $\mathbf{\Phi}^\intercal \mathbf{r} = 0$, or equivalently, $\mathbf{\Phi}^\intercal\mathbf{\Phi w} = \mathbf{\Phi}^\intercal \mathbf{y}$, or again equivalently, $P\mathbf{y} = \mathbf{\Phi w}$.



References

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

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons.

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

Papoulis, A. and Pillai, S. U. (2002). *Probability, random variables and stochastic processes*. Mc Graw Hill, fourth edition.

Trefethen, L. n. and Bau III, D. (1997). Numerical linear algebra.

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.