# Normal Distributions 

Robert L. Wolpert<br>Department of Statistical Science<br>Duke University, Durham, NC, USA

## 1 Jeffreys

Let $p \in \mathbb{N}=\{1,2, \ldots\}, \mu \in \mathbb{R}^{p}$, and $\Lambda \in \mathcal{S}_{p}^{+}=\operatorname{HPD}(p)$, the space of $p \times p$ Hermitian Positive-Definite (hence symmetric) matrices, with inverse $\mathbb{E}=\Lambda^{-1}$. Then the log likelihood for a simple random sample $\left\{X_{i}\right\}_{i \leq n} \stackrel{\mathrm{iid}}{\sim} \operatorname{No}\left(\mu, \Lambda^{-1}\right)$ is:

$$
\begin{align*}
\ell(\mu, \Lambda \mid x) & =\frac{n}{2} \log |\Lambda|-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{p}\left(X_{i j}-\mu_{j}\right) \Lambda_{j k}\left(X_{i k}-\mu_{k}\right)+c \\
& =\frac{n}{2} \log |\Lambda|-\frac{1}{2} \operatorname{tr} S \Lambda+c \tag{1}
\end{align*}
$$

where $S=\left(X-\mathbf{1} \mu^{\prime}\right)^{\prime}\left(X-\mathbf{1} \mu^{\prime}\right)$ for a vector $\mathbf{1}=(1, \cdots, 1)^{\prime} \in \mathbb{R}^{n}$. Recall

$$
\begin{equation*}
|\Lambda|=\sum_{j=1}^{p} \Lambda_{j k} m_{j k}=\sum_{k=1}^{p} \Lambda_{j k} m_{j k} \quad \text { and } \quad \Lambda_{j k}^{-1}=|\Lambda|^{-1} m_{j k} \tag{2}
\end{equation*}
$$

where $M=\left\{m_{j k}\right\}$ is the cofactor matrix with entries $(-1)^{j-k}$ times the determinant $\left|\Lambda_{-j-k}\right|$ of the $(p-1) \times(p-1)$ matrix constructed by removing the $j^{\prime}$ th row and $k^{\prime}$ th column from $\Lambda$. By symmetry,

$$
\frac{\partial|\Lambda|}{\partial \Lambda_{j k}}=\frac{\partial}{\partial \Lambda_{j k}}\left[\Lambda_{j j} m_{j j}+2 \sum_{k>j} \Lambda_{j k} m_{j k}\right]= \begin{cases}2 m_{j k} & \text { if } j \neq k \\ m_{j j} & \text { o.w }\end{cases}
$$

and

$$
\frac{\partial \operatorname{tr} S \Lambda}{\partial \Lambda_{j k}}= \begin{cases}2 S_{j k} & \text { if } j \neq k \\ S_{j j} & \text { o.w. }\end{cases}
$$

so the $\Lambda$ score is ${ }^{1}$

$$
\begin{aligned}
Z_{\Lambda}:=\frac{\partial \ell}{\partial \Lambda} & =\frac{n}{2} \frac{1}{|\Lambda|}\left\{2|\Lambda| \Lambda^{-1}-\operatorname{diag}\left(|\Lambda| \Lambda^{-1}\right)\right\}-\frac{1}{2}\{2 S-\operatorname{diag}(S) \mid\} \\
& =\left(n \Lambda^{-1}-S\right)-\frac{1}{2} \operatorname{diag}\left(n \Lambda^{-1}-S\right)
\end{aligned}
$$

The Fisher Information for $\Lambda$ is the $\binom{p}{2} \times\binom{ p}{2}$ matrix whose $(j k),(\ell m)$ entry is $\mathrm{E}\left[Z_{\Lambda} Z_{\Lambda}^{\prime}\right]$. Press (1982, p.79) says its determinant is proportional to $|\Lambda|^{-(p+1)}$ but I haven't worked through that. The $p=2$ or $p=3$ cases might be worth trying - Volunteers? Similarly the $\mu$-score for fixed $\Lambda$ is

$$
\begin{equation*}
Z_{\mu}:=\frac{\partial}{\partial \mu} \ell=\Lambda\left(X-\mathbf{1} \mu^{\prime}\right)^{\prime} \mathbf{1} \tag{3}
\end{equation*}
$$

so the MLE is $\hat{\Lambda}=n S^{-1}$ (or $\hat{\Sigma}=\frac{1}{n} X^{\prime} \mathbf{1}$ ). The Fisher Information for $\mu$ is

$$
I_{\mu}=\mathrm{E}\left[-\frac{\partial}{\partial \mu} Z_{\mu}\right]=n \Lambda,
$$

a constant (in $\mu$ ), so the Jeffreys' Rule prior distribution for $\mu$ is the improper uniform distribution $\pi_{J}(d \mu) \propto d \mu$ on $\mathbb{R}^{p}$.

## References

Press, S. J. (1982), Applied Multivariate Analysis using Bayesian and Frequentist Methods of Inference, New York, NY: Dover, 2nd edition.

[^0]
## 2 Conjugate Prior: The Wishart Distribution

The likelihood for a sample $\mathbf{x}=\left\{X^{(m)}\right\}_{m \leq n} \subset \mathbb{R}^{p}$ of size $n \in \mathbb{N}$ is

$$
\begin{aligned}
f(\mathbf{x} \mid \mu, \Lambda) & =|2 \pi \mathbb{\Psi}|^{-n / 2} \exp \left\{-\frac{1}{2} \sum\left(X^{(m)}-\mu\right)^{\prime} \Lambda\left(X^{(m)}-\mu\right)\right\} \\
& =|\Lambda / 2 \pi|^{n / 2} \exp \left\{-\frac{1}{2} \sum\left(X^{(m)}-\bar{x}\right)^{\prime} \Lambda\left(X^{(m)}-\bar{x}\right)-\frac{n}{2}(\bar{x}-\mu)^{\prime} \Lambda(\bar{x}-\mu)\right\} \\
& \propto|\Lambda|^{\frac{1}{2}} \exp \left\{-\frac{n}{2}(\mu-\bar{x})^{\prime} \Lambda(\mu-\bar{x})\right\} \\
& \times|\Lambda|^{\frac{n-1}{2}} \exp \left\{-\frac{1}{2}\left(\sum\left(X^{(m)}-\bar{x}\right)^{\prime} \Lambda\left(X^{(m)}-\bar{x}\right)\right)\right\} \\
& =\left[|\Lambda|^{\frac{1}{2}} \exp \left\{-\frac{n}{2}(\mu-\bar{x})^{\prime} \Lambda(\mu-\bar{x})\right\}\right] \times|\Lambda|^{(n-1) / 2} \exp \left\{-\frac{1}{2} \operatorname{tr} \Lambda S\right\},
\end{aligned}
$$

where

$$
S=\sum_{m \leq n}\left(X^{(m)}-\bar{x}\right)\left(X^{(m)}-\bar{x}\right)^{\prime}=X^{\prime}\left[I_{n}-\frac{1}{n} \mathbf{1}_{n} \mathbf{1}_{n}^{\prime}\right] X
$$

where $\mathbf{1}_{n}$ is the $n \times 1$ column vector of ones and $I_{n}$ the $n \times n$ identity. A random matrix $W$ has the "Wishart distribution with scale $G \in \operatorname{HPD}(p)$ and $\nu>0$ degrees of freedom" (written $\left.W \sim \mathrm{Wi}_{p}(G, \nu)\right)$ if it has density function

$$
p(W \mid G, \nu) \propto|W|^{(\nu-p-1) / 2} \exp \left\{-\frac{1}{2} \operatorname{tr} G^{-1} W\right\}
$$

The proportionality constant is pretty ugly but almost never needed-

$$
\text { const }^{-1}=\pi^{p(p-1) / 4}|2 G|^{\nu / 2} \prod_{0 \leq j<p} \Gamma\left(\frac{\nu-j}{2}\right) .
$$

The density formula only works for $\nu>p$ degrees of freedom, but the distribution is well-defined for all $\nu>0$ (it just doesn't have a density for $\nu \leq p$ ). For example, it can be characterized by its characteristic function

$$
\mathrm{E}\{\exp (i \operatorname{tr} W \Theta)\}=|I-2 i \Theta G|^{-\nu / 2}
$$

If $\nu$ is an integer and $Y$ is a $\nu \times p$ matrix whose rows are independent $\operatorname{No}(0, G)$ random vectors, then

$$
Y^{\prime} Y \sim \mathrm{Wi}_{p}(G, \nu)
$$

The Wishart is a $p \times p$ generalization of the univariate $\chi_{\nu}^{2}=\mathrm{Ga}(\nu / 2,1 / 2)$ distribution.

### 2.1 A few properties

The matrix $G$ plays the role of a scale: if $W \sim \mathrm{Wi}_{p}(G, \nu)$ and $C$ is any $q \times p$ matrix then $U:=C W C^{\prime}$ is a $q \times q$ random matrix with distribution $U \sim \mathrm{Wi}_{q}\left(C G C^{\prime}, \nu\right)$. For $q=1$ with $C=u^{\prime}$ for some $u \in \mathbb{R}^{p}$, it follows that $u^{\prime} W u \sim \sigma^{2} \chi_{\nu}^{2}=\mathrm{Ga}\left(\frac{\nu}{2}, \frac{1}{2 \sigma^{2}}\right)$ where $\sigma^{2}=u^{\prime} G u$. If $G$ is full-rank and we choose a $p \times p$ square-root $C$ such that $C^{\prime} C=G^{-1}$, then $C W C^{\prime} \sim \mathrm{Wi}_{p}(I, \nu)$ has identity scale matrix; conversely, if we take $A$ such that $A A^{\prime}=G$ and take $U \sim \mathrm{Wi}_{p}(I, \nu)$, then $W:=A U A^{\prime} \sim \mathrm{Wi}_{p}(G, \nu)$.

If we partition $W \sim \mathrm{Wi}_{p}(G, \nu)$ into an upper-left $q \times q$ block for some $1 \leq q \leq p$,

$$
W=\left[\begin{array}{ll}
W_{11} & W_{12} \\
W_{21} & W_{22}
\end{array}\right]
$$

and similarly

$$
G=\left[\begin{array}{ll}
G_{11} & G_{12} \\
G_{21} & G_{22}
\end{array}\right]
$$

then

$$
W_{11} \sim \mathrm{Wi}_{q}\left(G_{11}, \nu\right)
$$

is independent of

$$
\begin{aligned}
W_{22 \cdot 1} & :=W_{22}-W_{21} W_{11}^{\dagger} W_{12} \\
& \sim \mathrm{Wi}_{p-q}\left(G_{22 \cdot 1}, \nu-p+q\right) .
\end{aligned}
$$

If we have several independent Wishart matrices $W_{\alpha} \sim \mathrm{Wi}_{p}\left(G, \nu_{\alpha}\right)$, all with the same dimension $p$ and scale $G \in \mathcal{S}_{p}^{+}$, then their element-wise sum $W_{+}:=\sum_{\alpha} W_{\alpha}$ is also Wishart: $W_{+} \sim \mathrm{Wi}_{p}\left(G, \nu_{+}\right)$, with $\nu_{+}:=\sum_{\alpha} \nu_{\alpha}$ degrees of freedom (this follows immediately from the ch.f. above).
The moments of $W \sim \mathrm{Wi}_{p}(G, \nu)$ are easily shown to be:

$$
\begin{aligned}
\mathrm{E}[W] & =\nu G \\
\mathrm{~V}\left[W_{i j}\right] & =\nu\left[g_{i j}^{2}+g_{i i} g_{j j}\right] \\
\operatorname{Cov}\left[W_{i j}, W_{k \ell}\right] & =\nu\left[g_{i k} g_{j \ell}+g_{i \ell} g_{j k}\right]
\end{aligned}
$$

To see this, either use the $W=\sum X_{\alpha} X_{\alpha}^{\prime}$ representation or use the $\log$ ch.f.

### 2.2 Bartlett Decomposition

Let $U \sim \mathrm{Wi}_{p}(I, \nu)$ have a Wishart distribution with identity scale, and let $U=L L^{\prime}$ be its Cholesky decomposition for a lower-triangular matrix

$$
L=\left[\begin{array}{cccccc}
\ell_{11} & 0 & 0 & 0 & \ldots & 0 \\
\ell_{21} & \ell_{22} & 0 & 0 & \ldots & 0 \\
\ell_{31} & \ell_{32} & \ell_{33} & 0 & \ldots & 0 \\
\ell_{41} & \ell_{42} & \ell_{43} & \ell_{44} & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\ell_{p 1} & \ell_{p 2} & \ell_{p 3} & \ell_{p 4} & \ldots & \ell_{p p}
\end{array}\right]
$$

then the $\left\{\ell_{i j}\right\}$ are all independent for $1 \leq i, j \leq p$, with

$$
\ell_{i j} \sim \begin{cases}0 & i<j \\ \chi_{\nu-i+1} & i=j \\ \operatorname{No}(0,1) & i>j\end{cases}
$$

where " $Y \sim \chi_{\nu}$ " means $Y^{2} \sim \chi_{\nu}^{2}$, i.e., the diagonal elements are square roots of $\chi^{2}$ variates. This is the preferred way to generate samples of $\mathrm{Wi}_{p}(G, \nu)$ random matrices- usually it's best to parametrize models (or algorithms) in terms of the Cholesky matrix $L$ instead of $U=L L^{\prime}$ or $W=A L L^{\prime} A^{\prime}$. Question: What happens if $\nu \leq p-1$ ? For example, what if $\nu=1$ but $p \geq 2$ ?

### 2.3 Hotelling's $T^{2}$

Recall that the ratio $t:=Z / \sqrt{Y / \nu}$ of a standard normal $Z \sim \operatorname{No}(0,1)$ random variable to the square root of a chi-squared variate $Y \sim \chi_{\nu}^{2}$, divided by its degrees of freedom $\nu$, has a "Student $t$ distribution with $\nu$ degrees of freedom." First proved by William Gossett as part of his work for the Guinness Brewery, this result is the key to testing hypotheses and constructing confidence intervals for the mean $\mu$ of univariate normally-distributed data $X_{\alpha} \stackrel{\text { iid }}{\sim} \operatorname{No}\left(\mu, \sigma^{2}\right)$ when the variance $\sigma^{2}$ is unknown, since it is a "pivotal quantity" (one whose distribution doesn't depend on $\mu$ or $\sigma^{2}$ )

$$
t:=\frac{(X-\bar{X})}{\sqrt{S /(n-1)}}
$$

for $\left\{X_{\alpha}\right\}_{1 \leq \alpha \leq n} \stackrel{\text { iid }}{\sim} \operatorname{No}\left(\mu, \sigma^{2}\right)$ with $S:=\sum\left(X_{\alpha}-\bar{X}\right)^{2}$ and $\nu=n-1$.
The square of the $t$ ratio,

$$
t^{2}=\frac{Z^{2}}{Y^{2} / \nu}
$$

is the ratio of two independent $\chi^{2}$ variables divided by their degrees of freedom, with 1 degree of freedom in the numerator and $\nu$ in the denominator, so $t^{2} \sim F_{\nu}^{1}$ has Snedecker's $F$ distribution.
Similarly, for a standard $p$-variate normal vector $Z \sim \mathrm{No}_{p}\left(0, I_{p}\right)$ and Wishart $W \sim \mathrm{Wi}_{p}\left(I_{p}, \nu\right)$, the quantity

$$
T^{2}:=\nu Z^{\prime} W^{-1} Z
$$

is said to have a "Hotelling's $T^{2}$ distribution." With a change of variables to $X:=\mu+A Z \sim$ $\mathrm{No}(\mu, G)$ and $M:=A W A^{\prime} \sim \mathrm{Wi}_{p}(G, \nu)$ with $G=A A^{\prime}$, one can show that

$$
T^{2}=\nu(X-\mu)^{\prime} M^{-1}(X-\mu)
$$

again has the $T_{\nu}^{2}$ distribution with $X \sim \operatorname{No}(\mu, G)$ and $M \sim \mathrm{Wi}_{p}(G, \nu)$ independent, so $T^{2}$ is also pivotal. Again there is a connection with the $F$ distribution: $T_{\nu}^{2}=\frac{\nu p}{\nu-p+1} F_{\nu-p+1}^{p}$, so

$$
\frac{\nu-p+1}{p}(x-\mu)^{\prime} M^{-1}(x-\mu) \sim F_{\nu-p+1}^{p} .
$$


[^0]:    ${ }^{1}$ Thanks to Anirban Bhattacharya and Jianyu Wang for this calculation.

