# Really Useful Proofs and Identities

Pete Bunch

October 16, 2012

## 1 Schur Complement Decomposition

This is an identity for the inverse of a block matrix which arises by doing block-wise Gaussian elimination.

$$\begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2}\\ \Sigma_{2,1} & \Sigma_{2,2}\end{bmatrix}^{-1} = \begin{bmatrix} I & 0\\ -\Sigma_{2,2}^{-1}\Sigma_{2,1} & I\end{bmatrix}\begin{bmatrix}\left(\Sigma_{1,1}-\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\right)^{-1} & 0\\ 0 & \Sigma_{2,2}^{-1}\end{bmatrix}\begin{bmatrix} I & -\Sigma_{1,2}\Sigma_{2,2}^{-1}\\ 0 & I\end{bmatrix}$$

Unsurprisingly, you can also do this the "other way round". The term $\left(\Sigma_{1,1}-\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\right)^{-1}$ is called the Schur complement. The two outer matrices are invertible and have determinant 1.

I use this most commonly on covariance matrices, which are symmetric, so a more useful but less general form is as follows.

$$\begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2}\\ \Sigma_{1,2}^{T} & \Sigma_{2,2}\end{bmatrix}^{-1} = \begin{bmatrix} I & 0\\ -\Sigma_{2,2}^{-1}\Sigma_{1,2}^{T} & I\end{bmatrix}\begin{bmatrix}\left(\Sigma_{1,1}-\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{1,2}^{T}\right)^{-1} & 0\\ 0 & \Sigma_{2,2}^{-1}\end{bmatrix}\begin{bmatrix} I & -\Sigma_{1,2}\Sigma_{2,2}^{-1}\\ 0 & I\end{bmatrix}$$

## 2 Woodbury Identity

Also known as the block matrix inversion formula or the matrix inversion lemma. This is derived by doing two Schur complement decompositions, one each way round, and then equating a diagonal block. Or by just multiplying the right hand side by the inverse of the left.

$$\left(A+DBC\right)^{-1} = A^{-1} - A^{-1}D\left(B^{-1}+CA^{-1}D\right)^{-1}CA^{-1}$$

Again, I use this most often when manipulating Gaussian, in which case $D=C^{T}$, so I like the following form.

$$\left(A+C^{T}BC\right)^{-1} = A^{-1} - A^{-1}C^{T}\left(B^{-1}+CA^{-1}C^{T}\right)^{-1}CA^{-1}$$

## 3 Gaussian Identities

The following notation is used for Gaussians.

$$\mathcal{N}(x|m,P) = ||2\pi P||^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[(x-m)^{T}P^{-1}(x-m)\right]\right\}dx$$

## 3.1 The Gaussian Integral

This is fundamental.

$$\int \exp\left\{-\frac{1}{2}\left[(x-m)^T P^{-1}(x-m)\right]\right\} dx = ||2\pi P||^{\frac{1}{2}}$$

## 3.2 Completing the Square

This is just a re-jigging of the Gaussian integral which is a super useful shortcut.

$$\begin{aligned}
&\int \exp\left\{-\frac{1}{2}\left[x^T \Upsilon x - 2\lambda^T x + c\right]\right\} dx \\
&= \int \exp\left\{-\frac{1}{2}\left[(x-\Upsilon^{-1}\lambda)^T \Upsilon (x-\Upsilon^{-1}\lambda) - \lambda^T \Upsilon^{-1}\lambda + c\right]\right\} dx \\
&= \left|\left|2\pi\Upsilon^{-1}\right|\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[c - \lambda^T \Upsilon^{-1}\lambda\right]\right\}
\end{aligned}$$

## 3.3 Inverting a Gaussian

Consider a Gaussian density over a variable $y$ conditional on a variable $x$, where $A$ is an invertible matrix. This can be rewritten as an unnormalised Gaussian density over $x$ conditional on $y$.

$$\mathcal{N}(y|Ax, Q) = ||A||^{-1} \mathcal{N}(x|A^{-1}y, A^{-1}QA^{-T})$$

Proof:

$$\begin{aligned}
\mathcal{N}(y|Ax, Q) &= ||2\pi Q||^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(y-Ax)^T Q^{-1}(y-Ax)\right]\right\} \\
&= ||2\pi Q||^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(x-A^{-1}y)^T A^T Q^{-1} A(x-A^{-1}y)\right]\right\} \\
&= \frac{\left|\left|2\pi\left(A^T Q^{-1} A\right)^{-1}\right|\right|^{\frac{1}{2}}}{||2\pi Q||^{\frac{1}{2}}} \mathcal{N}(x|A^{-1}y, \left(A^T Q^{-1} A\right)^{-1}) \\
&= ||A||^{-1} \mathcal{N}(x|A^{-1}y, A^{-1}QA^{-T})
\end{aligned}$$

## 3.4 Conditioning a Gaussian

This uses the Schur complement decomposition on the covariance matrix of a joint Gaussian.

$$\begin{aligned}
\mathcal{N}(x|m, P) &= \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} P_1 & C \\ C^T & P_2 \end{bmatrix}\right) \\
&= \mathcal{N}(x_1|m_1 + CP_2^{-1}(x_2 - m_2), P_1 - CP_2^{-1}C^T)\mathcal{N}(x_2|m_2, P_2)
\end{aligned}$$

Proof:

$$
\begin{aligned}
\mathcal{N}(x|m,\,P) &= \mathcal{N}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\middle|\begin{bmatrix}m_1\\m_2\end{bmatrix},\begin{bmatrix}P_1 & C\\ C^T & P_2\end{bmatrix}\right)\\
&= \left|\left|2\pi\begin{bmatrix}P_1 & C\\ C^T & P_2\end{bmatrix}^{-1}\right|\right|^{1/2}\exp\left\{-\frac{1}{2}\begin{bmatrix}x_1-m_1\\x_2-m_2\end{bmatrix}^T\begin{bmatrix}P_1 & C\\ C^T & P_2\end{bmatrix}^{-1}\begin{bmatrix}x_1-m_1\\x_2-m_2\end{bmatrix}\right\}\\
&= \left|\left|2\pi\begin{bmatrix}(P_1-CP_2^{-1}C^T)^{-1} & 0\\ 0 & P_2^{-1}\end{bmatrix}\right|\right|^{1/2}\\
&\quad\times\exp\left\{-\frac{1}{2}\begin{bmatrix}x_1-m_1\\x_2-m_2\end{bmatrix}^T\begin{bmatrix}I & 0\\ -P_2^{-1}C^T & I\end{bmatrix}\begin{bmatrix}(P_1-CP_2^{-1}C^T)^{-1} & 0\\ 0 & P_2^{-1}\end{bmatrix}\begin{bmatrix}I & -CP_2^{-1}\\ 0 & I\end{bmatrix}\begin{bmatrix}x_1-m_1\\x_2-m_2\end{bmatrix}\right\}\\
&= \left|\left|2\pi(P_1-CP_2^{-1}C^T)\right|\right|^{-1/2}\exp\left\{-\frac{1}{2}\left[(x_1-m_1)-CP_2^{-1}(x_2-m_2)\right]^T(P_1-CP_2^{-1}C^T)^{-1}\left[(x_1-m_1)-CP_2^{-1}(x_2-m_2)\right]\right\}\\
&\quad\times\left|\left|2\pi P_2\right|\right|^{-1/2}\exp\left\{-\frac{1}{2}\left[(x_2-m_2)^TP_2^{-1}(x_2-m_2)\right]\right\}\\
&= \mathcal{N}(x_1|m_1+CP_2^{-1}(x_2-m_2),\,P_1-CP_2^{-1}C^T)\mathcal{N}(x_2|m_2,\,P_2)
\end{aligned}
$$

This can also be done the "other way around".

## 3.5 Unconditioning a Gaussian

We can reverse the conditioning process to give us this handy little formula.

$$\mathcal{N}(y|Ax+b,Q)\mathcal{N}(x|m,P)=\mathcal{N}\left(\begin{bmatrix}y\\x\end{bmatrix}\middle|\begin{bmatrix}Am+b\\m\end{bmatrix},\begin{bmatrix}Q+APA^T & AP\\ PA^T & P\end{bmatrix}\right)$$

Proof:

$$\mathcal{N}\left(\begin{bmatrix}y\\x\end{bmatrix}\middle|\begin{bmatrix}\zeta\\\xi\end{bmatrix},\begin{bmatrix}\Sigma_1 & \Sigma_{1,2}\\ \Sigma_{1,2}^T & \Sigma_2\end{bmatrix}\right)=\mathcal{N}(y|\zeta+\Sigma_{1,2}\Sigma_2^{-1}(x-\xi),\Sigma_1-\Sigma_{1,2}\Sigma_2^{-1}\Sigma_{1,2}^T)\mathcal{N}(x|\xi,\Sigma_2)$$

First equate the easy terms.

$$\begin{aligned}\xi&=m\\ \Sigma_2&=P\end{aligned}$$

Now equate the means and rearrange.

$$\begin{aligned}\zeta+\Sigma_{1,2}P^{-1}(x-m)&=Ax+b\\ (A-\Sigma_{1,2}P^{-1})x&=\zeta-\Sigma_{1,2}P^{-1}m-b\end{aligned}$$

For a valid Gaussian, $\zeta$ must be independent of $x$, which means we can split this up.

$$\begin{aligned}\Sigma_{1,2}&=AP\\ \xi&=Am+b\end{aligned}$$

Finally, equate the variances.

$$\Sigma_1=APA^T+Q$$

And that's it.

### 3.6 Switching a Gaussian Product

If we use the unconditioning formula then the conditioning formula, we get the following switching trick. Useful for a Kalman update.

$$\begin{aligned}&\mathcal{N}(y|Ax+b,Q)\mathcal{N}(x|m,P)\\&=\mathcal{N}(y|Am+b,APA^T+Q)\mathcal{N}(x|m+PA^T(APA^T+Q)^{-1}(y-Am-b),P-PA^T(APA^T+Q)^{-1}AP)\end{aligned}$$

### 3.7 Marginalising a Gaussian Product

This one follows by switching then doing the integral. Useful for a Kalman prediction.

$$\int \mathcal{N}(y|Ax+b,Q)\mathcal{N}(x|m,P)dx=\mathcal{N}(y|Am+b,APA^T+Q)$$

### 3.8 Representing a Degenerate Gaussian

If a Gaussian has a degenerate covariance matrix, then the standard form for the density does not work, because the determinant is zero. For many proofs and derivations, it is then helpful to write the density in the following form.

$$\mathcal{N}(x|m,P)=\int \mathcal{N}(\epsilon|0,1)\delta_{m+F\epsilon}(x)d\epsilon$$

$P=FF^T$ is a rank factorisation of the covariance matrix. So $F$ is a "tall" matrix.

Proof:

$$\begin{aligned}\int \mathcal{N}(\epsilon|0,1)\delta_{m+F\epsilon}(x)d\epsilon&=\lim_{h\to 0}\int \mathcal{N}(\epsilon|0,1)\mathcal{N}(x|m+F\epsilon,hI)d\epsilon\\&=\lim_{h\to 0}\mathcal{N}(x|m,FF^T+hI)\\&=\mathcal{N}(x|m,P)\end{aligned}$$

## 4 Kalman Filters

### 4.1 Exact Kalman Filter for Linear Gaussian Models

The Kalman filter is the solution to the Bayesian filtering recursions for linear Gaussian dynamics. The transition and observation densities are Gaussian with linear functions for means.

$$\begin{aligned}p(x_k|x_{k-1})&=\mathcal{N}(x_k|Ax_{k-1},Q)\\p(y_k|x_k)&=\mathcal{N}(y_k|Cx_k,R)\end{aligned}$$

Assume the filtering and predictive filtering densities take the following forms.

$$\begin{aligned}p(x_k|y_{1:k})&=\mathcal{N}(x_k|m_k,P_k)\\p(x_k|y_{1:k-1})&=\mathcal{N}(x_k|\hat{m}_k,\hat{P}_k)\end{aligned}$$

The Bayesian filtering recursions are as follows.
Prediction:

$$\begin{aligned} p(x_k|y_{1:k-1}) &= \int p(x_k|x_{k-1})p(x_{k-1}|y_{1:k-1})dx_{k-1} \\ &= \int \mathcal{N}(x_k|Ax_{k-1}, Q)\mathcal{N}(x_{k-1}|m_{k-1}, P_{k-1}) \\ &= \mathcal{N}(x_k|Am_{k-1}, AP_{k-1}A^T + Q) \end{aligned}$$

Update:

$$\begin{aligned} p(x_k|y_{1:k}) &= \frac{p(y_k|x_k)p(x_k|y_{1:k})}{\int p(y_k|x_k)p(x_k|y_{1:k})dx_{k-1}} \\ &= \frac{\mathcal{N}(y_k|Cx_k, R)\mathcal{N}(x_k|\hat{m}_k, \hat{P}_k)}{\int \mathcal{N}(y_k|Cx_k, R)\mathcal{N}(x_k|\hat{m}_k, \hat{P}_k)dx_k} \\ &= \frac{\mathcal{N}(y|C\hat{m}_k, C\hat{P}_kC^T + R)\mathcal{N}(x_k|m_k, P_k)}{\mathcal{N}(y|C\hat{m}_k, C\hat{P}_kC^T + R)} \\ &= \mathcal{N}(x_k|m_k, P_k), \end{aligned}$$

where the updated mean and covariance are given by,

$$\begin{aligned} m_k &= \hat{m}_k + \hat{P}_kC^T(C\hat{P}_kC^T + R)^{-1}(y - C\hat{m}_k) \\ P_k &= \hat{P}_k - \hat{P}_kC^T(C\hat{P}_kC^T + R)^{-1}C\hat{P}_k. \end{aligned}$$

Hence the derivation of the Kalman filter is simply an application of the marginalisation formula (twice) and the switching formula (once).

## 4.2 Extended Kalman Filter for Nonlinear-Gaussian Models

The Kalman filter can be modified for use with nonlinear systems of the following form.

$$\begin{aligned} p(x_k|x_{k-1}) &= \mathcal{N}(x_k|f(x_{k-1}), Q) \\ p(y_k|x_k) &= \mathcal{N}(y_k|h(x_k), R) \end{aligned}$$

The solution is approximate. We replace $f(x_{k-1})$ and $h(x_k)$ with first order Taylor series approximations about the points $\bar{x}_{k-1}$ and $\bar{x}_k$ respectively.

$$\begin{aligned} f(x_{k-1}) &\approx f(\bar{x}_{k-1}) + \underbrace{\left.\frac{\partial f}{\partial x_{k-1}}\right|_{x_{k-1}=\bar{x}_{k-1}}}_{F_k}(x_{k-1} - \bar{x}_{k-1}) \\ h(x_k) &\approx h(\bar{x}_k) + \underbrace{\left.\frac{\partial h}{\partial x_k}\right|_{x_k=\bar{x}_k}}_{H_k}(x_k - \bar{x}_k) \end{aligned}$$

Assume the filtering and predictive filtering densities take the following forms.

$$p(x_k|y_{1:k}) = \mathcal{N}(x_k|m_k, P_k)$$
$$p(x_k|y_{1:k-1}) = \mathcal{N}(x_k|\hat{m}_k, \hat{P}_k)$$

The Bayesian filtering recursions are as follows.
Prediction:

$$\begin{aligned} p(x_k|y_{1:k-1}) &= \int p(x_k|x_{k-1})p(x_{k-1}|y_{1:k-1})dx_{k-1} \\ &\approx \int \mathcal{N}(x_k|f(\bar{x}_{k-1}) + F_k(x_{k-1} - \bar{x}_{k-1}), Q)\mathcal{N}(x_{k-1}|m_{k-1}, P_{k-1}) \\ &= \mathcal{N}(x_k|f(\bar{x}_{k-1}) + F_k(m_{k-1} - \bar{x}_{k-1}), F_k P_{k-1} F_k^T + Q) \\ &= \mathcal{N}(x_k|f(\bar{x}_{k-1}), F_k P_{k-1} F_k^T + Q) \end{aligned}$$

The last line follows if we choose $\bar{x}_{k-1} = m_{k-1}$, which is a natural choice.
Update:

$$\begin{aligned} p(x_k|y_{1:k}) &= \frac{p(y_k|x_k)p(x_k|y_{1:k})}{\int p(y_k|x_k)p(x_k|y_{1:k})dx_{k-1}} \\ &\approx \frac{\mathcal{N}(y_k|h(\bar{x}_k) + H_k(x_k - \bar{x}_k), R)\mathcal{N}(x_k|\hat{m}_k, \hat{P}_k)}{\int \mathcal{N}(y_k|h(\bar{x}_k) + H_k(x_k - \bar{x}_k), R)\mathcal{N}(x_k|\hat{m}_k, \hat{P}_k)dx_k} \\ &= \mathcal{N}(x_k|m_k, P_k), \end{aligned}$$

where the updated mean and covariance are given by,

$$\begin{aligned} m_k &= \hat{m}_k + \hat{P}_k H_k^T (H_k \hat{P}_k H_k^T + R)^{-1}(y - h(\bar{x}_k) + H_k(\bar{x}_k - \hat{m}_k)) \\ P_k &= \hat{P}_k - \hat{P}_k H_k^T (H_k \hat{P}_k H_k^T + R)^{-1} H_k \hat{P}_k. \end{aligned}$$

As before, the mean update can be simplified by choosing $\bar{x}_k = \hat{m}_k$.

# 5 Conjugate Priors

## 5.1 Unknown Variance of a Univariate Gaussian

We want to estimate the variance of a Gaussian density for a random state variable.

$$p(x|\sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$$

The conjugate prior for the unknown variance of a Gaussian, $\sigma^2$ is an inverse gamma distribution. However, the gamma distribution is better documented and also (crucially) has a built in MATLAB implementation, so it's preferable to deal with that. Therefore we define a reciprocal variable.

$$\tau = \frac{1}{\sigma^2}$$

Here we work with the shape/scale parameterisation of the gamma distribution.

$$p(\tau) \propto \tau^{-a_0-1} \exp\left\{-\frac{b_0}{\tau}\right\}$$

The posterior is also a gamma distribution.

$$\begin{aligned} p(\tau|x_{1:N}) &= \prod_{i=1}^{N} p(x_i|\tau)p(\tau) \\ &\propto \prod_{i=1}^{N} \left[\tau^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\tau(x_i-\mu_i)^2\right\}\right] \tau^{-a_0-1} \exp\left\{-\frac{b_0}{\tau}\right\} \\ &= \tau^{a_0+\frac{N}{2}-1} \exp\left\{-\frac{1}{\tau}\left[\frac{1}{b_0} + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu_i)^2\right]\right\} \end{aligned} \tag{1}$$

The update rule is as follows.

$$\begin{aligned} a &= a_0 + \frac{N}{2} \\ \frac{1}{b} &= \frac{1}{b_0} + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu_i)^2 \end{aligned}$$