Changing the Basis of distributions within the exponential family

Marius Hobbhahn
Department of Computer Science
University of Tübingen
Tübingen, Germany
marius.hobbhahn@gmail.com

1 Introduction

normal distributions are the best, lets transform everything to normal

2 Background

2.1 Change of Variable for Probability Density Function

1D

Let X have a continuous density f_X. Let $g : \mathbb{R} \to \mathbb{R}$ be piece-wise strictly monotone and continuously differentiable, i.e. there exists intervals $I_1, I_2, ..., I_n$ which partition $\mathbb{R}$ such that g is strictly monotone and continuously differentiable on the interior of each I_i. For each $i, g : I_i \to \mathbb{R}$ is invertible on $g(I_i)$ and let g_i^{-1} be the inverse function. Let $Y = g(X)$ and $\wedge = \{y|y = g(x), x \in \mathbb{R}\}$ be the range g. Then the density function f_Y of Y exists and is given by

$$f_Y(y) = \sum_{i=1}^{n} f_X(g_i^{-1}(y)) \left| \frac{\partial g_i^{-1}(y)}{\partial y} \right| \mathbf{1}_\wedge(y) \tag{1}$$

Higher dim

TODO: Jacobian. But how do the intervals work?

2.2 Laplace Approximation

The Laplace approximation (LPA) is a tool to fit a normal distribution to the PDF of a given other distribution. The only constraints for the other distribution are: one peak (mode/ point of maximum) and twice differentiable. Laplace proposed a simple 2-term Taylor expansion on the log pdf. If $\hat{\theta}$ denotes the mode of a pdf $h(\theta)$, then it is also the mode of the log-pdf $q(\theta) = \log h(\theta)$. The 2-term Taylor expansion of $q(\theta)$ therefore is:

$$\begin{aligned} q(\theta) &\approx q(\hat{\theta}) + q'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})q''(\hat{\theta})(\theta - \hat{\theta}) && (2) \\ &= q(\hat{\theta}) + 0 + \frac{1}{2}(\theta - \hat{\theta})q''(\hat{\theta})(\theta - \hat{\theta}) \qquad [\text{since } q'(\theta) = 0] && (3) \\ &= c - \frac{(\theta - \mu)^2}{2\sigma^2} && (4) \end{aligned}$$

where c is a constant, $\mu = \hat{\theta}$ and $\sigma^2 = \{-q''(\hat{\theta})\}^{-1}$. The right hand side of the last line matches the log-pdf of a normal distribution $N(\mu, \sigma^2)$. Therefore the pdf $h(\theta)$ is approximated by the pdf of the normal distribution $N(\mu, \sigma^2)$ where $\mu = \hat{\theta}$ and $\sigma^2 = \{-q''(\hat{\theta})\}^{-1}$. Note, that even though this derivation is done for the one dimensional case only, it is also true for the multidimensional case. The second derivative just becomes the Hessian of the pdf at the mode.

2.3 Exponential Family

A pdf that can be written in the form

$$p(x|w) = h(x) \cdot \exp\left(\phi(x)^\top w - \log Z(w)\right), \qquad \text{where} \qquad Z(w) := \int_{\mathbf{X}} \exp\left(\phi(x)^\top w\right) \, dh(x)$$

is called an exponential family where $\phi(x) : \mathbb{X} \to \mathbb{R}^d$ is called sufficient statistics, $w \in \mathbb{R}^d$: natural parameters, $\log Z(w) : \mathbb{R}^d \to \mathbb{R}$: (log) partition function (normalization constant), and $h(x) : \mathbb{X} \to \mathbb{R}_+$: base measure.

2.4 Chi2 <-> Normal

It is already well-known that the Chi-squared distribution describes the sum of independent, standard normal random variables. To introduce a certain 'trick' we show the forward and backward transformation between chi2 and normal.
Let X be normal with $\mu = 0, \sigma^2 = 1$. Let $Y = X^2$ and therefore $g(x) = x^2$, which is neither monotone nor injective. Take $I_1 = (-\infty, 0)$ and $I_2 = [0, +\infty)$. Then g is monotone and injective on I_1 and I_2 and $I_1 \cup I_2 = \mathbb{R}$. $g(I_1) = (0, \infty)$ and $g(I_2) = [0, \infty)$. Then $g_1^{-1} : [0, \infty) \to \mathbb{R}$ by $g_1^{-1}(y) = -\sqrt{y}$ and $g_2^{-1} : [0, \infty) \to \mathbb{R}$ by $g_2^{-1}(y) = \sqrt{y}$. Then

$$\left| \frac{\partial g_i^{-1}(y)}{\partial y} \right| = \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{2\sqrt{y}}$$

Applying Equation 1 we can transform a normal distribution to a chi-squared.

$$\begin{aligned} f_Y(y) &= f_X(g_1^{-1}(y)) \left| \frac{\partial g_1^{-1}(y)}{\partial y} \right| \mathbf{1}_{\wedge}(y) + f_X(g_2^{-1}(y)) \left| \frac{\partial g_2^{-1}(y)}{\partial y} \right| \mathbf{1}_{\wedge}(y) \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{y}{2}) \frac{1}{2\sqrt{y}} + \frac{1}{\sqrt{2\pi}} \exp(-\frac{y}{2}) \frac{1}{2\sqrt{y}} \qquad (y > 0) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} \exp(-\frac{y}{2}) \end{aligned} \tag{5}$$

The 'trick' was to split up the variable transformation in two parts to adjust for the fact that the space of the chi-squared and the Normal are different. We can reverse the same procedure to get from a chi-squared to a normal distribution. We keep the variable names from before. Let $X = \sqrt{Y}$ and therefore $h(x) = \sqrt{x}$. Then $h_1^{-1} : \mathbb{R} \to (-\infty, 0)$ by $h_1^{-1}(x) = -x^2$ and $h_2^{-1} : \mathbb{R} \to [0, \infty)$ by $h_2^{-1}(x) = x^2$. Then

$$\left| \frac{\partial h_i^{-1}(y)}{\partial y} \right| = |2y|$$

and

$$\begin{aligned} f_X(x) &= f_y(h_1^{-1}(x)) \left| \frac{\partial h_1^{-1}(y)}{\partial y} \right| \mathbf{1}_{\wedge}(y) + f_y(h_2^{-1}(x)) \left| \frac{\partial h_2^{-1}(y)}{\partial y} \right| \mathbf{1}_{\wedge}(y) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{2\sqrt{x^2}} \exp(-\frac{x^2}{2}) |2x| \mathbf{1}_{(-\infty,0)}(x) + \frac{1}{\sqrt{2\pi}} \frac{1}{2\sqrt{x^2}} \exp(-\frac{x^2}{2}) |2x| \mathbf{1}_{[0,\infty)}(x) \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \end{aligned} \tag{6}$$

which is defined on the entirety of $\mathbb{R}$.

2.4.1 A generalization to matrices

"In general, an $n \times n$ matrix with n distinct nonzero eigenvalues has 2^n square roots. Such a matrix, A, has a decomposition VDV^{-1} where V is the matrix whose columns are eigenvectors of A and D is the diagonal matrix whose diagonal elements are the corresponding n eigenvalues λ_i. Thus the square roots of A are given by $VD^{\frac{1}{2}}V{-}1$, where $D^{\frac{1}{2}}$ is any square root matrix of D, which, for distinct eigenvalues, must be diagonal with diagonal elements equal to square roots of the diagonal elements of D; since there are two possible choices for a square root of each diagonal element of D, there are 2^n choices for the matrix $D^{\frac{1}{2}}$" - wikipedia.

The Chi-squared distribution can be seen as the 1D special case of the matrix transformation and therefore has $2^1 = 2$ possible options for $\lambda^{\frac{1}{2}}$, namely $-\lambda^{\frac{1}{2}}$ and $+\lambda^{\frac{1}{2}}$. A higher dimensional functions such as Wishart and inverse Wishart have 2^n different possibilities which have to be accounted for within the transformation.

An alternative way to see the variable transform

Another way to interpret the transformation from Chi2 to normal and back is by introducing the sign function. Say we have a variable X which is distributed according to a Gaussian with $\mu = 0, \sigma = 1$ as in the above setting. We introduce the following transformation $g(a, b)$:

$$\begin{aligned}
A &= \text{sign}(X) \\
B &= |X| \\
W &= \text{sign}(X) \\
Z &= \sqrt{|X|} \cdot |W|
\end{aligned}$$

with the inverse $g^{-1}(w, z)$.

The determinant of the Jacobian of this transformation is as follows

$$\begin{aligned}
\det(J) &= \left| \begin{bmatrix} \frac{\partial w}{\partial a} & \frac{\partial w}{\partial b} \\ \frac{\partial z}{\partial a} & \frac{\partial z}{\partial b} \end{bmatrix} \right| \\
&= \left| \begin{bmatrix} 1 & 0 \\ 0 & \frac{|w|}{2\sqrt{z}} \end{bmatrix} \right| \\
&= \frac{|w|}{2\sqrt{z}}
\end{aligned}$$

The joint distribution of w and z can now be seen as

$$\begin{aligned}
f_{w,z}(w, z) &= f_{a,b}(g^{-z}(w, z)) \det(J) \\
&= f_a(w) f_b(w, z) \frac{|w|}{2\sqrt{z}} \\
&= f_a(w) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{|w|^2 z}{2}\right) \frac{|w|}{2\sqrt{z}}
\end{aligned}$$

Since we know that W can only take the values -1 and 1 we can easily integrate over W.

$$\begin{aligned}
f_z &= \int dw f_{w,z} \\
&= \int dw f_a(w) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{|w|^2 z}{2}\right) \frac{|w|}{2\sqrt{z}} \\
&= 1 \cdot \int dw \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{|w|^2 z}{2}\right) \frac{|w|}{2\sqrt{z}} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{|-1|^2 z}{2}\right) \frac{|-1|}{2\sqrt{z}} + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{|1|^2 z}{2}\right) \frac{|1|}{2\sqrt{z}} \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{z}} \exp\left(-\frac{z}{2}\right) \qquad z > 0
\end{aligned}$$

Similarly we can construct a transformation for the inverse transformation, i.e. from Chi2 to Gaussian. Assume we have a variable X which is distributed according to a Chi2 distribution. Our transform is

$$\begin{aligned}
A &= \text{sign}(X) \\
B &= X \\
W &= \text{sign}(X) \\
Z &= X^2 \cdot \sqrt{(}W)
\end{aligned}$$

Then we get a determinant of the Jacobian

$$\begin{aligned}
\det(J) &= \left| \begin{bmatrix} \frac{\partial w}{\partial a} & \frac{\partial w}{\partial b} \\ \frac{\partial z}{\partial a} & \frac{\partial z}{\partial b} \end{bmatrix} \right| \\
&= \left| \begin{bmatrix} 1 & 0 \\ 0 & 2z\sqrt{w} \end{bmatrix} \right| \\
&= 2z\sqrt{w}
\end{aligned}$$

And the joint distribution becomes

$$\begin{aligned}
f_{w,z} &= f_{a,b}(g^{-z}(w,z)) \det(J) \\
&= f_a(w) f_b(w,z) 2z\sqrt{w} \\
&= f_a(w) \frac{1}{z^2 w} \exp\left(-\frac{z^2 w}{2}\right) 2z\sqrt{w}
\end{aligned}$$

We can marginalize over W as follows. Note that X is distributed according to a Chi2 and therefore only allows positive values. W is therefore only 1.

TODO: this kinda feels like cheating AND has to be refined. No variables should be assigned twice.

3 Laplace Propagation

If we have a distribution $p(x)$ in the 'standard base' x, i.e. as we usually use it, we can transform this distribution to another basis y via the change of variable for PDF formula. The resulting distribution is $p_y(x(y))$ where $x(y)$ is the inverse transform of denoted as $g^{-1}(x)$.

We define Laplace Propagation as the procedure of finding a different base y for $p(x)$ in which $p_y(x(y))$ is as close to a Gaussian as possible. In this base a Laplace approximation is performed to yield a Gaussian $q(y)$. The mean μ for q is computed by setting the first derivative $\frac{\partial p_y(x(y))}{\partial y}$ to 0 and solving for y. The covariance matrix Σ is computed by inverting the Hessian $H = \frac{\partial^2 p(x(y))}{\partial^2 y}$, multiplying it with -1 and inserting the mode μ. Lastly, a 'Bridge' is created that transforms the parameters θ of $p_(x(y))$ to the parameters μ, Σ of $q(x(y))$.

TODO: the part above should stand out in some way. Maybe we should put a fancy box around it

While there are no restrictions for the choice of the basis transform (e.g. could be done by a Neural Network), in this paper we choose transforms that fulfill the following desiderata as much as possible: a) the new space of the distribution is $\mathbb{R}$ (or the $d-$dimensional equivalent) since this implies that the resultung Gaussian from the Laplace approximation is well-defined. b) The sufficient statistics in the new basis are either x in the first entry or x^2 in the second entry since the sufficient statistics of the Gaussian are (x, x^2) and it is therefore likely closer to the Normal distribution. c) The base measure $h(x)$ is 1 since this implies that there are no resctrictions on the parameters θ of the distribution in the new base. In the standard basis the Beta distribution's parameters α, β have to be larger than 1 to yield a uni-modal distribution. Multi-modal distributions yield bad or no valid Laplace approximations and we would therefore prefer $h(x) = 1$. d) Lastly, all 'Bridges' should be available in closed-form because it often implies fast computation.

4 General overview

Table 1: Overview of new bases.

Distribution	$g(x)$	$g^{-1}(x)$	$\phi(x)$	new $\phi(y)$	Standard space	New space
Exponential	log	exp	x	$(y, \exp(y))$	$\mathbb{R}_+$	$\mathbb{R}$
Gamma	log	exp	$(\log(x), x)$	$(\log(y), y^2)$	$\mathbb{R}_+$	$\mathbb{R}$
Gamma	sqrt	sqr	$(\log(x), x)$	$(y, \exp(y))$	$\mathbb{R}_+$	$\mathbb{R}$
Inv. Gamma	log	exp	$(\log(x), x)$	$(\log(y), y^2)$	$\mathbb{R}_+$	$\mathbb{R}$
Inv. Gamma	sqrt	sqr	$(\log(x), x)$	$(y, \exp(y))$	$\mathbb{R}_+$	$\mathbb{R}$
Chi2	log	exp	$(\log(x), x)$	$(\log(y), y^2)$	$\mathbb{R}_+$	$\mathbb{R}$
Chi2	sqrt	sqr	$(\log(x), x)$	$(y, \exp(y))$	$\mathbb{R}_+$	$\mathbb{R}$
Beta	logit	logistic	$(\log(x), \log(1-x))$	$(\log(\sigma(y))), (1 - \log(\sigma(y)))$	$\mathbb{P}$	$\mathbb{R}$
Dirichlet	-	softmax	$(\log(x_i))$	$\log(\pi_i(y))$	$\mathbb{P}^d$	$\mathbb{R}^d$
Wishart	logm	expm	$(\mathrm{logm}(X), X)$	$(Y, \mathrm{expm}(Y))$	$\mathbb{R}^{d\times d}_{++}$	$\mathbb{R}^{d\times d}$
Wishart	sqrtm	sqrm	$(\mathrm{logm}(X), X)$	$(\mathrm{logm}(Y), Y^2)$	$\mathbb{R}^{d\times d}_{++}$	$\mathbb{R}^{d\times d}$
Inv. Wishart	logm	expm	$(\mathrm{logm}(X), X)$	$(Y, \mathrm{expm}(Y))$	$\mathbb{R}^{d\times d}_{++}$	$\mathbb{R}^{d\times d}$
Inv. Wishart	sqrtm	sqrm	$(\mathrm{logm}(X), X)$	$(\mathrm{logm}(Y), Y^2)$	$\mathbb{R}^{d\times d}_{++}$	$\mathbb{R}^{d\times d}$

NOTE:MOST DISTRIBUTIONS ACTUALLY HAVE TWO VALID TRANSFORMS. ONE FOR X AND ONE FOR X^2 AS SUFFICIENT STATISTICS

TODO: full table of Bridges:

Distances:

Table 2: Transformations/Bridges. * means imperfect approximation

Distribution	Basis	$\theta \to \mathcal{N}$	$\mathcal{N} \to \theta$
Exponential	log	$\mu = \log(\frac{1}{\lambda})$ $\sigma^2 = 1$	$\lambda = \frac{1}{\exp(x)}$
Gamma	log	$\mu = \log\left(\frac{\alpha}{\lambda}\right)$ $\sigma^2 = \frac{1}{\alpha}$	$\alpha = \frac{1}{\sigma^2}$ $\lambda = \frac{1}{\exp(\mu)\sigma^2}$
Gamma	sqrt	$\mu = \sqrt{\frac{\alpha - 0.5}{\lambda}}$ $\sigma^2 = \frac{1}{4\lambda}$	$\alpha = \frac{\mu^2}{4\sigma^2} - 0.5$ $\lambda = \frac{4}{\sigma^2}$
Inv. Gamma	log	$\mu = \log\left(\frac{\lambda}{\alpha}\right)$ $\sigma^2 = \frac{1}{\alpha}$	$\alpha = \frac{1}{\sigma^2}$ $\lambda = \frac{\exp(\mu)}{\sigma^2}$
Inv. Gamma	sqrt	$\mu =$ $\sigma^2 =$	$\alpha =$ $\lambda =$
Chi-squared	log	$\mu =$ $\sigma^2 =$	$\alpha =$ $\lambda =$
Chi-squared	sqrt	$\mu =$ $\sigma^2 =$	$\alpha =$ $\lambda =$
Beta	logit	$\mu = \log(\frac{\alpha}{\beta})$ $\sigma^2 = \frac{\alpha+\beta}{\alpha\beta}$	$\alpha = \frac{\exp(\mu)+1}{\sigma^2}$ $\beta = \frac{\exp(-\mu)+1}{\sigma^2}$
Dirichlet	softmax^{-1}	$\mu_k = \log \alpha_k - \frac{1}{K} \sum_{l=1}^{K} \log \alpha_l$ $\Sigma_{k\ell} = \delta_{k\ell} \frac{1}{\alpha_k} - \frac{1}{K} \left[\frac{1}{\alpha_k} + \frac{1}{\alpha_\ell} - \frac{1}{K} \sum_{u=1}^{K} \frac{1}{\alpha_u} \right]$	$\alpha_k = \frac{1}{\Sigma_{kk}} \left(1 - \frac{2}{K} + \frac{e^{\mu_k}}{K^2} \sum_{l=1}^{K} e^{-\mu_l} \right)$
Wishart	logm	$\mu =$ $\Sigma =$	$V =$
Wishart*	sqrtm	$\mu =$ $\Sigma =$	$V =$
Inv. Wishart	logm	$\mu =$ $\Sigma =$	$V =$
Inv. Wishart*	sqrtm	$\mu =$ $\Sigma =$	$V =$

5 Exponential Distribution

5.1 Standard Exponential Distribution

The PDF of the exponential distribution is

$$p(x|\lambda) = \lambda \exp(-\lambda x) \tag{7}$$

which can be written as

$$p(x|\lambda) = \exp\left[-\lambda x + \log \lambda\right] \tag{8}$$

with $h(x) = 1$, $\phi(x) = x$, $w = -\lambda$ and $Z(\lambda) = -\log \lambda$

5.1.1 Laplace Approximation of the Exponential Distribution

$$\begin{aligned} \text{log-pdf: } & (\log \lambda - \lambda x) \\ \text{1st derivative: } & -\lambda \\ \text{2nd derivative: } & 0 \end{aligned}$$

The Laplace Approximation is not defined since the second derivative is not positive.

Table 3: Distances. MMD and KL div are normed to 1 in the standard basis for better comparison.

Distribution	Basis	MMD $\downarrow$	KL divergence $\downarrow$
Exponential	standard	∞	∞
Exponential	log	1	0.157
Exponential	sqrt	0.054	1
Gamma	standard	1	1
Gamma	log	0.196	0.147
Gamma	sqrt	0.031	0.406
Inv. Gamma	standard	1	1
Inv. Gamma	log	-	-
Inv. Gamma	sqrt	-	-
Chi2	standard	1	1
Chi2	log	-	-
Chi2	sqrt	-	-
Beta	standard	1	1
Beta	logit	-	-
Dirichlet	standard	1	1
Dirichlet	inverse softmax	-	-
Wishart	standard	1	1
Wishart	logm	-	-
Wishart	sqrtm	-	-
Inv. Wishart	standard	1	1
Inv. Wishart	logm	-	-
Inv. Wishart	sqrtm	-	-

5.2 Log-transformed Exponential Distribution

We choose $X = \log(Y)$ and therefore $g(x) = \log(x)$, and $x(y) = g^{-1}(y) = \exp(y)$. Also, $\left|\frac{\partial x(y)}{\partial y}\right| = \exp(y)$. It follows that the new pdf is

$$\begin{align} \mathcal{E}_{Y\,\log}(y;\lambda) &= \lambda\exp(-\lambda x(y)) \cdot \exp(y) \tag{9} \\ &= \lambda\exp(-\lambda\exp(y) + y) \nonumber \\ &= \exp\left[-\lambda\exp(y) + y + \log\lambda\right] \tag{10} \end{align}$$

with $h(y) = 1$, $\phi(x) = (y, \exp(y))$, $w = (1, -\lambda)$ and $Z(\lambda) = \log\lambda$

5.2.1 Laplace Approximation of the log-transformed Exponential Distribution

$$\begin{aligned} \text{log-pdf:}\ & -\lambda\exp(y) + y + \log\lambda \\ \text{1st derivative:}\ & \lambda - \exp(y) + 1 \\ \text{mode:}\ & y = \log(1/\lambda) \\ \text{2nd derivative:}\ & -\lambda\exp(y) \\ \text{insert mode:}\ & -\lambda\exp(1/\lambda) = -1 \\ \text{invert \& times -1:}\ & \sigma^2 = 1 \end{aligned}$$

Therefore the Laplace approximation in the transformed basis is given by $\mathcal{N}(y, \log(1/\lambda), 1)$.

5.3 The Bridge for the log-transformed Exponential Distribution

We have already found μ and σ. The inverse transformation is easily found through the mode $\mu = \log(1/\lambda) \Leftrightarrow \lambda = 1/\exp(\mu)$. In summary:

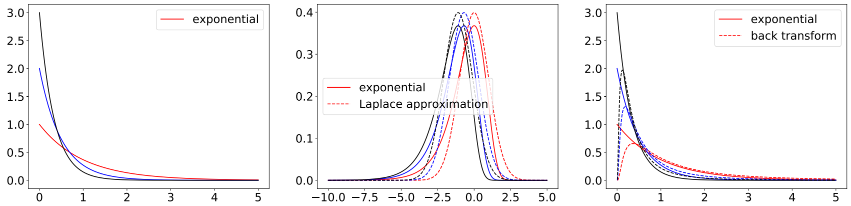

Figure 1: exponential comparison

$$\mu = \log(1/\lambda) \tag{11}$$
$$\sigma = 1 \tag{12}$$
$$\lambda = 1/\exp(\mu) \tag{13}$$

5.4 Sqrt-transformed Exponential Distribution

We choose $X = \sqrt{Y}$ and therefore $g(x) = \sqrt{x}$, and $x(y) = g^{-1}(y) = y^2$. Also, $\left|\frac{\partial x(y)}{\partial y}\right| = 2y$. It follows that the new pdf is

$$\begin{aligned} \mathcal{E}_{Y\,sqrt}(y;\lambda) &= \lambda \exp(-\lambda y^2) \cdot 2y \qquad (14)\\ &= 2 \cdot \exp\left[\log(y) - \lambda y^2 + \log\lambda\right] \\ & \qquad (15) \end{aligned}$$

with $h(y) = 2$, $\phi(y) = (\log(y), y^2))$, $w = (1, -\lambda)$ and $Z(\lambda) = \log\lambda$

5.4.1 Laplace Approximation of the sqrt-transformed Exponential Distribution

$$\begin{aligned} \text{log-pdf:}\ & \log(2y) - \lambda y^2 + \log\lambda \\ \text{1st derivative:}\ & \frac{1}{y} - 2\lambda y \\ \text{mode:}\ & y = \sqrt{\frac{1}{2\lambda}} \\ \text{2nd derivative:}\ & -\frac{1}{y^2} - 2\lambda \\ \text{insert mode:}\ & -\frac{1}{\frac{1}{2\lambda}} - 2\lambda = -4\lambda \\ \text{invert \& times -1:}\ & \sigma^2 = \frac{1}{4\lambda} \end{aligned}$$

5.5 The Bridge for the sqrt-transformed Exponential Distribution

$$\mu = \sqrt{\frac{1}{2\lambda}} \tag{16}$$
$$\sigma^2 = \frac{1}{4\lambda} \tag{17}$$
$$\lambda = \frac{1}{2\mu^2} \tag{18}$$

6 Gamma Distribution

6.1 Standard Gamma Distribution

$$\mathcal{G}_X(x, \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot x^{(\alpha-1)} \cdot e^{(-\lambda x)} \tag{19}$$

where $\Gamma(\alpha)$ is the Gamma function. This can be written as

$$\begin{aligned}\mathcal{G}_X(x, \alpha, \lambda) &= \exp\left[(\alpha-1)\log(x) - \lambda x + \alpha \log(\lambda) - \log(\Gamma(\alpha))\right] & (20)\\ &= \frac{1}{x}\exp\left[\alpha\log(x) - \lambda x + \alpha\log(\lambda) - \log(\Gamma(\alpha))\right] & (21)\end{aligned}$$

with $h(x) = \frac{1}{x}$, $\phi(x) = (\log x, x)$, $w = (\alpha, -\lambda)$ and $Z(\alpha, \lambda) = \log(\Gamma(\alpha)) - \alpha\log(\lambda)$.

6.1.1 Laplace Approximation of the Gamma Distribution

To get the LPA of the Gamma function in the standard basis we need its mode and the second derivative of the log-pdf. The mode is already known to be $\hat{\theta} = \frac{\alpha-1}{\lambda}$. For the second derivative of the log-pdf we take the log-pdf and derive it twice and insert the mode for x:

$$\begin{aligned}\text{log-pdf: } &\log\left(\frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot x^{(\alpha-1)} \cdot e^{(-\lambda x)}\right)\\ &= \alpha \cdot \log(\lambda) - \log(\Gamma(\alpha)) + (\alpha-1)\log(x) - \lambda x\\ \text{1st derivative: } &\frac{(\alpha-1)}{x} - \lambda\\ \text{mode: } &\frac{(\alpha-1)}{x} - \lambda = 0 \Leftrightarrow x = \frac{\alpha-1}{\lambda}\\ \text{2nd derivative: } &-\frac{(\alpha-1)}{x^2}\\ \text{insert mode: } &-\frac{(\alpha-1)}{(\frac{\alpha-1}{\lambda})^2} = -\frac{\lambda^2}{\alpha-1}\\ \text{invert and times -1: } &\sigma^2 = \frac{\alpha-1}{\lambda^2}\end{aligned}$$

The LPA of the Gamma distribution is therefore approximately distributed according to the pdf of $\mathcal{N}(\frac{\alpha-1}{\lambda}, \frac{\alpha-1}{\lambda^2})$.

6.2 Log-Transform of the Gamma Distribution

6.2.1 Log-Transformation

We transform the Gamma Distribution with the Log-Transformation, i.e. $Y = \log(X), g(x) = \log(x), x(y) = g^{-1}(x) = \exp(x)$. Also, $\left|\frac{\partial x(y)}{\partial y}\right| = \exp(y)$. The transformed pdf is

$$\begin{aligned}\mathcal{G}_{Y\_\log}(y, \alpha, \lambda) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot x(y)^{(\alpha-1)} \cdot e^{(-\lambda x(y))} \cdot \exp(y) & (22)\\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \exp(y)^\alpha \cdot e^{(-\lambda\exp(y))} \qquad = \exp\left[\alpha y - \lambda\exp(y) - \Gamma(\alpha) + \alpha\log(\lambda)\right]\end{aligned}$$

with exponential family parameters $h(y) = 1$, $\phi(y) = (y, \exp(y)), \eta = (\alpha, -\lambda)$ and $Z(\alpha, \lambda) = \log(\Gamma(\alpha)) - \alpha\log(\lambda)$.

6.2.2 Laplace Approximation of the log-transformed Gamma Distribution

To get the LPA of the Gamma distribution in the transformed basis we need to calculate its mode and the second derivative of the log-pdf. To get the mode we take the first derivative and set it to zero.

$$\begin{aligned}
\text{log-pdf: } &= \alpha \log(\lambda) - \log(\Gamma(\alpha)) + \alpha y - \lambda \exp(y) \\
\text{1st derivative: } & \alpha - \lambda \exp(y) \\
\text{mode: } & \alpha - \lambda \exp(y) = 0 \Leftrightarrow y = \log\left(\frac{\alpha}{\lambda}\right) \\
\text{2nd derivative: } & - \lambda \exp(y) \\
\text{insert mode: } & - \lambda \exp(\log\left(\frac{\alpha}{\lambda}\right)) = -\alpha \\
\text{invert and times -1: } & \sigma^2 = \frac{1}{\alpha}
\end{aligned}$$

Therefore the LPA now is $N(\log\left(\frac{\alpha}{\lambda}\right), \frac{1}{\alpha})$.

6.2.3 The bridge for the log-transformation

We already know how to get μ and σ from λ and α. To invert we calculate $\mu = \log(\alpha/\lambda) \Leftrightarrow \lambda = \alpha/\exp(\mu)$ and insert $\alpha = \sigma^2$. In summary we have

$$\mu = \log\left(\frac{\alpha}{\lambda}\right) \tag{23}$$

$$\sigma^2 = \frac{1}{\alpha} \tag{24}$$

$$\lambda = \frac{1}{\exp(\mu)\sigma^2} \tag{25}$$

$$\alpha = \frac{1}{\sigma^2} \tag{26}$$

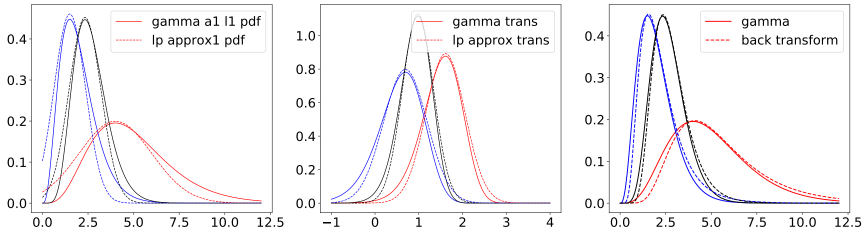

Figure 2: gamma comparison

6.3 Sqrt-Transform of the Gamma Distribution

6.3.1 Sqrt-Transformation

We transform the Gamma Distribution with the sqrt-transformation, i.e. $Y = \sqrt{X}, g(x) = \sqrt{x}, x_1(y) = g_1^{-1}(y) = -y^2, x_2(y) = g_2^{-1}(y) = y^2$ and $\left|\frac{\partial x_i(y)}{\partial y}\right| = \left|\frac{\partial g_i^{-1}(y)}{\partial y}\right| = |2y|$. We use the same 'trick' as in Subsection 2.4 to split up the transformation in two parts.

$$\begin{aligned}\mathcal{G}_Y(y) &= \frac{1}{2}\cdot\mathcal{G}_X(x_1(y))\left|\frac{\partial x_1(y)}{\partial y}\right|\mathbf{1}_{\wedge}(y) + \frac{1}{2}\cdot\mathcal{G}_X(x_2(y))\left|\frac{\partial x_2(y)}{\partial y}\right|\mathbf{1}_{\wedge}(y)\\ &= \frac{1}{2y^2}\exp[\alpha\log(y^2) - \lambda y^2 - A(\alpha,\lambda)]|2y|\mathbf{1}_{(-\infty,0)}(y) + \frac{1}{2y^2}\exp[\alpha\log(y^2) - \lambda y^2 - A(\alpha,\lambda)]|2y|\mathbf{1}_{[0,\infty)}(y)\\ &= \frac{1}{\sqrt{y}}\exp[2\alpha\log(y) - \lambda y^2 - A(\alpha,\lambda)]\mathbf{1}_{(-\infty,+\infty)}(y) \qquad (27)\\ &= \frac{1}{\sqrt{y}}\exp[2\alpha\log(y) - \lambda y^2 - A(\alpha,\lambda)]\end{aligned}$$

which is defined on the entirety of $\mathbb{R}$ and is an exponential family with $h(y) = \frac{1}{\sqrt{y}}$, $\phi(y) = (\log(y), y^2)$, $w = (2\alpha, -\lambda)$ and $Z(\alpha,\lambda) = \log(\Gamma(\alpha)) - \alpha\log(\lambda)$.

6.3.2 Laplace Approximation of the sqrt-transformed Gamma Distribution

To get the LPA of the Gamma distribution in the transformed basis we need to calculate its mode and the second derivative of the log-pdf. To get the mode we take the first derivative and set it to zero.

$$\begin{aligned}\text{log-pdf: } & (2\alpha - 1)\log(y) - \lambda y^2 + \alpha\log(\lambda) - \log(\Gamma(\alpha))\\ \text{1st derivative: } & \frac{2\alpha - 1}{y} - 2\lambda y\\ \text{mode: } & \frac{2\alpha - 1}{y} - 2\lambda y = 0 \Leftrightarrow y = \sqrt{\frac{\alpha - 0.5}{\lambda}}\\ \text{2nd derivative: } & -\frac{2\alpha - 1}{x^2} - 2\lambda\\ \text{insert mode: } & -\frac{2\alpha - 1}{\frac{\alpha - 0.5}{\lambda}} - 2\lambda = -4\lambda\\ \text{invert and times -1: } & \sigma^2 = \frac{1}{4\lambda}\end{aligned}$$

Therefore the LPA now is $\mathcal{N}\left(\sqrt{\frac{\alpha-0.5}{\lambda}}, \frac{1}{4\lambda}\right)$.

6.3.3 The bridge for the sqrt-transformation

We already know how to get μ and σ from λ and α. To invert we calculate $\mu = \sqrt{\frac{\alpha-0.5}{\lambda}} \Leftrightarrow \alpha = \frac{\mu^2}{\lambda} - 0.5$ and insert $\lambda = \frac{4}{\sigma^2}$. In summary we have

$$\mu = \sqrt{\frac{\alpha - 0.5}{\lambda}} \qquad (28)$$

$$\sigma^2 = \frac{1}{4\lambda} \qquad (29)$$

$$\lambda = \frac{4}{\sigma^2} \qquad (30)$$

$$\alpha = \frac{4\mu^2}{\sigma^2} + 0.5 \qquad (31)$$

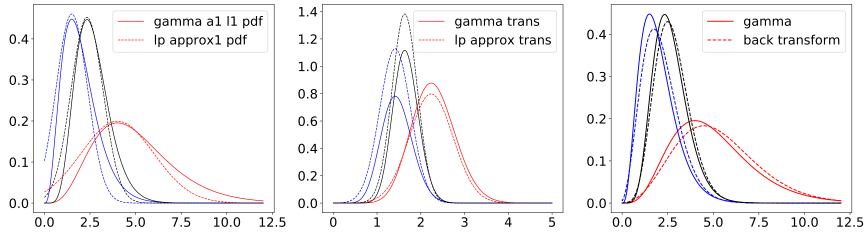

Figure 3: gamma comparison square

7 Inverse Gamma Distribution

7.1 Standard Inverse Gamma Distribution

The pdf of the inverse gamma is

$$f(x, \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp(-\frac{\lambda}{x}) \tag{32}$$

where Γ is the Gamma function. This can be rewritten as

$$f(x, \alpha, \lambda) = \exp\left[(-\alpha - 1)\log(x) - \lambda/x + \alpha \log(\lambda) - \log\Gamma(\alpha)\right] \tag{33}$$

where $T = (\log(x), x), \eta = (-\alpha - 1, -\lambda)$ and $A(\alpha, \lambda) = \log\Gamma(\alpha) - \alpha \log \lambda$.

7.1.1 Laplace Approximation of the standard inverse gamma distribution

$$\begin{aligned}
\text{log-pdf: } & (-\alpha - 1)\log(x) - \lambda/x + \alpha \log(\lambda) - \log\Gamma(\alpha) \\
\text{1st derivative: } & \frac{-\alpha - 1}{x} + \frac{\lambda}{x^2} \\
\text{mode: } & \frac{-\alpha - 1}{x} + \frac{\lambda}{x^2} = 0 \Leftrightarrow x = \frac{\lambda}{a+1} \\
\text{2nd derivative: } & \frac{\alpha + 1}{x^2} - 2\frac{\lambda}{x^3} \\
\text{insert mode: } & \frac{\alpha + 1}{{\frac{\lambda}{a+1}}^2} - 2\frac{\lambda}{{\frac{\lambda}{a+1}}^3} = -\frac{(\alpha+1)^3}{\lambda^2} \\
\text{invert and times -1: } & \sigma^2 = \frac{\lambda^2}{(\alpha+1)^3}
\end{aligned}$$

7.2 Sqrt-Transform of the inverse Gamma distribution

TODO: Double-Check this whole sqrt business.

7.2.1 Laplace Approximation of the sqrt-transformed Inverse Gamma Distribution

$$\begin{aligned}
\text{log-pdf: } & -2\alpha\log(x) - \frac{\lambda}{x^2} + \alpha\log\lambda - \log\Lambda(\alpha) \\
\text{1st derivative: } & -\frac{2\alpha}{x^2} + 2\frac{\lambda}{x^3} \\
\text{mode: } & x = \sqrt{\frac{\lambda}{\alpha}} \\
\text{2nd derivative: } & \frac{2\alpha}{x^2} - 6\frac{\lambda}{x^4} \\
\text{insert mode: } & -4\frac{\alpha^2}{\lambda} \\
\text{invert and times -1: } & \sigma^2 = \frac{\lambda}{4\alpha^2}
\end{aligned}$$

7.2.2 The Bridge for the sqrt-transformed Inverse Gamma Distribution

$$\mu = \sqrt{\frac{\lambda}{\alpha}} \tag{34}$$

$$\sigma^2 = \frac{\lambda}{4\alpha^2} \tag{35}$$

$$\alpha = \frac{\mu^2}{4\sigma^2} \tag{36}$$

$$\lambda = \frac{\mu^4}{4\sigma^2} \tag{37}$$

$$\tag{38}$$

7.3 Log-Transform of the inverse Gamma distribution

We choose $g(x) = \log(x)$, and thereby $g^{-1}(x) = \exp(x)$. It follows that the new pdf is

$$f_t(x, \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \exp(x)^{-\alpha} \exp(-\lambda/\exp(x)) \tag{39}$$

which can be written as

$$f_t(x, \alpha, \lambda) = \exp\left[-\alpha x - \frac{\lambda}{\exp(x)} + \alpha\log\lambda - \log\Lambda(\alpha)\right] \tag{40}$$

with $T = (x, 1/\exp(x)), \eta(-\alpha, \lambda)$ and $A(\alpha, \lambda) = \log\Gamma(\alpha) - \alpha\log\lambda$.

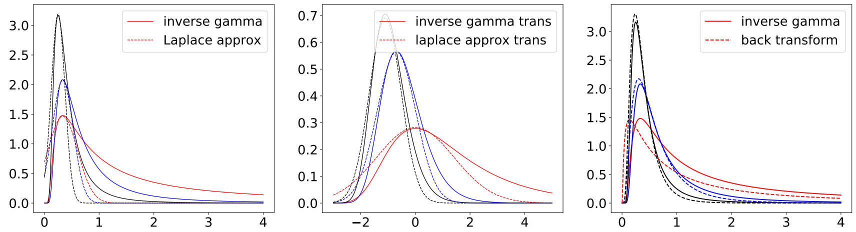

Figure 4: inverse gamma comparison

7.3.1 Laplace Approximation of the log-transformed Inverse Gamma Distribution

$$\begin{aligned}
\text{log-pdf:} \; & -\alpha x - \frac{\lambda}{\exp(x)} + \alpha \log \lambda - \log \Lambda(\alpha) \\
\text{1st derivative:} \; & -\alpha + \frac{\lambda}{\exp(x)} \\
\text{mode:} \; & -\alpha + \frac{\lambda}{\exp(x)} = 0 \Leftrightarrow x = \log(\lambda/\alpha) \\
\text{2nd derivative:} \; & -\frac{\lambda}{\exp(x)} \\
\text{insert mode:} \; & -\frac{\lambda}{\exp(\log(\lambda/\alpha))} = -\alpha \\
\text{invert and times -1:} \; & \sigma^2 = \frac{1}{\alpha}
\end{aligned}$$

7.3.2 The Bridge for the log-transformed Inverse Gamma Distribution

$$\mu = \log\left(\frac{\lambda}{\alpha}\right) \tag{41}$$

$$\sigma^2 = \frac{1}{\alpha} \tag{42}$$

$$\alpha = \frac{1}{\sigma^2} \tag{43}$$

$$\lambda = \frac{\exp(\mu)}{\sigma^2} \tag{44}$$

$$\tag{45}$$

8 Chi-squared Distribution

8.1 Standard Chi-squared distribution

The pdf is

$$f(x,k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp(-x/2) \tag{46}$$

which can be written as

$$f(x,k) = \exp\left[(k/2-1)\log(x) - x/2 - \log(2^{k/2}\Gamma(k/2))\right] \tag{47}$$

with $T = (\log(x), x), \eta = (k/2-1)$ and $A(k) = \log(2^{k/2}\Gamma(k/2))$.

8.1.1 Laplace approximation of the standard Chi-squared distribution

$$\begin{aligned}
\text{log-pdf: } & (k/2-1)\log(x) - x/2 - \log(2^{k/2}\Gamma(k/2)) \\
\text{1st derivative: } & \frac{k/2-1}{x} - \frac{1}{2} \\
\text{mode: } & \frac{k/2-1}{x} - \frac{1}{2} = 0 \Leftrightarrow x = k-2 \\
\text{2nd derivative: } & -\frac{k/2-1}{x^2} \\
\text{insert mode: } & -\frac{k/2-1}{(k-2)^2} \\
\text{invert and times -1: } & \sigma^2 = \frac{(k-2)^2}{k/2-1}
\end{aligned}$$

8.2 Log-Transformed Chi-squared distribution

we transform the distribution with $g(x) = \log(x)$, i.e. $g^{-1}(x) = \exp(x)$. The new pdf becomes

$$f_t(x,k) = \frac{1}{2^{k/2}\Gamma(k/2)} \exp(x)^{k/2-1} \exp(-\exp(x)/2) \tag{48}$$

which can be written as

$$f_t(x,k) = \exp\left[(k/2)x - \frac{\exp(x)}{2} - \log(2^{k/2}\Gamma(k/2))\right] \tag{49}$$

meaning $T = (x, \exp(x)), \eta = (k/2)$ and $A(k) = \log(2^{k/2}\Gamma(k/2))$.

8.2.1 Laplace approximation of the log-transformed Chi-squared distribution

$$\begin{aligned}
\text{log-pdf: } & (k/2)x - \frac{\exp(x)}{2} - \log(2^{k/2}\Gamma(k/2)) \\
\text{1st derivative: } & k/2 - \frac{\exp(x)}{2} \\
\text{mode: } & k/2 - \frac{\exp(x)}{2} = 0 \Leftrightarrow x = \log(k) \\
\text{2nd derivative: } & -\frac{\exp(x)}{2} \\
\text{insert mode: } & -k/2 \\
\text{invert and times -1: } & \sigma^2 = 2/k
\end{aligned}$$

8.2.2 The Bridge for log-transform

$$\mu = \log(k) \tag{50}$$

$$\sigma^2 = 2/k \tag{51}$$

$$k = \exp(\mu) \tag{52}$$

8.3 Sqrt-Transformed Chi-squared distribution

we transform the distribution with $g(x) = \sqrt{x}$, i.e. $g^{-1}(x) = x^2$. The new pdf becomes

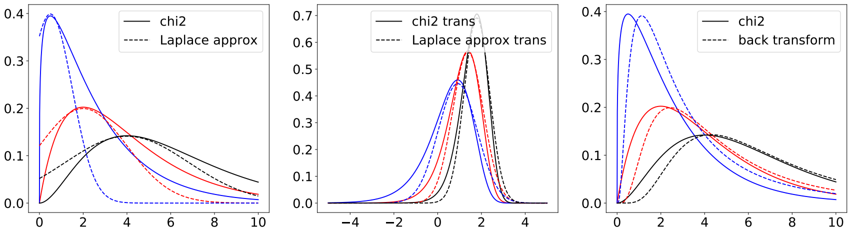

Figure 5: chi2 comparison log transform

$$f_t(x,k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{2(k/2-1)} \exp(-x^2/2)2x \tag{53}$$
$$= \frac{1}{2^{k/2}\Gamma(k/2)} x^{k} \exp(-x^2/2)$$

which can be written as

$$f_t(x,k) = \exp\left[(k\log(x) - \frac{x^2}{2} - \log(2^{k/2}\Gamma(k/2))\right] \tag{54}$$

meaning $T = (\log(x), x^2), \eta = (k, 1/2)$ and $A(k) = \log(2^{k/2}\Gamma(k/2))$.

8.3.1 Laplace approximation of the sqrt-transformed Chi-squared distribution

$$\begin{aligned}
\text{log-pdf: } & (k\log(x) - \frac{x^2}{2} - \log(2^{k/2}\Gamma(k/2)) \\
\text{1st derivative: } & \frac{k}{x} - x \\
\text{mode: } & \frac{k}{x} - x = 0 \Leftrightarrow x = \sqrt{k} \\
\text{2nd derivative: } & -\frac{k}{x^2} - 1 \\
\text{insert mode: } & -\frac{k}{k} - 1 \\
\text{invert and times -1: } & \sigma^2 = 1/2
\end{aligned}$$

8.3.2 The Bridge for sqrt-transform

$$\mu = \sqrt{k} \tag{55}$$
$$\sigma^2 = 1/2 \tag{56}$$
$$k = \mu^2 \tag{57}$$

TODO:THE BRIDGE BACK LOOKS A BIT WEIRD

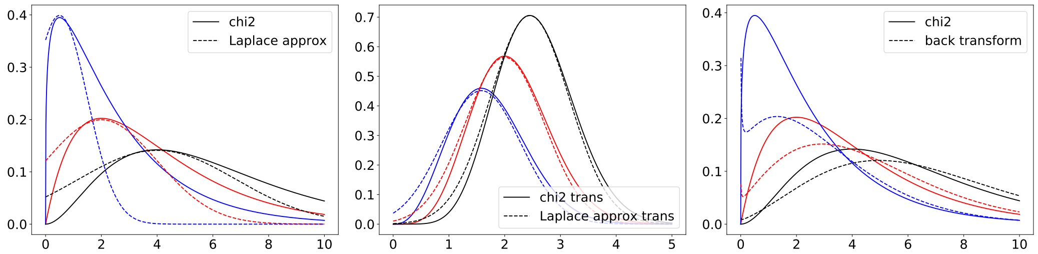

Figure 6: chi2 sqrt comparison

9 Beta Distribution

9.1 Standard Beta Distribution

The pdf of the Beta distribution in the standard basis is

$$f(x, \alpha, \beta) = \frac{x^{(\alpha-1)} \cdot (1-x)^{(\beta-1)}}{B(\alpha, \beta)} \tag{58}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(x)$ is the Gamma function. This can be written as

$$\begin{aligned} f(x, \alpha, \beta) &= \exp\left[(\alpha-1)\log(x) + (\beta-1)\log(1-x) - \log(B(\alpha, \beta))\right] && (59) \\ &= \frac{1}{x(1-x)} \exp\left[\alpha \log(x) + \beta \log(1-x) - \log(B(\alpha, \beta))\right] && (60) \end{aligned}$$

With $h(x) = \frac{1}{x(1-x)}, T = (\log(x), \log(1-x), \eta = (\alpha, \beta)$ and $A(\alpha, \beta) = \log(B(\alpha, \beta)))$.

9.1.1 Laplace approximation of the standard Beta distribution

To get the Laplace approximation we need the mode and Hessian. To get the mode we use the first derivative of the log-pdf and set it to zero. To get the Covariance we use the Hessian at the mode, multiply it with -1 and invert it.

$$\begin{aligned} \text{log-pdf: } & \log\left(\frac{x^{(\alpha-1)} \cdot (1-x)^{(\beta-1)}}{B(\alpha, \beta)}\right) \\ & = (\alpha-1)\log(x) + (\beta-1)\log(1-x) - \log(B(\alpha, \beta))) \\ \text{1st derivative: } & \frac{(\alpha-1)}{x} - \frac{(\beta-1)}{1-x} \\ \text{mode: } & \frac{(\alpha-1)}{x} - \frac{(\beta-1)}{1-x} = 0 \Leftrightarrow x = \frac{\alpha-1}{\alpha+\beta-2} \\ \text{2nd derivative: } & \frac{\alpha-1}{x^2} + \frac{\beta-1}{(1-x)^2} \\ \text{insert mode: } & \frac{\alpha-1}{\frac{\alpha-1}{\alpha+\beta-2}^2} + \frac{\beta-1}{(1-\frac{\alpha-1}{\alpha+\beta-2})^2} = \frac{(\alpha+\beta-2)^3}{(\alpha-1)(\beta-1)} \\ \text{invert: } & \frac{(\alpha-1)(\beta-1)}{(\alpha+\beta-2)^3} \end{aligned}$$

The Beta distribution in standard basis is therefore approximated by $N(\mu = \frac{\alpha-1}{\alpha+\beta-2}, \sigma^2 = \frac{(\alpha-1)(\beta-1)}{(\alpha+\beta-2)^3})$.

9.2 Logit-Transform of the Beta distribution

We transform the Beta distribution using $g(x) = \log(\frac{x}{1-x})$. Therefore $g^{-1}(x) = \sigma(x) = \frac{1}{1+\exp(-x)}$. This yields the following pdf

$$\begin{aligned} f_t(x, \alpha, \beta) &= \frac{1}{\sigma(x)(1-\sigma(x))} \exp\left[\alpha \log(\sigma(x)) + \beta \log(1-\sigma(x)) - \log(B(\alpha, \beta))\right] \cdot (\sigma(x)(1-\sigma(x)) \tag{61} \\ &= \exp\left[\alpha \log(x) + \beta \log(1-x) - \log(B(\alpha, \beta))\right] \end{aligned}$$

Which has $h(x) = 1, T = (\log(\sigma(x)), \log(1-\sigma(x)), \eta = (\alpha, \beta)$ and $A(\alpha, \beta) = \log(B(\alpha, \beta))$.

9.2.1 Laplace approximation of the logit transformed Beta distribution

mode and variance of log pdf blablabla

$$\begin{aligned} \text{log-pdf: } & \log\left(\frac{\sigma(x)^\alpha \cdot (1-\sigma(x)^\beta)}{B(\alpha, \beta)}\right) \\ & = \alpha \log(\sigma(x)) + \beta \log(1-\sigma(x)) - \log(B(\alpha, \beta)) \\ \text{1st derivative: } & \alpha(1-\sigma(x)) - \beta\sigma(x) \\ \text{mode: } & \alpha(1-\sigma(x)) - \beta\sigma(x) = 0 \Leftrightarrow x = -\log(\frac{\beta}{\alpha}) \\ \text{2nd derivative: } & (\alpha + \beta)\sigma(x)(1-\sigma(x)) \\ \text{insert mode: } & (\alpha + \beta)\sigma(-\log(\frac{\beta}{\alpha}))(1-\sigma(-\log(\frac{\beta}{\alpha}))) = \frac{\alpha\beta}{\alpha+\beta} \\ \text{invert: } & \frac{\alpha+\beta}{\alpha\beta} \end{aligned}$$

The LPA is therefore $\mathcal{N}(\mu = -\log(\frac{\beta}{\alpha}), \sigma^2 = \frac{\alpha+\beta}{\alpha\beta})$.

9.2.2 The Bridge for the logit transformation

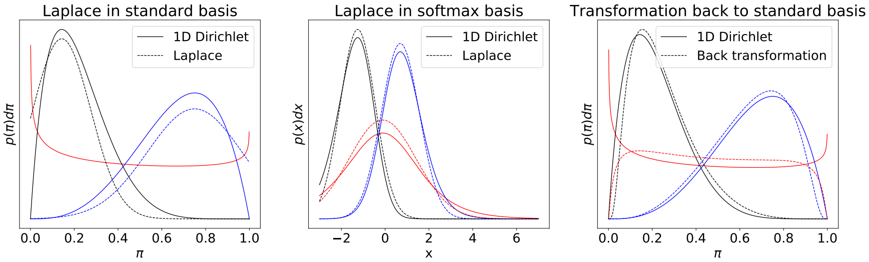

Figure 7: beta comparison

10 Dirichlet Distribution

10.1 Standard Dirichlet distribution

The pdf for the Dirichlet distribution in the standard basis (i.e. probability space) is

$$\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{62}$$

$$= \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{63}$$

$$= \exp\left[\sum_k (\alpha_k - 1)\log(\pi_k) - \log(B(\alpha))\right] \tag{64}$$

$$= \frac{1}{\prod_k \pi_k} \exp\left[\sum_k \alpha_k \log(\pi_k) - \log(B(\alpha))\right] \tag{65}$$

$$\tag{66}$$

with sufficient statistics $\phi(x_i) = \log(x_i)$, natural parameters $w_i = \alpha_i$, base measure $h(x) = \prod_k x_k$, and partition function $Z(w) = \log(B(\alpha))$.

10.1.1 Laplace approximation of the standard Dirichlet distribution

$$\text{log-pdf: } f = \sum_k \alpha_k \log(\pi_k) - \log(B(\alpha))$$

$$\text{1st derivative: } \frac{\partial f}{\partial x_i} = \frac{\alpha_i - 1}{x_i}$$

$$\text{mode: } x_i = \frac{(\alpha_i - 1)}{\sum_k \alpha_k}$$

$$\text{2nd derivative: } \frac{\partial^2 f}{\partial x_i \partial x_j} = -\delta_{ij}\frac{(\alpha_i - 1)}{x_i^2}$$

$$\text{insert mode: } -\delta_{ij}\frac{(\sum_k \alpha_k)^2}{(\alpha_i - 1)}$$

$$\text{invert and times -1: } \Sigma_{ij} = \delta_{ij}\frac{\alpha_i - 1}{(\sum_k \alpha_k)^2}$$

Which yields a diagonal Covariance matrix for the Laplace approximation.

10.2 Softmax-Transform of the Dirichlet distribution

We aim to transform the basis of this distribution from base $\mathbf{y}$ via the softmax transform to be in the new base π:

$$\pi_k(\mathbf{y}) := \frac{\exp(y_k)}{\sum_{l=1}^{K} \exp(y_l)}\,, \tag{67}$$

TODO:David J. MacKay has already done a transformation for the determinant but in a slightly different fashion.

The softmax transform has no analytic inverse $\pi_k^{-1}(y)$ but it is not necessary for our computation since we assume $\pi(y)$ to be the inverse transformation already (i.e. $g^{-1}(y)$). However, our transformation is from a variable in $\mathbb{R}^d$ (which has d degrees of freedom) to a variable that is in $\mathbb{P}^d$ (which has $d-1$ degrees of freedom). To account for the difference in size of the two spaces we create a helper variable for the transformation as described in the following.

We want to transform K variables y_i from $\mathbb{R}^d$ to $\tau_i = \exp(y_i)$. For τ_i to be equal to π_i we need to ensure that it sums to 1, $u = \sum_i \tau_i = 1$. With the helper-variable u our variable transform $g(\pi, u)$ becomes

$$p_{y,u}(\pi(y), u) = p_{\pi,u}(\pi(y), u)|\det g(\pi(y), u)| \tag{68}$$

with

$$\begin{aligned}
|\det g(\pi(y), u)| &= \left|\det \begin{pmatrix} \frac{\partial \tau_1}{\partial y_1} & \cdots & \frac{\partial \tau_k}{\partial y_1} & \frac{\partial u}{\partial y_1} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\partial \tau_1}{\partial y_k} & \cdots & \frac{\partial \tau_k}{\partial y_k} & \frac{\partial u}{\partial y_k} \\ \frac{\partial \tau_1}{\partial u} & \cdots & \frac{\partial \tau_k}{\partial u} & \frac{\partial u}{\partial u} \end{pmatrix}\right| && (69) \\
&= \left|\det \begin{pmatrix} \tau_1 = \exp(y_1) & \cdots & 0 & \tau_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \tau_k & \tau_k \\ 0 & \cdots & 0 & 1 \end{pmatrix}\right| && (70) \\
&= \prod_i^K \tau_i && (71)
\end{aligned}$$

To get $p_y(\pi(y))$ we have to integrate out u.

$$\begin{aligned}
p_y(\pi(y)) &= \int_{-\infty}^{\infty} p_{\pi,u}(\pi(y), u)|\det g(\pi, u)|du && (72) \\
&= \int_{-\infty}^{\infty} p_\pi(\pi(y)) \prod_i^K \tau_i du && (73) \\
&= p_\pi(\pi(y)) \int_0^{\infty} \prod_i^K \tau_i \delta(u-1)du && (74) \\
&= p_\pi(\pi(y)) \int_0^{\infty} \prod_i^K \tau_i \frac{u}{u} \delta(u-1)du && (75) \\
&= p_\pi(\pi(y)) \int_0^{\infty} \underbrace{\prod_i^K \pi_i u}_{f(u)} \, \delta(u-1)du && (76) \\
&= p_\pi(\pi(y)) \cdot \prod_i^K \pi_i(y) && (77) \\
&= \frac{1}{\prod_k \pi_k(y)} \exp\left[\sum_k \alpha_k \log(\pi_k(y)) - \log(B(\alpha))\right] \prod_k^K \pi_k(y) && (78) \\
&= \exp\left[\sum_k \alpha_k \log(\pi(y_k)) - \log(B(\alpha))\right] && (79)
\end{aligned}$$

where we used the fact that $u > 0$ since its a sum of exponentials, $\frac{\tau_i(y)}{u} = \pi_i(y)$, and multiplied with $\delta(u-1)$ since this transformation is only valid if $\sum_i \tau_i = u = 1$ because otherwise it is not a probability space. Additionally, we use

$$\int_{-\infty}^{\infty} f(x)\delta(x-t)dx = f(t) \tag{80}$$

which is known as the shifting property or sampling property of the Dirac delta function δ. Using all of the above we get the pdf of the Dirichlet distribution in the new basis $\mathbf{y}$:

$$\mathrm{Dir}_{\mathbf{y}}(\boldsymbol{\pi}(\mathbf{y})|\boldsymbol{\alpha}) := \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\pi_k(\mathbf{y})^{\alpha_k} \tag{81}$$

$$= \exp\left[\sum_k \alpha_k \log(\pi(y_k)) - \log(B(\alpha))\right] \tag{82}$$

with sufficient statistics $\phi(y_i) = \log(\pi_i(y))$, natural parameters $w_i = \alpha_i$, base measure $h(y) = 1$ and normalizing constant $Z = \log(B(\alpha))$.

10.2.1 Laplace approximation of the softmax-transformed Dirichlet distribution

TODO: mention that this has been done by Philipp in his PhD thesis.

Through the figures of the 1D Dirichlet approximation in the main paper we have already established that the mode of the Dirichlet lies at the mean of the Gaussian distribution and therefore $\boldsymbol{\pi}(\mathbf{y}) = \frac{\alpha}{\sum_i \alpha_i}$. Additionally, the elements of $\mathbf{y}$ must sum to zero. These two constraints combined yield only one possible solution for $\boldsymbol{\mu}$.

$$\mu_k = \log\alpha_k - \frac{1}{K}\sum_{l=1}^{K}\log\alpha_l \tag{83}$$

Calculating the covariance matrix $\boldsymbol{\Sigma}$ is more complicated but layed out in the following. The logarithm of the Dirichlet is, up to additive constants

$$\log p_y(y|\alpha) = \sum_k \alpha_k \pi_k \tag{84}$$

Using π_k as the softmax of $\mathbf{y}$ as shown in Equation 67 we can find the elements of the Hessian $\mathbf{L}$

$$L_{kl} = \hat{\alpha}(\delta_{kl}\hat{\pi}_k - \hat{\pi}_k\hat{\pi}_l) \tag{85}$$

where $\hat{\boldsymbol{\alpha}} := \sum_k \alpha_k$ and $\hat{\boldsymbol{\pi}} = \frac{\alpha_k}{\hat{\alpha}}$ for the value of $\boldsymbol{\pi}$ at the mode. Analytically inverting $\mathbf{L}$ is done via a lengthy derivation using the fact that we can write $\mathbf{L} = \mathbf{A} + \mathbf{X}\mathbf{B}\mathbf{X}^\top$ and inverting it with the Schur-complement. This process results in the inverse of the Hessian

$$L_{kl}^{-1} = \delta_{kl}\frac{1}{\alpha_k} - \frac{1}{K}\left[\frac{1}{\alpha_k} + \frac{1}{\alpha_l} - \frac{1}{K}\left(\sum_u^K \frac{1}{\alpha_u}\right)\right] \tag{86}$$

We are mostly interested in the diagonal elements, since we desire a sparse encoding for computational reasons and we otherwise needed to map a $K \times K$ covariance matrix to a $K \times 1$ Dirichlet parameter vector which would be a very overdetermined mapping. Note that K is a scalar not a matrix. The diagonal elements of $\boldsymbol{\Sigma} = \mathbf{L}^{-1}$ can be calculated as

$$\Sigma_{kk} = \frac{1}{\alpha_k}\left(1 - \frac{2}{K}\right) + \frac{1}{K^2}\sum_l^k \frac{1}{\alpha_l}. \tag{87}$$

To invert this mapping we transform Equation 83 to

$$\alpha_k = e^{\mu_k}\prod_l^K \alpha_l^{1/K} \tag{88}$$

by applying the logarithm and re-ordering some parts. Inserting this into Equation 87 and re-arranging yields

$$\prod_{l}^{K} \alpha_l^{1/K} = \frac{1}{\mathbf{\Sigma}_{kk}} \left[e^{-\mu} \left(1 - \frac{2}{K} \right) + \frac{1}{K^2} \sum_{u}^{K} e^{-\mu_u} \right] \tag{89}$$

which can be re-inserted into Equation 88 to give

$$\alpha_k = \frac{1}{\Sigma_k k} \left(1 - \frac{2}{K} + \frac{e^{-\mu_k}}{K^2} \sum_{l}^{K} e^{-\mu_k} \right) \tag{90}$$

which is the final mapping. With Equations 83 and 87 we are able to map from Dirichlet to Gaussian and with Equation 90 we are able to map the inverse direction.

10.2.2 The Bridge for the inverse-softmax transform

In summary we get the following forward and backward transformations between $\mathbf{y} \in \mathbb{R}^d$ and $\pi \in \mathbb{P}^d$.

$$\mu_k = \log \alpha_k - \frac{1}{K} \sum_{l=1}^{K} \log \alpha_l \,, \tag{91}$$

$$\Sigma_{k\ell} = \delta_{k\ell} \frac{1}{\alpha_k} - \frac{1}{K} \left[\frac{1}{\alpha_k} + \frac{1}{\alpha_\ell} - \frac{1}{K} \sum_{u=1}^{K} \frac{1}{\alpha_u} \right] . \tag{92}$$

The corresponding derivations require care because the Gaussian parameter space is evidently larger than that of the Dirichlet and not fully identified by the transformation. A pseudo-inverse of this map was provided by **?**. It maps the Gaussian parameters to those of the Dirichlet as

$$\alpha_k = \frac{1}{\Sigma_{kk}} \left(1 - \frac{2}{K} + \frac{e^{\mu_k}}{K^2} \sum_{l=1}^{K} e^{-\mu_l} \right) \tag{93}$$

11 Wishart Distribution

11.1 Interlude: Box-product and Kronecker-product

Kronecker-product: $A \otimes B \in \mathbb{R}^{(m_1 m_2) \times (n_1 n_2)}$ is defined by $(A \otimes B)_{(i-1)m_2 + j, (k-1)n_2 + l} = a_{il} b_{jk} = (A \otimes B)_{(ij)(kl)}$.

Box-product: $A \boxtimes B \in \mathbb{R}^{(m_1 m_2) \times (n_1 n_2)}$ is defined by $(A \boxtimes B)_{(i-1)m_2 + j, (k-1)n_1 + l} = a_{ik} b_{jl} = (A \boxtimes B)_{(ij)(kl)}$.

I found this box-product only in two sources, one of which is this: `https://researcher.watson.ibm.com/researcher/files/us-pederao/ADTalk.pdf` but it generally seems to be very helpful for matrix derivations with transposed matrices.

11.2 Standard Wishart distribution

the pdf of the Wishart is

$$f(X; n, p, V) = \frac{1}{2^{np/2} \left|\mathbf{V}\right|^{n/2} \Gamma_p\left(\frac{n}{2}\right)} |\mathbf{X}|^{(n-p-1)/2} e^{-(1/2) \operatorname{tr}(\mathbf{V}^{-1} \mathbf{X})} \tag{94}$$

which can be written as

$$f(X;n,p,V) = \exp\left[(n-p-1)/2\log(|X|) - (1/2)\operatorname{tr}(\mathbf{V}^{-1}\mathbf{X}) - \log\left(2^{np/2}\left|\mathbf{V}\right|^{n/2}\Gamma_p\left(\frac{n}{2}\right)\right)\right] \tag{95}$$

with $T = (\log(X), X), \eta = ((n-p-1)/2, V^{-1})$ and $A(n,p,V) = \log\left(2^{np/2}\left|\mathbf{V}\right|^{n/2}\Gamma_p\left(\frac{n}{2}\right)\right)$

11.2.1 Laplace Approximation of the standard Wishart distribution

Using $\frac{\partial \det(X)}{\partial X} = \det(X)(X^{-1})^\top$ and $\frac{\partial}{\partial X}Tr(AX^\top) = A$ we can calculate the mode by setting the first derivative of the log-pdf to zero

$$\begin{aligned}
\frac{\partial \log f(X;n,p,V)}{\partial X} &= \frac{(n-p-1)\det(X)(X^{-\top})}{2\det(X)} - \frac{V^{-1}}{2} \\
\Rightarrow 0 &= \frac{(n-p-1)X^{-1}}{2} - \frac{V^{-1}}{2} \\
\Leftrightarrow \frac{(n-p-1)X^{-1}}{2} &= \frac{V^{-1}}{2} \\
\Leftrightarrow X &= (n-p-1)V
\end{aligned}$$

Using the fact that $\frac{\partial X^{-T}}{\partial X} = X^{-T} \boxtimes X^{-1}$ where $\boxtimes$ is the Box-product we compute the second derivative as

$$\frac{\partial^2 \log f(X;n,p,V)}{\partial^2 X} = -\frac{(n-p-1)}{2}X^{-\top} \boxtimes X^{-1}$$

Using $(\alpha A)^{-1} = \alpha^{-1}A^{-1}$, the linearity of the Kronecker product to pull out scalars and $X^{-1} \boxtimes X^{-1} = (X \boxtimes X)^{-1}$ to insert the mode and invert we get:

$$\begin{aligned}
-\frac{(n-p-1)}{2}X^{-1} \boxtimes X^{-1} &= -\frac{(n-p-1)}{2}\frac{1}{(n-p-1)}V^{-1} \otimes \frac{1}{(n-p-1)}V^{-1} \\
&= -\frac{1}{2(n-p-1)}(V \boxtimes V)^{-1} \\
\Rightarrow \Sigma &= 2(n-p-1)(V \boxtimes V)
\end{aligned}$$

In summary, the Laplace approximation of a Wishart distribution in the standard basis is $\mathcal{N}\left(X;(n-p-1)V, 2(n-p-1)(V \boxtimes V)\right)$, where the representation of the symmetric positive definite matrices has been changed from $\mathbb{R}^{n\times n}$ to $\mathbb{R}^{n^2}$.

11.3 Logm-Transformed Wishart distribution

we transform the distribution with $g(X) = \operatorname{logm}(X)$, i.e. $g^{-1}(X) = \operatorname{expm}(X)$, where $\operatorname{expm}(X)$ is the matrix exponential and $\operatorname{logm}(X)$ is the matrix logarithm of X. The new pdf becomes

$$\begin{aligned}
f(X;n,p,V) &= \frac{1}{2^{np/2}\left|\mathbf{V}\right|^{n/2}\Gamma_p\left(\frac{n}{2}\right)}\left|\operatorname{expm}\mathbf{X}\right|^{(n-p-1)/2}e^{-(1/2)\operatorname{tr}(\mathbf{V}^{-1}\operatorname{expm}\mathbf{X})} \cdot \left|\operatorname{expm}X\right| \\
&= \frac{1}{2^{np/2}\left|\mathbf{V}\right|^{n/2}\Gamma_p\left(\frac{n}{2}\right)}\left|\operatorname{expm}\mathbf{X}\right|^{(n-p+1)/2}e^{-(1/2)\operatorname{tr}(\mathbf{V}^{-1}\operatorname{expm}\mathbf{X})} \\
&= \exp\left[C + \frac{(n-p+1)}{2}\log(\left|\operatorname{expm}\mathbf{X}\right|) - \frac{1}{2}\operatorname{tr}(\mathbf{V}^{-1}\operatorname{expm}\mathbf{X})\right]
\end{aligned}$$

with blablabla as expfam values.

11.3.1 Laplace Approximation of the logm-transformed Wishart distribution

To compute the first derivative we use the following

$$\frac{\partial \log(\det(\mathrm{expm}(X)))}{\partial X} = \frac{\partial \log(\det(\mathrm{expm}(X)))}{\partial \det(\mathrm{expm}(X))} \cdot \frac{\partial \det(\mathrm{expm}(X))}{\partial \,\mathrm{expm}(X)} \cdot \frac{\partial \,\mathrm{expm}(X)}{\partial X} \tag{96}$$

$$= \frac{1}{\det(\mathrm{expm}(X))} \cdot \det(\mathrm{expm}(X))\,\mathrm{expm}(X)^{-\top} \cdot \mathrm{expm}(X) \tag{97}$$

$$= I_p \tag{98}$$

where I_p is the identity matrix of size p and we use the fact that the matrix logaritm of a symmetric matrix is symmetric, implying $\mathrm{expm}(X)^{-\top} = \mathrm{expm}(X)^{-1}$. With this we get the

$$\frac{\partial \log W_{log}}{\partial X} = \frac{\partial}{\partial X}\left[C + \frac{(n-p+1)}{2}\log(|\mathrm{expm}\,\mathbf{X}|) - \frac{1}{2}\,\mathrm{tr}(\mathbf{V}^{-1}\mathrm{expm}\,\mathbf{X})\right] \tag{99}$$

$$= \frac{(n-p+1)}{2} I_p - \frac{1}{2} V^{-1}\mathrm{expm}\,\mathbf{X} \tag{100}$$

By setting this to zero we get a mode of

$$0 = \frac{(n-p+1)}{2} I_p - \frac{1}{2} V^{-1}\mathrm{expm}\,\mathbf{X} \tag{101}$$

$$\Leftrightarrow (n-p+1) I_p = V^{-1}\mathrm{expm}\,\mathbf{X} \tag{102}$$

$$\Leftrightarrow X = \mathrm{logm}((n-p+1)V) \tag{103}$$

For the second derivative we use the fact that

$$\frac{\partial (B\,\mathrm{expm}(X))_{kl}}{\partial X_{ij}} = ... \tag{104}$$

$$\Leftrightarrow \frac{\partial B\,\mathrm{expm}(X)}{\partial X} = (B\,\mathrm{expm}(X) \otimes I_p) \tag{105}$$

yielding

$$\frac{\partial^2 \log W_{log}}{\partial^2 X} = \frac{\partial \log W_{log}}{\partial X} \frac{(n-p+1)}{2} I_p - \frac{1}{2} V^{-1}\mathrm{expm}\,\mathbf{X} \tag{106}$$

$$= -\frac{1}{2}(V^{-1}\mathrm{expm}\,\mathbf{X} \otimes I_p) \tag{107}$$

$$\overset{\text{mode}}{\Rightarrow} -\frac{1}{2}((n-p+1)V^{-1}V \otimes I_p) \tag{108}$$

$$= -\frac{(n-p+1)}{2}(I_p \otimes I_p) \tag{109}$$

$$\Leftrightarrow \Sigma = \frac{2}{n-p+1} I_{p^2} \tag{110}$$

where I_{p^2} is an Identity matrix of size p^2.

11.3.2 The Bridge for logm-tranform

μ and Σ are already given by the Laplace approximation. Inverting the mode yields an estimate for V.

$$\mu = \mathrm{logm}((n-p+1)V) \Leftrightarrow \mathrm{expm}(\mu) = (n-p+1)V \Leftrightarrow V = \frac{\mathrm{expm}(\mu)}{(n-p+1)} \tag{111}$$

where μ and V are reshaped to a matrix of size $p \times p$. In summary this yields

$$\mu = \text{logm}((n-p+1)V) \tag{112}$$
$$\Sigma = \frac{2}{n-p+1} I_{p^2} \tag{113}$$
$$V = \frac{\text{expm}(\mu)}{(n-p+1)} \tag{114}$$

11.4 Sqrtm-Transformed Wishart distribution

we transform the distribution with $g(X) = \text{sqrtm}(X) = X^{\frac{1}{2}}$, i.e. $g^{-1}(X) = X^2$, where sqrtm(X) is the square root of the matrix. The new pdf becomes

$$f_t(X; n, p, V) = \frac{1}{2^{np/2} \left|\mathbf{V}\right|^{n/2} \Gamma_p\left(\frac{n}{2}\right)} \left|\mathbf{X^2}\right|^{(n-p-1)/2} e^{-(1/2)\,\text{tr}(\mathbf{V}^{-1}\mathbf{X^2})} \cdot |2X| \tag{115}$$
$$= \frac{1}{2^{np/2} \left|\mathbf{V}\right|^{n/2} \Gamma_p\left(\frac{n}{2}\right)} |\mathbf{X}|^{2(n-p-1)/2} e^{-(1/2)\,\text{tr}(\mathbf{V}^{-1}\mathbf{X^2})} \cdot 2^p |X| \tag{116}$$
$$= \frac{1}{2^{np/2} \left|\mathbf{V}\right|^{n/2} \Gamma_p\left(\frac{n}{2}\right)} |\mathbf{X}|^{(n-p)} e^{-(1/2)\,\text{tr}(\mathbf{V}^{-1}\mathbf{X^2})} \tag{117}$$

where we drop the 2^p in line (4) because there are 2^p matrices that are a root of X (I have explained this in more detailed in another version of the current draft). This can be rewritten as

$$f_t(X; n, p, V) = \exp\left[(n-p)\log(|X|) - (1/2)\,\text{tr}(\mathbf{V}^{-1}\mathbf{X^2}) - \log\left(2^{np/2}\left|\mathbf{V}\right|^{n/2}\Gamma_p\left(\frac{n}{2}\right)\right)\right] \tag{118}$$

with $T = (\log(X), X^2), \eta = ((n-p), V^{-1})$ and $A(n,p,V) = \log\left(2^{np/2}\left|\mathbf{V}\right|^{n/2}\Gamma_p\left(\frac{n}{2}\right)\right)$

11.4.1 Laplace Approximation of the sqrtm-transformed Wishart distribution

Using $\frac{\partial\,\det(X)}{\partial X} = \det(X)(X^{-1})^\top$ and $\frac{\partial}{\partial X} Tr(AX^2) = (AX + XA)^T$ we can calculate the mode by setting the first derivative of the log-pdf to zero

$$\begin{aligned}
\frac{\partial \log f_t(X; n, p, V)}{\partial X} &= \frac{(n-p)\det(X)(X^{-\top})}{\det(X)} - \frac{(V^{-1}X + XV^{-1})^\top}{2} \\
\Rightarrow 0 &= (n-p)X^{-\top} - \frac{(V^{-1}X + XV^{-1})^\top}{2} \\
\Leftrightarrow (n-p)X^{-\top} &= \frac{(V^{-1}X + XV^{-1})^\top}{2} \\
\Leftrightarrow (n-p)X^{-1} &= \frac{(V^{-1}X + XV^{-1})}{2} \\
\Leftrightarrow X &= ???
\end{aligned}$$

THIS IS WHERE SOLVING FOR X GETS COMPLICATED. Maybe we can rewrite it with Kronecker products and vectorized matrices like for the Sylvester equation and these laws `https://en.wikipedia.org/wiki/Vectorization_(mathematics)#Compatibility_with_Kronecker_products`.

So far I have found the following relationships that don't get me any further to the solution of X:

$$
\begin{aligned}
(n-p)X^{-1} &= \frac{(V^{-1}X + XV^{-1})}{2} \\
&\Leftrightarrow C = BXX + XBX \\
&\Leftrightarrow C = (I_p \otimes BX)\text{vec}X + (B^T X^T \otimes I_p)\text{vec}X \\
&\Leftrightarrow C = (B^T X^T \oplus BX)\text{vec}X \\
&\Leftrightarrow C = (BX \oplus BX)\text{vec}X
\end{aligned}
$$

Computing the second derivative by using $\frac{\partial}{\partial X} X^{-1} = -X^{-1} \otimes X^{-1}$, $\frac{\partial}{\partial X}(AX + XA)^\top = I \boxtimes A + A \boxtimes I$:

$$
\begin{aligned}
\frac{\partial^2 \log f_t(X; n, p, V)}{\partial^2 X} &= \frac{\partial}{\partial X}\left[(n-p)X^{-\top} - \frac{(V^{-1}X + XV^{-1})^\top}{2}\right] \\
&= -(n-p)(X^{-\top} \otimes X^{-1}) + \frac{1}{2}(I_p \boxtimes V^{-1} + V^{-1} \boxtimes I_p) \\
&\Rightarrow -(n-p)\left[\sqrt{\frac{1}{(n-p)}}V^{-\frac{1}{2}} \otimes \sqrt{\frac{1}{(n-p)}}V^{-\frac{1}{2}}\right] + \frac{1}{2}\left[I_p \boxtimes V^{-1} + V^{-1} \boxtimes I_p\right] \\
&= -\left(V^{-\frac{1}{2}} \otimes V^{-1}\right) + \frac{1}{2}\left[I_p \boxtimes V^{-1} + V^{-1} \boxtimes I_p\right] \\
&\overset{\cdot-1}{\Rightarrow} \left(V^{-\frac{1}{2}} \otimes V^{-\frac{1}{2}}\right) - \frac{1}{2}\left[I_p \boxtimes V^{-1} + V^{-1} \boxtimes I_p\right] \\
\Rightarrow \Sigma &= \left[V^{-\frac{1}{2}} \otimes V^{-\frac{1}{2}} - \frac{1}{2}\left(I_p \boxtimes V^{-1} + V^{-1} \boxtimes I_p\right)\right]^{-1}
\end{aligned}
$$

We can assume that X is symmetric because the root of a symmetric positive definite matrix is symmetric. This is the solution if we assume that the mode is given by $X = \sqrt{(n-p)}V^{\frac{1}{2}}$.

11.4.2 The Bridge for sqrtm-tranform

we use $\mu = ((n-p)V)^{\frac{1}{2}} \Leftrightarrow \mu^2 = (n-p)V \Leftrightarrow V = \frac{\mu^2}{(n-p)}$. Remember that μ is reshaped to be the same size as V even though we usually think of it in vector-form.

$$
\mu = ((n-p)V)^{\frac{1}{2}} \tag{119}
$$

$$
\Sigma = (V \otimes V)^{\frac{1}{2}} - \tilde{V}^{-1} \tag{120}
$$

$$
V = \frac{\mu^2}{(n-p)} \tag{121}
$$

QUESTION: DO WE ASSUME WE KNOW THE n?

12 Inverse Wishart Distribution

12.1 Standard Inverse Wishart distribution

the pdf of the Inverse Wishart is

$$
f_{\mathbf{x}}(\mathbf{x}; \mathbf{\Psi}, \nu) = \frac{|\mathbf{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})} |\mathbf{x}|^{-(\nu+p+1)/2} e^{-\frac{1}{2}\operatorname{tr}(\mathbf{\Psi}\mathbf{x}^{-1})} \tag{122}
$$

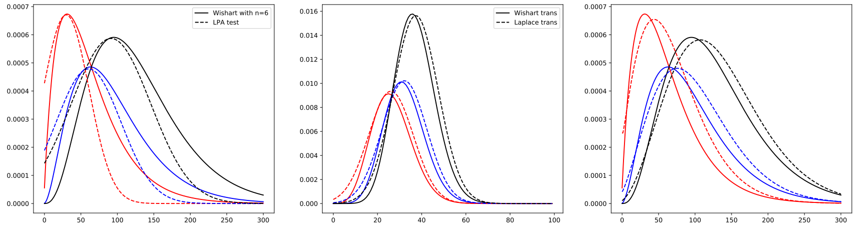

Figure 8: wishart comparison for sqrtm

which can be written as

$$f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\Psi}, \nu) = \exp\left[-(\nu + p + 1)/2 \log(|x|) - \frac{1}{2}\text{tr}(\Psi x^{-1}) + \log(\frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})})\right] \tag{123}$$

with $T = (\log(x), x^{-1}), \eta = (-(\nu + p + 1)/2, \Psi)$ and $A(n, p, V) = -\log(\frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})})$

12.1.1 Laplace Approximation of the standard inverse Wishart distribution

Using ... we can calculate the mode by setting the first derivative of the log-pdf to zero:

$$\begin{aligned}
\frac{\partial \log f_{\mathbf{X}}(\mathbf{X}; \boldsymbol{\Psi}, \nu)}{\partial X} &= \frac{-(\nu + p + 1)\det(X)X^{-\top}}{2\det(X)} + \frac{(X^{-1}\Psi X^{-1})^\top}{2} \\
&= \frac{-(\nu + p + 1)X^{-\top}}{2} + \frac{(X^{-1}\Psi X^{-1})^\top}{2} \\
\Rightarrow 0 &= \frac{-(\nu + p + 1)X^{-\top}}{2} + \frac{(X^{-1}\Psi X^{-1})^\top}{2} \\
\Leftrightarrow (\nu + p + 1)X^{-1} &= X^{-1}\Psi X^{-1} \\
\Leftrightarrow (\nu + p + 1) &= X^{-1}\Psi \\
\Leftrightarrow X &= \frac{1}{\nu + p + 1}\Psi
\end{aligned}$$

Using

$$\frac{\partial (XBX)_{kl}}{\partial X_{ij}} = \delta_{ki}(BX)_{lj} + \delta_{lj}(XB)_{ki} \tag{124}$$

$$\frac{\partial X^{-1}}{\partial X} = -(X^{-1} \otimes X^{-1}) \tag{125}$$

$$\frac{\partial (X^{-1}BX^{-1})}{\partial X} = \frac{\partial (X^{-1}BX^{-1})}{\partial X^{-1}} \frac{\partial X^{-1}}{\partial X} = -[\delta_{ki}(BX^{-1})_{lj} + \delta_{lj}(X^{-1}B)_{ki}](X^{-1} \otimes X^{-1}) \tag{126}$$

$$(AB)^{-1} = B^{-1}A^{-1} \tag{127}$$

we can get the covariance matrix by inverting the Hessian and multiplying with -1.

$$\begin{aligned}
\frac{\partial^2 \log f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\Psi},\nu)}{\partial^2 X} &= \frac{(\nu+p+1)}{2}(X^{-1}\otimes X^{-1})^\top - \frac{[\delta_{ki}(\Psi X^{-1})_{lj} + \delta_{lj}(X^{-1}\Psi)_{ki}]}{2}(X^{-1}\otimes X^{-1})^\top \\
&= \left\{\frac{(\nu+p+1)}{2} I_{n^2} - \frac{[\delta_{ki}(\Psi X^{-1})_{lj} + \delta_{lj}(X^{-1}\Psi)_{ki}]}{2}\right\}(X^{-1}\otimes X^{-1})^\top \\
&\overset{\text{insert mode}}{=} \left\{\frac{(\nu+p+1)}{2} I_{n^2} - \frac{[\delta_{ki}((\nu+p+1)\Psi\Psi^{-1})_{lj} + \delta_{lj}((\nu+p+1)\Psi^{-1}\Psi)_{ki}]}{2}\right\}(\nu+p+1)^2(\psi\otimes \\
&= \left\{\frac{(\nu+p+1)}{2} I_{n^2} - \underbrace{\frac{[\delta_{ki}((\nu+p+1)I_n)_{lj} + \delta_{lj}((\nu+p+1)I_n)_{ki}]}{2}}_{=(\nu+p+1)I_{n^2}}\right\}(\nu+p+1)^2(\psi\otimes\psi)^{-\top} \\
&= \underbrace{-\frac{1}{2}(\nu+p+1)I_{n^2}}_{A}\,\underbrace{(\nu+p+1)^2(\psi\otimes\psi)^{-\top}}_{B} \\
&\overset{\text{invert}}{\Rightarrow} -\frac{1}{(\nu+p+1)^2}(\psi\otimes\psi)^\top \frac{2}{(\nu+p+1)} I_{n^2} \\
&= \overset{\cdot^{-1}}{\Rightarrow} \frac{2}{(\nu+p+1)^3}(\Psi\otimes\Psi)^\top
\end{aligned}$$

where I_n and I_n^2 are the identity matrix of size n and n^2 respectively. We can also ignore the transpose since we are dealing with symmetric positive definite matrices when it comes to the inverse Wishart distribution.

12.2 Logm-Transformed inverse Wishart distribution

we transform the distribution with $g(X) = \text{logm}(X)$, i.e. $g^{-1}(X) = \text{expm}(X)$, where $\mathrm{expm}(X)$ is the matrix exponential. The new pdf becomes

$$\begin{aligned}
W_{logm}(\mathbf{X};\boldsymbol{\Psi},\nu) &= \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})} |\mathrm{expm}\,\mathbf{X}|^{-(\nu+p+1)/2} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}(\mathrm{expm}\,\mathbf{X})^{-1})} \cdot |\,\mathrm{expm}(\mathbf{X})| && (128) \\
&= \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})} |\mathrm{expm}\,\mathbf{X}|^{-(\nu+p-1)/2} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}(\mathrm{expm}\,\mathbf{X})^{-1})} && (129) \\
&= \exp\left[C - (\nu+p-1)/2 \log|\mathrm{expm}\,\mathbf{X}| - -\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\,\mathrm{expm}(-\mathbf{X}))\right] && (130)
\end{aligned}$$

with expfam stats blablabla.

12.2.1 Laplace Approximation of the logm-transformed inverse Wishart distribution

For the first derivative we use the same ideas as for the Wishart (TODO:link).

$$\begin{aligned}
\frac{\partial \log W_{logm}}{\partial X} &= \frac{\partial}{\partial X} - (\nu+p-1)/2 \log|\mathrm{expm}\,\mathbf{X}| - -\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\,\mathrm{expm}(-\mathbf{X})) && (131) \\
&= -\frac{(\nu+p-1)}{2} + \frac{1}{2}\boldsymbol{\Psi}\,\mathrm{expm}(-\mathbf{X}) && (132)
\end{aligned}$$

which yields the mode by setting it to zero and solving for X.

$$
\begin{aligned}
0 &= -\frac{(\nu+p-1)}{2} + \frac{1}{2}\boldsymbol{\Psi}\,\text{expm}(-\mathbf{X}) && (133)\\
\Leftrightarrow (\nu+p-1)I_p &= \boldsymbol{\Psi}(\text{expm}(\mathbf{X}))^{-1} && (134)\\
\Leftrightarrow X &= \text{logm}\left(\frac{\boldsymbol{\Psi}}{n+p-1}\right) && (135)
\end{aligned}
$$

For the second derivative we also use the same ideas as for the Wishart (TODO: link + rewrite).

$$
\begin{aligned}
\frac{\partial^2 \log W_{sqrt}}{\partial^2 X} &= \frac{\partial}{\partial X}\left[-\frac{(\nu+p-1)}{2} + \frac{1}{2}\boldsymbol{\Psi}\,\text{expm}(-\mathbf{X})\right] && (136)\\
&= -\frac{1}{2}\left[\boldsymbol{\Psi}\,\text{expm}(\mathbf{X})^{-1} \otimes I_p\right] && (137)\\
&\overset{\text{mode}}{\Rightarrow} -\frac{1}{2}\left[\boldsymbol{\Psi}\frac{\boldsymbol{\Psi}^{-1}}{(\nu+p-1)} \otimes I_p\right] && (138)\\
&= -\frac{1}{2(\nu+p-1)} I_{p\times p} && (139)\\
\Leftrightarrow \Sigma &= 2(\nu+p-1)I_{p\times p} && (140)
\end{aligned}
$$

12.2.2 The Bridge for the logm-transformed inverse Wishart distribution

We get μ and Ψ from the Laplace approximation and determine V by inverting μ

$$
\mu = \text{logm}\left(\frac{\boldsymbol{\Psi}}{n+p-1}\right) \Leftrightarrow \text{expm}(\mu) = \frac{\boldsymbol{\Psi}}{n+p-1} \Leftrightarrow \boldsymbol{\Psi} = \text{expm}(\mu)(n+p-1) \qquad (141)
$$

In summary:

$$
\begin{aligned}
\mu &= \text{logm}\left(\frac{\boldsymbol{\Psi}}{n+p-1}\right) && (142)\\
\Sigma &= 2(\nu+p-1)I_{p\times p} && (143)\\
\boldsymbol{\Psi} &= (n+p-1)\,\text{expm}(\mu) && (144)
\end{aligned}
$$

where $\boldsymbol{\Psi}$ is reshaped to a $p \times p$ matrix.

12.3 Sqrtm-Transformed inverse Wishart distribution

we transform the distribution with $g(X) = \text{sqrtm}(X) = X^{\frac{1}{2}}$, i.e. $g^{-1}(X) = X^2$, where sqrtm(X) is the square root of the matrix. The new pdf becomes

$$
\begin{aligned}
f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\Psi},\nu) &= \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})}\left|\mathbf{x}^{\mathbf{2}}\right|^{-(\nu+p+1)/2} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\mathbf{x}^{-2})}|2x|\\
&= \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})}|\mathbf{x}|^{-(\nu+p+1)} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\mathbf{x}^{-2})}2^p|x|\\
&= \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})}|\mathbf{x}|^{-(\nu+p)} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\mathbf{x}^{-2})}2^p
\end{aligned}
$$

which can be rewritten as

$$\exp\left[-(\nu+p)\log(|X|) - \frac{1}{2}\text{tr}(\Psi X^{-2}) + \log(C)\right]$$

with $T = (\log(|X|), X^{-2}), \nu = (-(\nu+p), \Psi)$ and $A = ...$

12.3.1 Laplace Approximation of the sqrtm-transformed inverse Wishart distribution

Using

$$\begin{aligned}
d(XBX) &= (dX)BX + XB(dX) = BX + XB \\
d(X^{-1}) &= -X^{-1}dXX^{-1} \\
\frac{\partial X^{-1}BX^{-1}}{\partial X} &= \frac{\partial X^{-1}BX^{-1}}{\partial X^{-1}}\frac{\partial X^{-1}}{\partial X} = (BX^{-1} + X^{-1}B)(X^{-2}) = BX^{-3} + X^{-1}BX^{-2} \\
\frac{\partial \text{tr}(BX^{-2})}{\partial X} &= \frac{\partial \text{tr}(X^{-1}BX^{-1})}{\partial X} = -\text{tr}(BX^{-3} + X^{-1}BX^{-2}) \\
&= -\text{tr}(BX^{-3}) - \text{tr}(X^{-1}BX^{-2}) = -2\text{tr}(BX^{-3}) = -2(BX^{-3})^\top
\end{aligned}$$

we can calculate the mode by setting the derivative of the log-pdf to zero:

$$\begin{aligned}
\frac{\partial \log f_t(X, \Psi, \nu)}{\partial X} &= \frac{-(\nu+p)\det(X)X^{-\top}}{\det(X)} + \frac{2(\Psi X^{-3})^\top}{2} \\
&= -(\nu+p)X^{-\top} + (\Psi X^{-3})^\top \\
\Rightarrow 0 &= -(\nu+p)X^{-\top} + (\Psi X^{-3})^\top \\
\Leftrightarrow (\nu+p)X^{-1} &= (\Psi X^{-3}) \\
\Leftrightarrow (\nu+p)I_n &= (\Psi X^{-2}) \\
\Leftrightarrow (\nu+p)\Psi^{-1} &= X^{-2} \\
\Leftrightarrow X &= \left(\frac{1}{(\nu+p)}\Psi\right)^{\frac{1}{2}}
\end{aligned}$$

Using

$$\begin{aligned}
\frac{\partial (XA)_{kl}}{\partial X_{ij}} &= \delta_{ki}A_{jl} \rightarrow \frac{\partial XA}{\partial X} = I \otimes A \\
\frac{\partial X^3}{\partial X} &= X^2 \otimes I_n + X \otimes X + I_n \otimes X \\
\frac{\partial X^{-1}}{\partial X} &= -X^{-1} \otimes X^{-1} \\
(\Psi X^{-3})^\top &= (X^{-3})^\top \Psi^\top \overset{\text{symmetry}}{=} X^{-3}\Psi \\
\frac{\partial (\Psi X^{-3})^\top}{\partial X} &= \frac{\partial (\Psi X^{-3})^\top}{\partial X} = \frac{\partial \Psi X^{-3})^\top}{\partial X^{-3}}\frac{\partial (X^{-1})^3}{\partial X^{-1}}\frac{\partial X^{-1}}{\partial X} \\
&= (I \otimes \Psi)(X^{-2} \otimes I_n + X^{-1} \otimes X^{-1} + I_n \otimes X^{-2})(-X^{-1} \otimes X^{-1})
\end{aligned}$$

where $\hat{\Psi} \in \mathbb{R}^{n^2 \times n^2}$ is the matrix given by $\delta_{ki}A_{jl}$.

we can calculate the covariance matrix by inverting the Hessian and multiplying it with -1.

$$\begin{aligned}
\frac{\partial^2 \log f_t(X,\Psi,\nu)}{\partial^2 X} &= \frac{\partial}{\partial X} - (\nu+p)X^{-\top} + (\Psi X^{-3})^{\top} \\
&\overset{\text{symmetry}}{=} (\nu+p)(X^{-1}\otimes X^{-1}) - (I\otimes\Psi)(X^{-2}\otimes I_n + X^{-1}\otimes X^{-1} + I_n\otimes X^{-2})(-X^{-1}\otimes X^{-1}) \\
&= (\nu+p)(X^{-1}\otimes X^{-1}) - (I\otimes\Psi)(X^{-2}\otimes I_n + X^{-1}\otimes X^{-1} + I_n\otimes X^{-2})(X^{-1}\otimes X^{-1}) \\
&= [(\nu+p)I_{n^2} - (X^{-2}\otimes\Psi + X^{-1}\otimes\Psi X^{-1} + I_n\otimes\Psi X^{-2})](X^{-1}\otimes X^{-1}) \\
&\overset{\text{insert mode}}{=} [(\nu+p)I_{n^2} - ((\nu+p)\Psi^{-1}\otimes\Psi + \sqrt{(\nu+p)}\Psi^{-\frac{1}{2}}\otimes\sqrt{(\nu+p)}\Psi\Psi^{-\frac{1}{2}} + I_n\otimes(\nu+p)\Psi\Psi^{-1})](\nu+p)(\Psi^{-\frac{1}{2}}\otimes\Psi^{-\frac{1}{2}}) \\
&= (\nu+p)[(\nu+p)I_{n^2} - (\nu+p)(\Psi^{-1}\otimes\Psi + \Psi^{-\frac{1}{2}}\otimes\Psi^{\frac{1}{2}} + I_n\otimes I_n)](\Psi^{-\frac{1}{2}}\otimes\Psi^{-\frac{1}{2}}) \\
&= -(\nu+p)^2[-I_{n^2} + \Psi^{-1}\otimes\Psi + \Psi^{-\frac{1}{2}}\otimes\Psi^{\frac{1}{2}} + I_{n^2}](\Psi^{-\frac{1}{2}}\otimes\Psi^{-\frac{1}{2}}) \\
&= -(\nu+p)^2[\Psi^{-1}\otimes\Psi + \Psi^{-\frac{1}{2}}\otimes\Psi^{\frac{1}{2}}](\Psi^{-\frac{1}{2}}\otimes\Psi^{-\frac{1}{2}}) \\
&= -(\nu+p)^2[\Psi^{-\frac{3}{2}}\otimes\Psi^{\frac{1}{2}} + \Psi^{-1}\otimes I_n] \\
&= -(\nu+p)^2[\Psi^{-\frac{1}{2}}\Psi^{-1}\otimes\Psi^{\frac{1}{2}}I_n + \Psi^{-1}\otimes I_n] \\
&= -(\nu+p)^2[(\Psi^{-\frac{1}{2}}\otimes\Psi^{\frac{1}{2}})(\Psi^{-1}\otimes I_n) + (\Psi^{-1}\otimes I_n)] \\
&= -(\nu+p)^2[(\Psi^{-\frac{1}{2}}\otimes\Psi^{\frac{1}{2}} + I_{n^2})(\Psi^{-1}\otimes I_n)] \\
&\overset{\text{invert}}{\Rightarrow} -\frac{1}{(\nu+p)^2}(\Psi^{1}\otimes I_n)(\Psi^{-\frac{1}{2}}\otimes\Psi^{\frac{1}{2}} + I_{n^2})^{-1} \\
&\overset{\cdot^{-1}}{\Rightarrow} \frac{1}{(\nu+p)^2}(\Psi^{1}\otimes I_n)(\Psi^{-\frac{1}{2}}\otimes\Psi^{\frac{1}{2}} + I_{n^2})^{-1}
\end{aligned}$$

TODO: add inversion for second part.

12.3.2 The Bridge for the sqrtm-transformed inverse Wishart distribution

13 Math

Some equations to explain what the hell is going on in our derivations

TODO

Notation:

$$\begin{aligned}
x &: \text{variable of the probability distribution in standard basis} \\
\pi &: \text{variable of the probability distribution in standard basis (if the space is a probability space)} \\
y &: \text{variable of the probability distribution in transformed basis} \\
u &: \text{helper variable for non-trivial transformations} \\
\phi(x) &: \text{sufficient statistics} \\
w &: \text{natural parameters} \\
Z(w) &: \text{partition function/normalization constant} \\
h(x) &: \text{base measure}
\end{aligned}$$