Gradient Calculations for Factor Analysis and Selected Variations

Jason D. M. Rennie
jrennie@gmail.com

September 12, 2006

Abstract

We derive the math (objective and gradients) for three closely-related Gaussian-based data models. All three are based on Factor Analysis (FA). The first is a simplification of FA; the second is FA itself; the third is our extension of FA to use a smooth trace norm prior on the factor loading matrix (in place of the hard rank constraint used by FA).

See [3] for some background on Factor Analysis (FA). See [1] for help with matrix calculus.

1 Diagonal Covariance Matrix

We begin with the problem of learning parameters for a multivariate normal with diagonal covariance matrix, $N(\boldsymbol{\mu}, \text{diag}(1/\boldsymbol{\psi}))$. We use $\boldsymbol{\mu} \in \mathbb{R}^d$ to parameterize the mean and $\boldsymbol{\psi} \in \mathbb{R}^d$ to parameterize (the diagonal of) the inverse covarance matrix. The likelihood of a set of n examples, $X \in \mathbb{R}^{n \times d}$, is

$$P(X|\boldsymbol{\mu}, \boldsymbol{\psi}) = \prod_i \frac{\prod_j \sqrt{\psi_j}}{(2\pi)^{d/2}} \exp\left(-\sum_j \frac{\psi_j (X_{ij} - \mu_j)^2}{2}\right). \tag{1}$$

The negative log-likelihood is

$$-\log P(X|\boldsymbol{\mu}, \boldsymbol{\psi}) = \frac{1}{2}\left(nd \log 2\pi + \sum_{i,j} \left[\psi_j (X_{ij} - \mu_j)^2 - \log \psi_j\right]\right). \tag{2}$$

1.1 Maximum Likelihood

Given a set of data, $X \in \mathbb{R}^{n \times d}$, one way of learning the parameters is via maximum likelihood (ML). This works well for the mean parameter, $\boldsymbol{\mu}$, but if one dimension of the data (one column of X) is constant, the empirical variance is null and the inverse variance does not exist (or is infinite). Though we can technically work around this issue, a model trained on data with a zero-variance dimension will assign zero density to any data with a value different than the mean in the zero-variance dimension. We see this as a compelling reason to use an alternate estimation method for the inverse covariance parameter, $\boldsymbol{\psi}$.

1.2 Maximum a Posteriori

An alternative to maximum likelihood (ML) is maximum a posteriori (MAP). Whereas ML selects the parameters which maximize the likelihood, MAP selects parameters to maximize the posterior. We maximize the posterior by maximizing a product of the likelihood and a prior. The posterior incorporates a parameter prior, in our case a prior on inverse variance. An important decision in MAP learning is the selection of the prior.

1.2.1 Selecting A Prior

The Wishart distribution,

$$P(W|V,n) = \frac{|W|^{(n-d-1)/2}|V|^{n/2}}{2^{nd/2}\Gamma_d(n/2)} \exp(-\mathbf{Tr}(VW/2)), \tag{3}$$

where $W \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{d\times d}$ are positive definite and $n \geq d$, is the conjugate prior for Gaussian inverse covariance, so it is a natural choice. However, the Wishart is normalized over the set of $d \times d$ positive definite matrices. We must re-normalize in order to use it for our problem where we assume a diagonal covariance matrix. Note that if W is diagonal, then off-diagonal entries of V have no impact on the likelihood—we might as well assume V to be diagonal too. Let $W \equiv \text{diag}(\boldsymbol{\psi})$ and $V \equiv \text{diag}(\mathbf{v})$. The pdf of our "diagonal" Wishart distribution is

$$P(\boldsymbol{\psi}|\mathbf{v},n) \propto \exp(-\sum_i \psi_i v_i)\prod_i \psi_i^{n-1} = \prod_i \psi_i^{n-1} e^{-v_i\psi_i} \tag{4}$$

Note that we have simplified the use of the n parameter. More importantly, note that, if we allow $n \in \mathbb{R}_+$, this distribution is simply a product of gammas. I.e. the normalization factor is

$$\int \prod_i \psi_i^{n-1} e^{-v_i\psi_i} d\mathbf{v} d\boldsymbol{\psi} = \prod_i \frac{v_i^n}{\Gamma(v_i)}. \tag{5}$$

We have shown that the Wishart is a generalization of the Gamma distribution and that a product of Gammas is the conjuate prior for the Gaussian inverse covariance when the covariance matrix is constrained to be diagonal. Hence, we utilize the product of Gammas as our parameter prior. Next, we discuss its use with MAP.

1.2.2 Maximizing the Posterior

Use of a Wishart prior (or product of Gammas) has the effect of adding "simulated" examples to our data set. We will use two numbers to parameterize the prior. The first, α, represents the effective sample size of the simulated examples; the second, β, represents the sum of squared differences to the mean of the simulated examples[1]. $\frac{\beta}{\alpha}$ represents the empirical variance of (each dimension of) the simulated examples. We use a non-traditional parameterization of the Gamma to align these meanings with the parameters,

$$P(\boldsymbol{\psi}|\alpha,\beta) = \prod_j \frac{(\beta/2)^{(\alpha/2+1)}}{\Gamma(\alpha/2+1)} \psi_j^{\alpha/2} e^{-\psi_j \beta/2}. \tag{6}$$

To reconcile this parameterization with equations 4 & 5, make the substitutions $\alpha/2 := n-1$ and $\beta/2 := v_i \ \forall i$.

The product of likelihood and prior gives us the joint distribution,

$$P(X,\boldsymbol{\psi}|\alpha,\beta) = P(X|\boldsymbol{\psi})P(\boldsymbol{\psi}|\alpha,\beta). \tag{7}$$

The parameters which maximize the joint are the same as those which maximize the posterior. The parameters of the prior, α and β, are known as "hyper-parameters."

1.2.3 Derivatives

Estimation of the parameters for MAP can easily be accomplished via gradient descent of the negative log-probability of the joint. Though, as we will find, the MAP parameter solution can be read off from the derivatives. The joint negative log-probability is

$$\begin{aligned} J &= -\log P(X,\boldsymbol{\psi}|\alpha,\beta) && (8)\\ &= C + \sum_{j=1}^d \frac{\psi_j}{2}\left(\beta + \sum_{i=1}^n (X_{ij}-\mu_j)^2\right) - \sum_{j=1}^d \frac{n+\alpha}{2}\log\psi_j, \end{aligned}$$

where $C = d\left[\log\Gamma(\alpha/2+1) - (\alpha/2+1)\log(\beta/2)\right] + \frac{nd}{2}\log(2\pi)$. The gradient with respect to the mean for dimension j is

$$\frac{\partial J}{\partial \mu_j} = \psi_j \sum_{i=1}^n (X_{ij}-\mu_j). \tag{9}$$

Note that this gradient is zero when μ_j is set to the empirical mean,

$$\boldsymbol{\mu}^* = \frac{1}{n}\sum_{i=1}^n X_i. \tag{10}$$

[1] We use a single β for all dimensions; when additional information about the data is available, one might instead allow a different value for each dimension, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)$.

This is due to the fact that we are simply maximizing likelihood with respect to the mean parameter. The gradient with respect to the inverse covariance parameter is

$$\frac{\partial J}{\partial \psi_j} = \frac{1}{2}\left[\beta + \sum_{i=1}^{n}(X_{ij} - \mu_j)^2 - \frac{n+\alpha}{\psi_j}\right]. \tag{11}$$

As with the mean parameter, we can directly solve for the inverse variance parameter, ψ_j by setting the gradient to zero, the only caveat being that the mean parameter value must be given. However, the inverse variance parameter is not set to the inverse of the empirical variance. Rather, it is set to a simulated inverse empirical variance, where α simulated examples are added to the data set; the α simulated examples each have a variance of $\frac{\beta}{\alpha}$. The MAP parameter setting is the inverse empirical variance of this combination of regular and simulated data,

$$\boldsymbol{\psi}^* = \frac{n+\alpha}{\beta + \sum_{i=1}^{n}(X_i - \boldsymbol{\mu}^*)^2}. \tag{12}$$

This is known to be a biased estimate of inverse variance. The unbiased estimate is achieved by reducing the example count by one (using $n + \alpha - 1$ in the numberator). However, we will use the biased estimate; our estimation of the hyper-parameters (next section) will compensate for error in the ML inverse variance estimate.

1.3 Hyper-Parameters

We can use leave-one-out cross-validation (LOOCV) on the training set to select values for the hyper-parameters, α and β. For LOOCV, we maximize the likelihood of the data, which is a product of the likelihoods for the individual examples. But, when we calculate the likelihood for a given example, we exclude that example in the data set used to calcualte the MAP parameters. Hence, the name, "leave one out."

Let $\boldsymbol{\mu}^{\setminus i}$ and $\boldsymbol{\psi}^{\setminus i}$ denote the leave-one-out estimates for example i. Then, the LOOCV negative log-likelihood is

$$\begin{aligned} J_{\mathrm{L}} &= -\log P(X|\boldsymbol{\mu}^{\setminus i}, \boldsymbol{\psi}^{\setminus i}) \\ &= \frac{nd}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\left[\psi_j^{\setminus i}\left(X_{ij} - \mu_j^{\setminus i}\right)^2 - \log \psi_j^{\setminus i}\right], \end{aligned} \tag{13}$$

where $\mu_j^{\setminus i} = \frac{1}{n-1}\sum_{k\neq i} X_{lj}$, and $\psi_j^{\setminus i} = \frac{n+a-1}{\beta + \sum_{k\neq i}\left(X_{kj} - \mu_j^{\setminus i}\right)^2}$. We can solve for the hyper-parameters α and β via gradient descent. The gradients are

$$\frac{\partial J_{\mathrm{L}}}{\partial \alpha} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\left[\frac{\left(X_{ij} - \mu_j^{\setminus i}\right)^2}{\beta + \sum_{k\neq i}\left(X_{kj} - \mu_j^{\setminus i}\right)^2} - \frac{1}{n+\alpha-1}\right], \tag{14}$$

and

$$\frac{\partial J_{\mathrm{L}}}{\partial \beta} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\left[\frac{1}{\beta + \sum_{k\neq i}\left(X_{kj} - \mu_j^{\setminus i}\right)^2} - \frac{(n+\alpha-1)\left(X_{ij} - \mu_j^{\setminus i}\right)^2}{\left[\beta + \sum_{k\neq i}\left(X_{kj} - \mu_j^{\setminus i}\right)^2\right]^2}\right]. \tag{15}$$

In other words, to achieve the minimum value of J_L, α and β must be chosen so that the average product of leave-one-out empirical variance with the leave-one-out inverse variance parameter is unity,

$$\frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{(n+\alpha-1)\left(X_{ij} - \mu_j^{\setminus i}\right)^2}{\beta + \sum_{k\neq i}\left(X_{kj} - \mu_j^{\setminus i}\right)^2} = 1. \tag{16}$$

The fact that the gradients yield only a single equation constraint indicates that the solution is under-determined. We find a solution by setting $\alpha = 1$ and solving for β.

2 Factor Analysis

2.1 Notation

a, α	scalar
$\mathbf{a}$, $\boldsymbol{\alpha}$	(row) vector
A	matrix
$\mathrm{diag}(A)$	diagonal of A, taken as a row vector
$\mathrm{diag}(\boldsymbol{\alpha})$	diagonal matrix, with diagonal elements taken from $\boldsymbol{\alpha}$
A_{ab}	scalar from a^{th} row, b^{th} column of A
A^{-1}_{ab}	scalar from a^{th} row, b^{th} column of A^{-1}
$A_{\cdot b}$	b^{th} column of A (as a column vector)
$A_{a\cdot}$	a^{th} row of A (as a row vector)
AB	matrix multiplication
$\mathbf{a}\mathbf{b}^T$	vector product
$A * B$, $\mathbf{a} * \mathbf{b}$	element-wise multiplication
A/B, $\mathbf{a}/\mathbf{b}$	element-wise division
$A + \mathbf{a}$, $A - \mathbf{a}$	add/subtract $\mathbf{a}$ from each row of A
$A * \mathbf{a}$, $\frac{A}{\mathbf{a}}$	multiply/divide each row of A by $\mathbf{a}$

2.2 Definitions, Notes, and Sizes

- $X \in \mathbb{R}^{n\times d}$
- $\boldsymbol{\sigma} \in \mathbb{R}^d$

- $\Lambda \in \mathbb{R}^{d \times k}$
- $S \equiv \Lambda\Lambda^T + \mathrm{diag}(\boldsymbol{\psi}) \in \mathbb{R}^{d \times d}$; S is symmetric, $S = S^T$

2.3 Introduction

Here we consider the Factor Analysis model where the covariance matrix is a sum of two matrices: (1) a diagonal matrix, and (2) a low-rank matrix. The purpose of this section is to establish the math necessary to learn parameters via gradient descent so that we can easily replace the hard rank constraint with a soft trace norm prior (§ 3).

We use $\boldsymbol{\sigma} \in \mathbb{R}^d$ to parameterize the diagonal matrix and $\Lambda \in \mathbb{R}^{d \times k}$ to parameterize the low-rank matrix. Λ is known as the "factor loading matrix." Our data likelihood is a Gaussian with covariance matrix $\Lambda\Lambda^T + \mathrm{diag}(\boldsymbol{\sigma}^2)$. Define $S \equiv \Lambda\Lambda^T + \mathrm{diag}(\boldsymbol{\sigma}^2)$. Note that S is symmetric ($S = S^T$). The likelihood of a set of n data points is

$$P(X|\boldsymbol{\mu}, \boldsymbol{\sigma}, \Lambda) = \frac{1}{(2\pi)^{nd/2}|S|^{n/2}} \exp\left(-\frac{1}{2}\sum_i (X_{i\cdot} - \boldsymbol{\mu}) S^{-1} (X_{i\cdot} - \boldsymbol{\mu})^T\right). \quad (17)$$

2.4 Learning Parameters

As with the diagonal-covariance Gaussian model, there is concern that data which is constant in a certain dimension will lead to a model which will reject any new data which differs in that dimension. To temper the model, we apply a common Gamma prior (6) to each entry of the diagonal of the inverse covariance matrix. The pdf is

$$\prod_j P(S_{jj}^{-1}|\alpha, \beta) = \prod_j \frac{(\beta/2)^{(\alpha/2+1)}}{\Gamma(\alpha/2+1)} (S_{jj}^{-1})^{\alpha/2} e^{-S_{jj}^{-1}\beta/2}. \quad (18)$$

As before, we do not use a prior on the mean parameter ($\boldsymbol{\mu}$). Note that the limited rank of Λ provides a form of regularization on the inverse covariance matrix, S^{-1}. Further regularization is not customarily used; we follow suit.

We learn parameters by maximizing the posterior (product of likelihood and prior) or equivalently, minimizing the negative log-posterior,

$$\begin{aligned} J &= P(X|\boldsymbol{\mu}, \boldsymbol{\sigma}, \Lambda) P(\mathrm{diag}(S^{-1})|\alpha, \beta) \quad (19) \\ &= C + \frac{1}{2}\left(n \log|S| + \sum_{a=1}^{n} (X_a - \boldsymbol{\mu}) S^{-1} (X_a - \boldsymbol{\mu})^T + \sum_{a=1}^{d} \left[\beta S_{aa}^{-1} - \alpha \log S_{aa}^{-1}\right]\right), \end{aligned}$$

where C is a constant (not a function of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, or Λ). We use gradient descent to optimize the parameters. This requires calculation of the first-order gradient of the objective with respect to the parameters. To simplify the full gradient calculation, we break the objective into four parts:

$$J = C + \frac{1}{2}(J_1 + J_2 + J_3 + J_4), \quad (20)$$

where

$$J_1 = n \log |S| \tag{21}$$

$$J_2 = \sum_a (X_a - \boldsymbol{\mu}) S^{-1} (X_a - \boldsymbol{\mu})^T = \sum_{a,b,c} [X_a - \boldsymbol{\mu}]_b S^{-1}_{bc} [X_a - \boldsymbol{\mu}]_c. \tag{22}$$

$$J_3 = \beta \sum_a S^{-1}_{aa} \tag{23}$$

$$J_4 = -\alpha \sum_a \log S^{-1}_{aa} \tag{24}$$

We first establish a number of intermediate partial derivative calculations:

- $\frac{\partial S_{kl}}{\partial \sigma_j} = \begin{cases} 2\sigma_j & \text{if } j = k = l, \\ 0 & \text{otherwise} \end{cases}$
- $\frac{\partial S_{kl}}{\partial \Lambda_{ij}} = \frac{\partial (\Lambda \Lambda^T)_{kl}}{\partial \Lambda_{ij}} = \frac{\partial (\sum_a \Lambda_{ka} \Lambda_{la})}{\partial \Lambda_{ij}} = \delta_{k=i} \Lambda_{lj} + \Lambda_{kj} \delta_{l=i}$
- $\frac{\partial a^T M a}{\partial M_{ij}} = \sum_{k,l} a_k \frac{\partial M_{kl}}{\partial M_{ij}} a_l = a_i a_j$
- $\frac{\partial (S^{-1})_{kl}}{\partial S_{ij}} = -\left(S^{-1} \frac{\partial S}{\partial S_{ij}} S^{-1} \right)_{kl} = -S^{-1}_{ki} S^{-1}_{jl}$ (page 8 of [4])
  Note: $\frac{\partial (ABA)_{kl}}{\partial B_{ij}} = \frac{\partial \left(\sum_{a,b} A_{ka} B_{ab} A_{bl} \right)}{\partial B_{ij}} = A_{ki} A_{jl}$
- $\frac{\partial \log |S|}{\partial S_{ij}} = S^{-1}_{ji}$ (page 7 of [4], eqn. 10)
  Note: $\partial \log |S| = \mathbf{Tr}(S^{-1} \partial S) = \sum_{i,j} S^{-1}_{ji} \partial S_{ij}$

Partial derivative of J_1 with respect to (wrt) $\boldsymbol{\sigma}$:

$$\frac{\partial J_1}{\partial \sigma_j} = n \sum_{k,l} \frac{\partial \log |S|}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \sigma_j} = 2n \frac{\partial \log |S|}{\partial S_{jj}} \sigma_j = 2n S^{-1}_{jj} \sigma_j \tag{25}$$

$$\frac{\partial J_1}{\partial \boldsymbol{\sigma}} = 2n \boldsymbol{\sigma} * \operatorname{diag}(S^{-1}) \tag{26}$$

Partial derivative of J_1 wrt Λ:

$$\frac{\partial J_1}{\partial \Lambda_{ij}} = n \sum_{k,l} \frac{\partial \log |S|}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \Lambda_{ij}} = n \sum_{k,l} S^{-1}_{lk} (\delta_{k=i} \Lambda_{lj} + \Lambda_{kj} \delta_{l=i}) \tag{27}$$

$$= n \left[\sum_l S^{-1}_{il} \Lambda_{lj} + \sum_k S^{-1}_{ik} \Lambda_{kj} \right] = 2n S^{-1}_{i\cdot} \Lambda_{\cdot j} \tag{28}$$

$$\frac{\partial J_1}{\partial \Lambda} = 2n S^{-1} \Lambda \tag{29}$$

Partial derivative of J_2 wrt $\boldsymbol{\mu}$:

$$\frac{\partial J_2}{\partial \mu_j} = -\sum_{a,c} S^{-1}_{jc} [X_a - \boldsymbol{\mu}]_c - \sum_{a,b} [X_a - \boldsymbol{\mu}]_b S^{-1}_{bj} = -2 \sum_{a,b} [X_a - \boldsymbol{\mu}]_b S^{-1}_{bj} \tag{30}$$

$$\frac{\partial J_2}{\partial \boldsymbol{\mu}} = -2 \sum_a [X_a - \boldsymbol{\mu}] S^{-1}_{\cdot j} \tag{31}$$

Partial derivative of J_2 wrt $\boldsymbol{\sigma}$:

$$\begin{aligned}
\frac{\partial J_2}{\partial \sigma_j} &= \sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c \frac{\partial S^{-1}_{bc}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \sigma_j} = 2 \sum_{a,b,c} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c \frac{\partial S^{-1}_{bc}}{\partial S_{jj}} \sigma_j \\
&= -2 \sum_{a,b,c} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bj} S^{-1}_{jc} \sigma_j = -2 \sum_{a} S^{-1}_{j.} [X_a - \boldsymbol{\mu}]^T [X_a - \boldsymbol{\mu}] S^{-1}_{.j} \sigma_j \\
&= -2 S^{-1}_{j.} [X - \boldsymbol{\mu}]^T [X - \boldsymbol{\mu}] S^{-1}_{.j} \sigma_j && (32) \\
\frac{\partial J_2}{\partial \boldsymbol{\sigma}} &= -2\boldsymbol{\sigma} * \operatorname{diag}(S^{-1} [X - \boldsymbol{\mu}]^T [X - \boldsymbol{\mu}] S^{-1}) && (33)
\end{aligned}$$

Partial derivative of J_2 wrt Λ:

$$\begin{aligned}
\frac{\partial J_2}{\partial \Lambda_{ij}} &= \sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c \frac{\partial S^{-1}_{bc}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \Lambda_{ij}} && (34) \\
&= - \sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bk} S^{-1}_{lc} (\delta_{k=i} \Lambda_{lj} + \Lambda_{kj} \delta_{l=i}) && (35) \\
&= - \sum_{a,b,c,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bi} S^{-1}_{lc} \Lambda_{lj} - \sum_{a,b,c,k} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bk} S^{-1}_{ic} \Lambda_{kj} \\
&= - \sum_{a,b,c,l} S^{-1}_{ib} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{cl} \Lambda_{lj} - \sum_{a,b,c,k} S^{-1}_{ic} [X_a - \boldsymbol{\mu}]_c [X_a - \boldsymbol{\mu}]_b S^{-1}_{bk} \Lambda_{kj} \\
&= -2 \sum_{a} S^{-1}_{i.} [X_a - \boldsymbol{\mu}]^T [X_a - \boldsymbol{\mu}] S^{-1} \Lambda_{.j} && (36) \\
&= -2 S^{-1}_{i.} [X - \boldsymbol{\mu}]^T [X - \boldsymbol{\mu}] S^{-1} \Lambda_{.j} && (37) \\
\frac{\partial J_2}{\partial \Lambda} &= -2 S^{-1} [X - \boldsymbol{\mu}]^T [X - \boldsymbol{\mu}] S^{-1} \Lambda && (38)
\end{aligned}$$

Partial derivative of J_3 wrt $\boldsymbol{\sigma}$:

$$\begin{aligned}
\frac{\partial J_3}{\partial \sigma_j} &= \beta \sum_{a,k,l} \frac{\partial S^{-1}_{aa}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \psi_j} = 2\beta \sum_{a} \frac{\partial S^{-1}_{aa}}{\partial S_{jj}} \sigma_j = -2\beta \sum_{a} S^{-1}_{aj} S^{-1}_{ja} \sigma_j && (39) \\
&= -2\beta S^{-1}_{j.} S^{-1}_{.j} \sigma_j && (40) \\
\frac{\partial J_3}{\partial \boldsymbol{\sigma}} &= -2\beta\boldsymbol{\sigma} * \operatorname{diag}(S^{-1} S^{-1}) = -2\beta\boldsymbol{\sigma} * \operatorname{diag}(S^{-2}) && (41)
\end{aligned}$$

Partial derivative of J_3 wrt Λ:

$$\begin{aligned}
\frac{\partial J_3}{\partial \Lambda_{ij}} &= \beta \sum_{a,k,l} \frac{\partial S^{-1}_{aa}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \Lambda_{ij}} = -\beta \sum_{a,k,l} S^{-1}_{ak} S^{-1}_{la} (\delta_{k=i} \Lambda_{lj} + \Lambda_{kj} \delta_{l=i}) && (42) \\
&= -\beta \left[\sum_{a,l} S^{-1}_{ia} S^{-1}_{al} \Lambda_{lj} + \sum_{a,k} S^{-1}_{ia} S^{-1}_{ak} \Lambda_{kj} \right] = -2\beta S^{-1} S^{-1} \Lambda && (43) \\
\frac{\partial J_3}{\partial \Lambda} &= -2\beta S^{-2} \Lambda && (44)
\end{aligned}$$

Partial derivative of J_4 wrt $\boldsymbol{\sigma}$:

$$\frac{\partial J_4}{\partial \sigma_j} = -\alpha \sum_a \frac{\partial \log S^{-1}_{aa}}{\partial S^{-1}_{aa}} \sum_{k,l} \frac{\partial S^{-1}_{aa}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \sigma_j} = -2\alpha \sum_a \frac{1}{S^{-1}_{aa}} \frac{\partial S^{-1}_{aa}}{\partial S_{jj}} \sigma_j \tag{45}$$

$$= 2\alpha \sum_a \frac{S^{-1}_{aj} S^{-1}_{ja}}{S^{-1}_{aa}} \sigma_j = 2\alpha \frac{S^{-1}_{j\cdot}}{\mathrm{diag}(S^{-1})} S^{-1}_{\cdot j} \sigma_j \tag{46}$$

$$\frac{\partial J_4}{\partial \boldsymbol{\sigma}} = 2\alpha \boldsymbol{\sigma} * \mathrm{diag}\left(\frac{S^{-1}}{\mathrm{diag}(S^{-1})} S^{-1} \right) \tag{47}$$

Partial derivative of J_4 wrt Λ:

$$\frac{\partial J_4}{\partial \Lambda_{ij}} = -\alpha \sum_a \frac{\partial \log S^{-1}_{aa}}{\partial S^{-1}_{aa}} \sum_{k,l} \frac{\partial S^{-1}_{aa}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \Lambda_{ij}} \tag{48}$$

$$= \alpha \sum_a \frac{1}{S^{-1}_{aa}} \sum_{k,l} S^{-1}_{ak} S^{-1}_{la} (\delta_{k=i} \Lambda_{lj} + \Lambda_{kj} \delta_{l=i}) \tag{49}$$

$$= \alpha \sum_a \frac{1}{S^{-1}_{aa}} \left[\sum_l S^{-1}_{ia} S^{-1}_{al} \Lambda_{lj} + \sum_k S^{-1}_{ia} S^{-1}_{ak} \Lambda_{kj} \right] \tag{50}$$

$$= 2\alpha \frac{S^{-1}_{i\cdot}}{\mathrm{diag}(S^{-1})} S^{-1} \Lambda_{\cdot j} \tag{51}$$

$$\frac{\partial J_4}{\partial \Lambda} = 2\alpha \frac{S^{-1}}{\mathrm{diag}(S^{-1})} S^{-1} \Lambda \tag{52}$$

3 Trace Norm Prior

Factor analysis utilizes a rank constraint on the factor loading matrix (Λ) to provide regularization so that the model does not overfit the data. While the constraint serves its purpose, there are two aspects that make it somewhat undesirable. The first is that the hard rank constraint introduces a non-convexity. Note that the convex combination of two rank-one matrices may yield a rank-two matrix. Optimization of the rank-constrained objective may be much more susceptible to local minima than the unconstrained objective. The second undesirable trait is that the parameter of the constraint is discrete. It is quite possible, if not likely, that the optimal (in terms of generalization) constraint be "between" two discrete values (in, for example, the case that we could "relax" the rank parameter).

Here we address these two issues by introducing a new form of regularization on the factor loading matrix. Instead of placing a hard constraint on the rank of the factor loading matrix, we introduce a prior which encourages low rank. We call this the "trace norm prior" [5] since it acts as a penalty on the trace norm[2] of the matrix,

$$P_\Sigma(X|\lambda) \propto \exp(-\lambda \|X\|_\Sigma). \tag{53}$$

[2] The trace norm of a matrix, X, denoted $\|X\|_\Sigma$ or $\|X\|_{\mathrm{KF}}$, is the sum of its singular values.

We include this prior in our posterior. The updated negative log-posterior is

$$\begin{aligned} J &= P(X|\boldsymbol{\mu}, \boldsymbol{\sigma}, \Lambda) P(\mathrm{diag}(S^{-1})|\alpha, \beta) P(\Lambda|\lambda) \qquad (54) \\ &= C + \frac{1}{2}\left(n \log|S| + \sum_{a=1}^{n} (X_a - \boldsymbol{\mu}) S^{-1} (X_a - \boldsymbol{\mu})^T + \sum_{a=1}^{d} \left[\beta S_{aa}^{-1} - \alpha \log S_{aa}^{-1}\right] \right) + \lambda \|\Lambda\|_\Sigma, \end{aligned}$$

where, again, C is constant with respect to $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and Λ.

The use of the trace norm for solving rank-constrained problems was introduced by [2] and has been successfully been applied to the task of collaborative filtering [7, 6]. Since we are maximizing the posterior and thus need not calculate the normalization constant, our use of the trace norm is nearly identical to that of [6].

As noted by [6], we cannot calculate a gradient for J_5 like we have the other J_i:

> $\|X\|_\Sigma$ is a complicated non-differentiable function for which it is not eash to find the subdif[fe]rential. Finding good descent directions for [the trace norm] is not easy.

But, as [6] conclude, we can substitute the trace norm with a variational bound (for which we can easily calculate the gradient) and optimize this alternate objective. Though it is unclear whether this methodology provides the same solution as direct optimization of the trace norm, results from [6] indicate that the variational bound is highly effective.

The variational bound is

$$\|UV^T\| \leq \frac{1}{2}\left(\|U\|_{\mathrm{Fro}}^2 + \|V\|_{\mathrm{Fro}}^2\right), \qquad (55)$$

where $\|U\|_{\mathrm{Fro}}$ is the Frobenius norm—the square root of sum of squared entries—of U. To apply this to our objective, we make the substitution $\Lambda \equiv UV^T$ and optimize U and V in place of Λ,

$$\begin{aligned} J(\boldsymbol{\mu}, \boldsymbol{\sigma}, UV^T) &\leq J'(\boldsymbol{\mu}, \boldsymbol{\sigma}, U, V) \qquad (56) \\ &= C + \frac{1}{2}\Bigg[n \log|S| + \sum_{a=1}^{n} (X_a - \boldsymbol{\mu}) S^{-1} (X_a - \boldsymbol{\mu})^T \\ &\qquad + \sum_{a=1}^{d} \left[\beta S_{aa}^{-1} - \alpha \log S_{aa}^{-1}\right] + \lambda\big(\|U\|_{\mathrm{Fro}}^2 + \|V\|_{\mathrm{Fro}}^2\big)\Bigg], \end{aligned}$$

where $S \equiv UV^TVU^T + \mathrm{diag}(\boldsymbol{\sigma})$. To update our organization of the objective, we define $J_5 \equiv \lambda(\|U\|_{\mathrm{Fro}}^2 + \|V\|_{\mathrm{Fro}}^2)$. Then, $J = C + \frac{1}{2}(J_1 + J_2 + J_3 + J_4 + J_5)$. All that remains is to calcualte the gradients of $J_1, \ldots, J_5$ with respect to U and V.

Note that

- $\frac{\partial S_{kl}}{\partial U_{ij}} = \frac{\partial (UV^TVU^T)_{kl}}{\partial U_{ij}} = \frac{\partial (\sum_{a,b,c} U_{ka} V_{ba} V_{bc} U_{lc})_{kl}}{\partial U_{ij}} = \sum_{a,b} (U_{la}\delta_{k=i} + U_{ka}\delta_{l=i}) V_{ba} V_{bj}$

- $\frac{\partial S_{kl}}{\partial V_{ij}} = \frac{\partial (UV^TVU^T)_{kl}}{\partial V_{ij}} = \frac{\partial (\sum_{a,b,c} U_{ka}V_{ba}V_{bc}U_{lc})_{kl}}{\partial V_{ij}} = \sum_a (U_{la}V_{ia}U_{kj} + U_{ka}V_{ia}U_{lj})$

Partial derivative of J_1 wrt U:

$$\frac{\partial J_1}{\partial U_{ij}} = n \sum_{k,l} \frac{\partial \log |S|}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial U_{ij}} = n \sum_{k,l} S^{-1}_{lk} \sum_{a,b} (U_{la}\delta_{k=i} + U_{ka}\delta_{l=i}) V_{ba} V_{bj} \tag{57}$$

$$= n \sum_{a,b,k,l} (S^{-1}_{kl} U_{la}\delta_{k=i} + S^{-1}_{lk} U_{ka}\delta_{l=i}) V_{ba} V_{bj} \tag{58}$$

$$\frac{\partial J_1}{\partial U} = 2nS^{-1}UV^TV \tag{59}$$

Partial derivative of J_1 wrt V:

$$\frac{\partial J_1}{\partial V_{ij}} = n \sum_{k,l} \frac{\partial \log |S|}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial V_{ij}} = n \sum_{k,l} S^{-1}_{lk} \sum_{a} (U_{la}V_{ia}U_{kj} + U_{ka}V_{ia}U_{lj}) \tag{60}$$

$$= n \sum_{a,k,l} (V_{ia}U^T_{al}S^{-1}_{lk}U_{kj} + V_{ia}U^T_{ak}S^{-1}_{kl}U_{lj}) \tag{61}$$

$$\frac{\partial J_1}{\partial V} = 2nVU^TS^{-1}U \tag{62}$$

Partial derivative of J_2 wrt U:

$$\frac{\partial J_2}{\partial U_{ij}} = \sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c \frac{\partial S^{-1}_{bc}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial U_{ij}} \tag{63}$$

$$= - \sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bk} S^{-1}_{lc} \sum_{\alpha,\beta} (U_{l\alpha}\delta_{k=i} + U_{k\alpha}\delta_{l=i}) V_{\beta\alpha} V_{\beta j}$$

$$= -2 \sum_{\alpha,\beta,a,b,c,l} S^{-1}_{ib} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{cl} U_{l\alpha} V^T_{\alpha\beta} V_{\beta j} \tag{64}$$

$$\frac{\partial J_2}{\partial U} = -2S^{-1}[X - \boldsymbol{\mu}]^T [X - \boldsymbol{\mu}] S^{-1}UV^TV \tag{65}$$

Partial derivative of J_2 wrt V:

$$\frac{\partial J_2}{\partial V_{ij}} = \sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c \frac{\partial S^{-1}_{bc}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial V_{ij}} \tag{66}$$

$$= - \sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bk} S^{-1}_{lc} \sum_{\alpha} (U_{l\alpha}V_{i\alpha}U_{kj} + U_{k\alpha}V_{i\alpha}U_{lj})$$

$$= -2 \sum_{\alpha,a,b,c,k,l} V_{i\alpha} U^T_{\alpha l} S^{-1}_{lc} [X_a - \boldsymbol{\mu}]_c [X_a - \boldsymbol{\mu}]_b S^{-1}_{bk} U_{kj} \tag{67}$$

$$\frac{\partial J_2}{\partial V} = -2VU^TS^{-1}[X - \boldsymbol{\mu}]^T [X - \boldsymbol{\mu}] S^{-1}U \tag{68}$$

Partial derivative of J_3 wrt U:

$$\begin{aligned}\frac{\partial J_3}{\partial U_{ij}} &= \beta \sum_{c,k,l} \frac{\partial S_{cc}^{-1}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial U_{ij}} = -\beta \sum_{c,k,l} S_{ck}^{-1} S_{lc}^{-1} \sum_{a,b} (U_{la}\delta_{k=i} + U_{ka}\delta_{l=i}) V_{ba} V_{bj} \\ &= -2\beta \sum_{c,a,b,l} S_{ic}^{-1} S_{cl}^{-1} U_{la} V_{ab}^T V_{bj} \qquad (69)\end{aligned}$$

$$\frac{\partial J_3}{\partial U} = -2\beta S^{-2} U V^T V \tag{70}$$

Partial derivative of J_3 wrt V:

$$\frac{\partial J_3}{\partial U_{ij}} = \beta \sum_{c,k,l} \frac{\partial S_{cc}^{-1}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial U_{ij}} = -\beta \sum_{c,k,l} S_{ck}^{-1} S_{lc}^{-1} \sum_{a} (U_{la} V_{ia} U_{kj} + U_{ka} V_{ia} U_{lj}) \tag{71}$$

$$= -2\beta \sum_{c,a,l,k} V_{ia} U_{al}^T S_{lc}^{-1} S_{ck}^{-1} U_{kj} \tag{72}$$

$$\frac{\partial J_3}{\partial U} = -2\beta V U^T S^{-2} U \tag{73}$$

Partial derivative of J_4 wrt U:

$$\frac{\partial J_4}{\partial U_{ij}} = -\alpha \sum_{c} \frac{\partial \log S_{cc}^{-1}}{\partial S_{cc}^{-1}} \sum_{k,l} \frac{\partial S_{cc}^{-1}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial U_{ij}} \tag{74}$$

$$= \sum_{c,k,l} \frac{1}{S_{cc}^{-1}} S_{ck}^{-1} S_{lc}^{-1} \sum_{a,b} (U_{la}\delta_{k=i} + U_{ka}\delta_{l=i}) V_{ba} V_{bj} \tag{75}$$

$$= 2\alpha \sum_{a,b,c,l} \frac{S_{ic}^{-1}}{S_{cc}^{-1}} S_{cl}^{-1} U_{la} V_{ab}^T V_{bj} \tag{76}$$

$$\frac{\partial J_4}{\partial U} = 2\alpha \frac{S^{-1}}{\mathrm{diag}(S^{-1})} S^{-1} U V^T V \tag{77}$$

Partial derivative of J_4 wrt V:

$$\frac{\partial J_4}{\partial V_{ij}} = -\alpha \sum_{c} \frac{\partial \log S_{cc}^{-1}}{\partial S_{cc}^{-1}} \sum_{k,l} \frac{\partial S_{cc}^{-1}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial V_{ij}} \tag{78}$$

$$= \sum_{c,k,l} \frac{1}{S_{cc}^{-1}} S_{ck}^{-1} S_{lc}^{-1} \sum_{a} (U_{la} V_{ia} U_{kj} + U_{ka} V_{ia} U_{lj}) \tag{79}$$

$$= 2\alpha \sum_{a,c,k,l} V_{ia} U_{al}^T \frac{S_{lc}^{-1}}{S_{cc}^{-1}} S_{ck}^{-1} U_{kj} \tag{80}$$

$$\frac{\partial J_4}{\partial U} = 2\alpha V U^T \frac{S^{-1}}{\mathrm{diag}(S^{-1})} S^{-1} U \tag{81}$$

Partial derivatives for J_5:

$$\frac{\partial J_5}{\partial U} = 2\lambda U \tag{82}$$

$$\frac{\partial J_5}{\partial V} = 2\lambda V. \tag{83}$$

Using a variant of gradient descent (conjugate gradients), we can learn parameters for our modified version of Factor Analysis which uses the trace norm prior in place of the rank constraint.

References

[1] M. Brookes. The matrix reference manual. http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html, 1998.

[2] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[3] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, 1996.

[4] K. B. Petersen and M. S. Pedersen. The matrix cookbook. http://matrixcookbook.com, February 2006.

[5] J. D. M. Rennie. Factor analysis with a trace norm prior. http://people.csail.mit.edu/jrennie/writing, June 2006.

[6] J. D. M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[7] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2005.