
An Approximate Inference Approach for the PCA Reconstruction Error

Manfred Opper

Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ
mo@ecs.soton.ac.uk

Abstract

The problem of computing a resample estimate for the reconstruction error in PCA is reformulated as an inference problem with the help of the replica method. Using the expectation consistent (EC) approximation, the intractable inference problem can be solved efficiently with two variational parameters. A perturbative correction to the result is computed and an alternative simplified derivation is also presented.

1 Introduction

This paper was motivated by recent joint work with Ole Winther on approximate inference techniques (the expectation consistent (EC) approximation [1] related to Tom Minka's EP [2] approach) which allows us to tackle high-dimensional sums and integrals that are necessary for Bayesian probabilistic inference.

I was looking for a nice model to which this approximation would apply. It had to be simple enough so that I would not be bogged down by large numerical simulations. But it had to be nontrivial enough to be of at least modest interest to Machine Learning. With the application of approximate inference to resampling in PCA I hope to be able to stress the following points:

- Approximate efficient inference techniques can be useful in areas of Machine Learning where one would not necessarily assume that they are applicable. This can happen when the underlying probabilistic model is not immediately visible but shows only up as a result of a mathematical transformation.
- Approximate inference methods can be highly robust allowing for analytic continuations of model parameters to the complex plane or even noninteger dimensionalities into areas where the real probabilistic model no longer exists.
- It is not always necessary to use a large number of variational parameters in order to get reasonable accuracy.
- Inference methods could be systematically improved using perturbative corrections.

The work was also stimulated by previous joint work with Dörthe Malzahn [3] on resam-

pling estimates for generalization errors of Gaussian process models and Supportvector-Machines.

2 Resampling estimators for PCA

Principal Component Analysis (PCA) is a well known and widely applied tool for data analysis. The goal is to project data vectors \mathbf{y} from a typically high (d -) dimensional space into an optimally chosen lower (q -) dimensional linear space with $q \ll d$, thereby minimizing the expected projection error $\varepsilon = E\|\mathbf{y} - P_q[\mathbf{y}]\|^2$, where $P_q[\mathbf{y}]$ denotes the projection. E stands for an expectation over the distribution of the data. In practice where the distribution is not available, one has to work with a data sample D_0 consisting of N vectors $\mathbf{y}_k = (y_k(1), y_k(2), \dots, y_k(d))^T$, $k = 1, \dots, N$. Arranging these vectors into a $(d \times N)$ data matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$. Assuming centered data, the optimal subspace is spanned by the eigenvectors \mathbf{u}_l of the $d \times d$ data covariance matrix $\mathbf{C} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T$ corresponding to the q largest eigenvalues λ_k . We will assume that these correspond to all eigenvectors $\lambda_k > \lambda$ above some threshold value λ .

After computing the PCA projection, one would be interested in finding out if the computed subspace represents actually the data well by estimating the average projection error on *novel data* \mathbf{y} (ie not contained in D_0) which are drawn from the same distribution.

Fixing the projection P_q , the error can be rewritten as

$$\varepsilon = \sum_{\lambda_l < \lambda} E \text{Tr} [\mathbf{y}\mathbf{y}^T \mathbf{u}_l \mathbf{u}_l^T] \quad (1)$$

where the expectation is only over \mathbf{y} and the training data are fixed. The *training error*

$$\varepsilon_t = \sum_{\lambda_l < \lambda} \lambda_l^2 \quad (2)$$

can be obtained without knowledge of the distribution but will usually only give an optimistically biased estimate for ε .

2.1 A resampling estimate for the error

New artificial data samples D of arbitrary size M can be created by resampling a number of data points from D_0 with or without replacement. A simple choice would be to choose all data independently with the same probability $1/N$, but other possibilities can be implemented within our formalism. Thus, some \mathbf{y}_i in D_0 will appear multiple times in D and others not at all. The idea of performing PCA on resampled data sets D and testing on the remaining data $D_0 \setminus D$, motivates the following definition of a *resample averaged* reconstruction error

$$\varepsilon_r = \frac{1}{N_0} E_D \left[\sum_{\mathbf{y}_i \notin D; \lambda_l < \lambda} \text{Tr} (\mathbf{y}_i \mathbf{y}_i^T \mathbf{u}_l \mathbf{u}_l^T) \right] \quad (3)$$

as a proxy for ε . E_D is the expectation over the resampling process. This is an estimator of the *bootstrap* type [3,4]. N_0 is the expected number of data in D_0 which are not contained in the random set D . The rest of the paper will discuss a method for efficiently approximating (3).

2.2 Basic formalism

We introduce ‘‘occupation numbers’’ s_i which state how many times \mathbf{y}_i is contained in D . We also introduce two matrices \mathbf{D} and \mathbf{C} . \mathbf{D} is *diagonal random matrix*

$$\mathbf{D}_{ii} = D_i = \frac{1}{\mu\Gamma}(s_i + \epsilon\delta_{s_i,0}) \quad \mathbf{C}(\epsilon) = \frac{\Gamma}{N}\mathbf{YDY}^T. \quad (4)$$

and $\mathbf{C}(0)$ is proportional to the covariance matrix of the *resampled* data. μ is the sampling rate, i.e. $\mu N = E_D[\sum_i s_i]$ is the expected number of data in D (counting multiplicities). The role of Γ will be explained later. Using ϵ , we can generate expressions that can be used in (3) to sum over the data which are not contained in the set D

$$\mathbf{C}'(0) = \frac{1}{\mu N} \sum_j \delta_{s_j,0} \mathbf{y}_j \mathbf{y}_j^T. \quad (5)$$

In the following λ_k and \mathbf{u}_l will always denote eigenvalues and eigenvectors of the data dependent (i.e. random) covariance matrix $\mathbf{C}(0)$.

The desired averages can be constructed from the $d \times d$ *matrix Green’s function*

$$\mathbf{G}(\Gamma) = (\mathbf{C}(0) + \Gamma\mathbf{I})^{-1} = \sum_k \frac{\mathbf{u}_k \mathbf{u}_k^T}{\lambda_k + \Gamma} \quad (6)$$

Using the well known representation of the *Dirac* δ distribution given by $\delta(x) = \lim_{\eta \rightarrow 0^+} \Im \frac{1}{\pi(x-i\eta)}$ where $i = \sqrt{-1}$ and \Im denotes the imaginary part, we get

$$\lim_{\eta \rightarrow 0^+} \frac{1}{\pi} \Im \mathbf{G}(\Gamma - i\eta) = \sum_k \mathbf{u}_k \mathbf{u}_k^T \delta(\lambda_k + \Gamma). \quad (7)$$

Hence, we have

$$\mathcal{E}_r = \mathcal{E}_r^0 + \int_{0^+}^{\lambda} d\lambda' \varepsilon_r(\lambda') \quad (8)$$

where

$$\varepsilon_r(\lambda) = \frac{1}{\pi} \lim_{\eta \rightarrow 0^+} \Im \frac{1}{N_0} E_D \left[\sum_j \delta_{s_j,0} \text{Tr}(\mathbf{y}_j \mathbf{y}_j^T \mathbf{G}(-\lambda - i\eta)) \right] \quad (9)$$

defines the *error density* from all eigenvalues > 0 and \mathcal{E}_r^0 is the contribution from the eigenspace with $\lambda_k = 0$. The latter can also be easily expressed from \mathbf{G} as

$$\mathcal{E}_r^0 = \lim_{\Gamma \rightarrow 0} \frac{1}{N_0} E_D \left[\sum_j \delta_{s_j,0} \text{Tr}(\mathbf{y}_j \mathbf{y}_j^T \Gamma \mathbf{G}(\Gamma)) \right] \quad (10)$$

We can also compute the resample averaged density of eigenvalues from

$$\rho(\lambda) = \frac{1}{\pi\mu N} \lim_{\eta \rightarrow 0^+} \text{Im} E_D [\text{Tr} \mathbf{G}(-\lambda - i\eta)] \quad (11)$$

3 A Gaussian probabilistic model

The matrix Green’s function for $\Gamma > 0$ can be generated from a Gaussian partition function Z . This is a well known construction in statistical physics, and has also been used within the NIPS community to study the distribution of eigenvalues for an average case analysis of PCA [5]. Its use for computing the expected reconstruction error is to my knowledge new.

With the $(N \times N)$ kernel matrix $\mathbf{K} = \frac{1}{N} \mathbf{Y}^T \mathbf{Y}$ we define the Gaussian partition function

$$Z = \int d\mathbf{x} \exp \left[-\frac{1}{2} \mathbf{x}^T (\mathbf{K}^{-1} + \mathbf{D}) \mathbf{x} \right] \quad (12)$$

$$= |\mathbf{K}|^{\frac{1}{2}} \Gamma^{d/2} (2\pi)^{(N-d)/2} \int d^d \mathbf{z} \exp \left[-\frac{1}{2} \mathbf{z}^T (\mathbf{C}(\epsilon) + \Gamma \mathbf{I}) \mathbf{z} \right]. \quad (13)$$

\mathbf{x} is an N dimensional integration variable. The equality can be easily shown by expressing the integrals as determinants.¹ The first representation (12) is useful for computing the resampling average and the second one connects directly to the definition of the matrix Green's function \mathbf{G} . Note, that by its dependence on the kernel matrix \mathbf{K} , a generalization to $d = \infty$ dimensional feature spaces and *kernel PCA* is straightforward. The partition function can then be understood as a certain Gaussian process expectation. We will not discuss this point further.

The *free energy* $F = -\ln Z$ enables us to generate the following quantities

$$-2 \frac{\partial \ln Z}{\partial \epsilon} \Big|_{\epsilon=0} = \frac{1}{\mu N} \sum_{j=1}^N \delta_{s_k, 0} \text{Tr} \mathbf{y}_j \mathbf{y}_j^T \mathbf{G}(\Gamma) \quad (14)$$

$$-2 \frac{\partial \ln Z}{\partial \Gamma} = \frac{d}{\Gamma} + \text{Tr} \mathbf{G}(\Gamma) \quad (15)$$

where we have used (5) for (14). While (14) will be used for (9) and (15) applies to the density of eigenvalues. Note that the definition of the partition function Z requires that $\Gamma > 0$, whereas the application to the reconstruction error (8) needs negative values $\Gamma = -\lambda < 0$. Hence, an analytic continuation of end results must be performed.

4 Resampling average and replicas

(14) and (15) shows that we can compute the desired resampling averages from the expected free energy $-E_D[\ln Z]$. This can be expressed using the ‘‘replica trick’’ of statistical physics (see e.g. [6]) using

$$E_D[\ln Z] = \lim_{n \rightarrow 0} \frac{1}{n} \ln E_D[Z^n], \quad (16)$$

where one attempts an approximate computation of $E_D[Z^n]$ for *integer* n and uses a continuation to real numbers at the end. The n times replicated and averaged partition function (12) can be written in the form

$$Z^{(n)} \doteq E_D[Z^n] = \int dx \psi_1(x) \psi_2(x) \quad (17)$$

where we set $x \doteq (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and

$$\psi_1(x) = E_D \left[\exp \left\{ -\frac{1}{2} \sum_{a=1}^n \mathbf{x}_a^T \mathbf{D} \mathbf{x}_a \right\} \right] \quad \psi_2(x) = \exp \left[-\frac{1}{2} \sum_{a=1}^n \mathbf{x}_a^T \mathbf{K}^{-1} \mathbf{x}_a \right] \quad (18)$$

The *unaveraged* partition function Z (12) is Gaussian, but the *averaged* $Z^{(n)}$ one is not. In fact it is for most interesting resample methods intractable.

¹If \mathbf{K} has zero eigenvalues, a division of $Z |\mathbf{K}|^{\frac{1}{2}}$ maybe necessary. This additive renormalization of the free energy $-\ln Z$ will not influence the subsequent computations.

5 Approximate inference

To approximate $Z^{(n)}$, we will use the EC approximation recently introduced by Opper & Winther [1]. For this method we need two auxiliary distributions

$$p_1(x) = \frac{1}{Z_1} \psi_1(x) e^{-\Lambda_1 x^T x} \quad p_0(x) = \frac{1}{Z_0} e^{-\frac{1}{2} \Lambda_0 x^T x}, \quad (19)$$

where Λ_1 and Λ_0 are ‘‘variational’’ parameters to be optimized. p_1 tries to mimic the intractable $p(x) \propto \psi_1(x) \psi_2(x)$, replacing the multivariate Gaussian ψ_2 by a simpler, i.e. tractable diagonal one. One may think of using a general diagonal matrix Λ_1 , but we will restrict ourselves in the present case to the simplest case of a spherical Gaussian with a *single parameter* Λ_1 .

The strategy is to split $Z^{(n)}$ into a product of Z_1 and a term that has to be further approximated:

$$\begin{aligned} Z^{(n)} &= Z_1 \int dx p_1(x) \psi_2(x) e^{\Lambda_1 x^T x} \\ &\approx Z_1 \int dx p_0(x) \psi_2(x) e^{\Lambda_1 x^T x} \equiv Z_{EC}^{(n)}(\Lambda_1, \Lambda_0). \end{aligned} \quad (20)$$

The approximation replaces the intractable average over p_1 by a tractable one over p_0 . To optimize Λ_1 and Λ_0 we argue as follows: We try to make p_0 as close as possible to p_1 by matching the moments $\langle x^T x \rangle_1 = \langle x^T x \rangle_0$. The index denotes the distribution which is used for averaging. By this step, Λ_0 becomes a function of Λ_1 . Second, since the true partition function $Z^{(n)}$ is *independent* of Λ_1 , we expect that a good approximation to $Z^{(n)}$ should be stationary with respect to variations of Λ_1 . Both conditions can be expressed by the requirement that $\ln Z_{EC}^{(n)}(\Lambda_1, \Lambda_0)$ must be stationary with respect to variations of Λ_1 and Λ_0 .

Within this EC approximation we can carry out the replica limit $E_D[\ln Z] \approx \ln Z_{EC} = \lim_{n \rightarrow 0} \frac{1}{n} \ln Z_{EC}^{(n)}$ and get after some calculations

$$\begin{aligned} -\ln Z_{EC} &= -E_D \left[\ln \int d\mathbf{x} e^{-\frac{1}{2} \mathbf{x}^T (\mathbf{D} + (\Lambda_0 - \Lambda) \mathbf{I}) \mathbf{x}} \right] - \\ &\quad - \ln \int d\mathbf{x} e^{-\frac{1}{2} \mathbf{x}^T (\mathbf{K}^{-1} + \Lambda \mathbf{I}) \mathbf{x}} + \ln \int d\mathbf{x} e^{-\frac{1}{2} \Lambda_0 \mathbf{x}^T \mathbf{x}} \end{aligned} \quad (21)$$

where we have set $\Lambda = \Lambda_0 - \Lambda_1$. The resampling average in the approximation (21) appears much simpler than in the original free energy, especially if we assume a sampling scheme where all data are sampled independently from each others. E.g., we could take *Poisson* probabilities $p(s) = e^{-\mu u} \frac{\mu^s}{s!}$ which gives a good approximation to the case of resampling μN points with replacement.

The variational equations which make (21) stationary are

$$E_D \left(\frac{1}{\Lambda_0 - \Lambda + D_i} \right) = \frac{1}{\Lambda_0} \quad \frac{1}{N} \sum_k \frac{\omega_k}{1 + \omega_k \Lambda} = \frac{1}{\Lambda_0} \quad (22)$$

where ω_k are the eigenvalues of the matrix \mathbf{K} . The variational equations can be solved without problems numerically in the region $\Gamma = -\lambda < 0$ where the original partition function does not exist. The resulting parameters Λ_0 and Λ will usually come out as *complex* numbers. Their corresponding values can be used for the computation of the resampling errors, e.g. we get

$$\mathcal{E}_b = -\frac{2\mu N}{N_0 \pi} \Im \frac{\partial \ln Z_{EC}}{\partial \epsilon} \Big|_{\epsilon=0} = \frac{N}{\pi \lambda} \Im \left(\frac{1}{\Lambda_0 - \Lambda} \right). \quad (23)$$

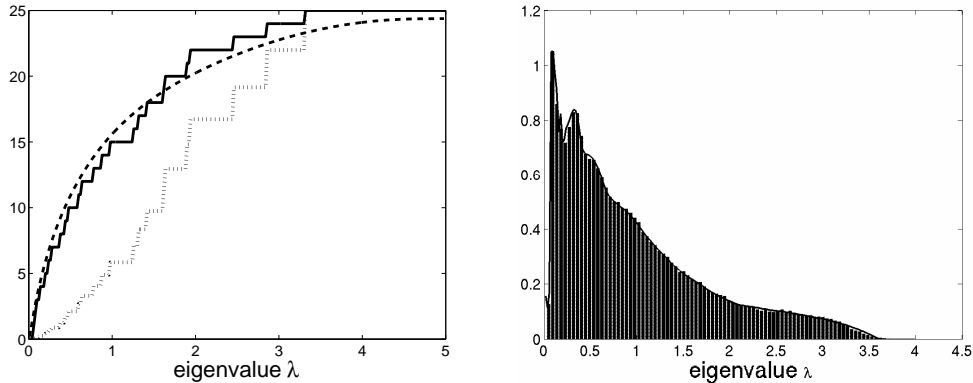


Figure 1: *Left*: Errors for PCA on $N = 32$ spherically Gaussian data with $d = 25$. Sampling rate $\mu = 3$. Smooth curve: approximate resampled error estimate, upper step function: true error. Lower step function: Training error. *Right*: Comparison of EC approximation (line) and simulation (histogram) of the resampled density of eigenvalues for $N = 50$ spherically Gaussian data of dimensionality $d = 25$. The sampling rate was $\mu = 3$.

6 Simulations

Eliminating the parameter Λ_0 from (22) it is possible to reduce the numerical computations to solving a nonlinear equation for a single *complex* parameter Λ which can be solved easily and fast by a Newton method. While the analytical results are based on *Poisson* statistics, the simulations of random resampling was performed by choosing a *fixed* number (equal to the expected number of the Poisson distribution) of data at random with replacement.

The first simulation was for a set of data generated at random from a spherical Gaussian. To show that resampling maybe useful, we give on the left hand side of (1) the reconstruction error as a function of the value of λ below which eigenvalues are discarded. The smooth function is the approximate resampling error ($3\times$ oversampled to leave not many data out of the samples) from our method. The upper step function gives the true reconstruction error (easy to calculate for spherical data) for the full sample. The lower step function is the training error. The right panel demonstrates the *accuracy* of the approximation on a similar set of data. We compare the analytical resampled density of states with the results of a true resampling experiment, where eigenvalues for many samples are counted into small bins. The theoretical curve follows closely the experiment.

Since the good accuracy might be attributed to the high symmetry of the toy data, we have also performed experiments on a set of $N = 100$ handwritten digits with $d = 784$. The results in (2) are promising. Although the density of eigenvalues is more accurate than the resampling error, the latter comes still out reasonable.

7 Corrections

I will show next that the EC approximation can be augmented by a perturbation expansion. Going back to (20), we can write

$$\frac{Z^{(n)}}{Z_1} = \int dx p_1(x) \psi_2(x) e^{\Lambda_1 x^T x} = \int dx \psi_2(x) e^{\frac{1}{2} \Lambda x^T x} \left\{ \int \frac{dk}{(2\pi)^{Nn}} e^{-ik^T x} \chi(k) \right\}$$

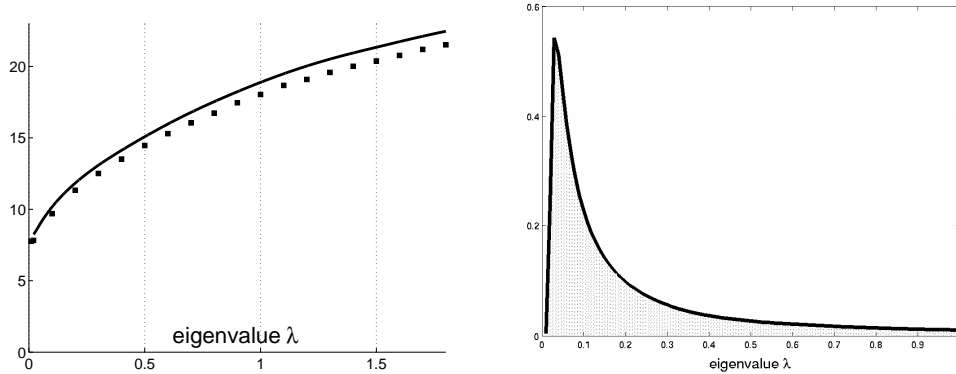


Figure 2: *Left*: Resampling error ($\mu = 1$) for PCA on a set of 100 handwritten digits (“5”) with $d = 784$. The approximation (line) for $\mu = 1$ is compared with simulations of the random resampling. *Right*: Resampled density of eigenvalues for the same data set. Only the nonzero eigenvalues are shown.

where $\chi(k) \doteq \int dx p_1(x) e^{-ik^T x}$ is the *characteristic function* of the density p_1 (19). $\ln \chi(k)$ is the *cumulant generating function*. Using the symmetries of the density p_1 , we can perform a power series expansion of $\ln \chi(k)$, which starts with a quadratic term (second cumulant)

$$\ln \chi(k) = -\frac{M_2}{2} k^T k + R(k), \quad (25)$$

where $M_2 = \langle \mathbf{x}_a^T \mathbf{x}_a \rangle_1$. It can be shown that if we neglect $R(k)$ (containing the higher order cumulants) and carry out the integral over k , we end up replacing p_1 by a simpler Gaussian p_0 with matching moments M_2 , i.e. the EC approximation. Higher order corrections to the free energy $-E_D[\ln Z] = -\ln Z_{EC} + \Delta F_1 + \dots$ can be obtained perturbatively by writing $\chi(k) = e^{-\frac{M_2}{2} k^T k} (1 + R(k) + \dots)$. This expansion is similar in spirit to *Edgeworth expansions* in statistics. The present case is more complicated by the extra dimensions introduced by the *replicating* of variables and the limit $n \rightarrow 0$. After a lengthy calculation one finds for the lowest order correction (containing the monomials in k of order 4) to the free energy:

$$\Delta F_1 = -\frac{1}{4} \sum_i \left(\Lambda_0 (\mathbf{K}^{-1} + \Lambda \mathbf{I})_{ii}^{-1} - 1 \right)^2 \times \sum_i E_D \left(\frac{\Lambda_0}{\Lambda_0 - \Lambda + D_i} - 1 \right)^2 \quad (26)$$

To illustrate the effect of ΔF_1 , I have used it so far only for obtaining a correction to the reconstruction error in the “zero-subspace” using (10) and (14) of the digit data. Preliminary results are not yet conclusive. For sufficiently small sampling rates corrections increase and reach a maximum at $\mu = 1.5$. For $m < 3$ corrections are quite accurate compared with the simulations (averaged over 5000 samples). E.g., with $\mu = 1$, we have $\mathcal{E}_r^0(th) = 7.0774$ from the theory and $\mathcal{E}_r^0(sim) = 7.485$. The difference of 0.4076 is well accounted for by the computed correction of 0.384. Similarly for $\mu = 3$, we had $\mathcal{E}_r^0(th) = 3.8035$ and $\mathcal{E}_r^0(sim) = 4.096$ with a difference of 2.93 compared to the theoretical correction of 0.2668. The predicted rapid decrease of corrections for larger μ is not found in the simulations. The difference between theory and simulation is rather increasing. For $\mu = 4$, we have $\mathcal{E}_r^0(sim) - \mathcal{E}_r^0(th) = 0.72$ compared to the predicted value of 0.05. This behaviour needs further investigation.

8 The calculation without replicas

Knowing with hindsight how the final EC result (21) looks like, we can rederive it using another method which does not rely on the “replica trick”. We first write down an exact expression for $-\ln Z$ *before* averaging. Expressing Gaussian integrals by determinants yields

$$-\ln Z = -\ln \int d\mathbf{x} e^{-\frac{1}{2}\mathbf{x}^T(\mathbf{D}+(\Lambda_0-\Lambda)\mathbf{I})\mathbf{x}} - \ln \int d\mathbf{x} e^{-\frac{1}{2}\mathbf{x}^T(\mathbf{K}^{-1}+\Lambda\mathbf{I})\mathbf{x}} + \quad (27)$$

$$+ \ln \int d\mathbf{x} e^{-\frac{1}{2}\Lambda_0\mathbf{x}^T\mathbf{x}} + \frac{1}{2} \ln \det(\mathbf{I} + \mathbf{r})$$

where the matrix \mathbf{r} has elements

$$\mathbf{r}_{ij} = \left(1 - \frac{\Lambda_0}{\Lambda_0 - \Lambda + D_i}\right) \left(\Lambda_0 (\mathbf{K}^{-1} + \Lambda\mathbf{I})^{-1} - \mathbf{I}\right)_{ij}. \quad (28)$$

The EC approximation is obtained by simply neglecting \mathbf{r} . Corrections to this are found by expanding

$$\ln \det(\mathbf{I} + \mathbf{r}) = \text{Tr} \ln(\mathbf{I} + \mathbf{r}) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \text{Tr}(\mathbf{r}^k) \quad (29)$$

The first order term in the expansion (29) vanishes after averaging (see (22)) and the second order term gives exactly the correction of the cumulant method (26).

9 Outlook

It will be interesting to extend the perturbative framework for the computation of corrections to inference approximations to other, more complex models. However, our results indicate that the use and convergence of such perturbation expansion needs to be critically investigated and may not give a clear indication of the accuracy of the approximation. The alternative derivation for our simple model could present an interesting ground for testing these ideas.

Acknowledgments

I would like to thank Ole Winther for the great collaboration on the EC approximation.

References

- [1] Manfred Opper and Ole Winther. Expectation consistent free energies for approximate inference. In *NIPS 17*, 2005.
- [2] T. P. Minka. Expectation propagation for approximate Bayesian inference. In *UAI 2001*, pages 362–369, 2001.
- [3] D. Malzahn and M. Opper. An approximate analytical approach to resampling averages. *Journal of Machine Learning Research*, pages 1151–1173, 2003.
- [4] B. Efron, R. J. Tibshirani. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57, Chapman & Hall, 1993.
- [5] D. C. Hoyle and M. Rattay. Limiting form of the sample covariance matrix eigenspectrum in PCA and kernel PCA. In *NIPS 16*, 2003.
- [6] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, 2001).